Supplementary Materials for Generative Model-Based Fusion for Improved Few-Shot Semantic Segmentation of Infrared Images

Junno Yun, Mehmet Akçakaya
University of Minnesota
{yun00049, akcakaya}@umn.edu

These supplementary materials provide detailed descriptions of the Few-Shot Segmentation (FSS) architectures for both the baseline and the proposed networks. Additionally, they include the ablation studies mentioned in the main text.

A. FSS Baseline Architecture

The architecture of the baseline is depicted in Fig. 1, featuring a meta-learner, a base learner, and an ensemble module. The base learner, PSPNet [15] with a ResNet-50/101 [5] backbone, is trained in a supervised manner on base classes that are already known, yielding the prediction F_{Base} from the query set. Its role is to predict regions of base classes in query images and suppress falsely activated regions of base categories in the meta learner output. The base learner trains the encoder to extract essential features from support and query sets, enabling the meta learner to focus on discerning relationships between these features.

The meta-learner leverages intermediate- and high-level features from the support and query sets to generate five key features to evaluate their relationship. All generated features are concatenated and processed through an atrous spatial pyramid pooling (ASPP) module [2], followed by a decoder to produce a binary meta prediction mask F_{Meta}.

The ensemble module (Fig. 2) integrates these two predictions to produce the final foreground and background probability maps, culminating in the generation of F_{final}. It initially estimates the scene differences between query-support image pairs by calculating the Gram matrices of the support and query images using low-level features $F_{\text{S}}^{low}, F_{\text{Q}}^{low} \in \mathbb{R}^{C \times H_l \times W_l}$ extracted from the shared encoder blocks during the training of the meta learner, where C, H_l, and W_l are the dimensions of the low-level features. The Gram matrices are calculated as follows:

$$G_S = R_S R_S^T \in \mathbb{R}^{C \times C} \tag{1}$$

$$G_Q = R_Q R_Q^T \in \mathbb{R}^{C \times C} \tag{2}$$

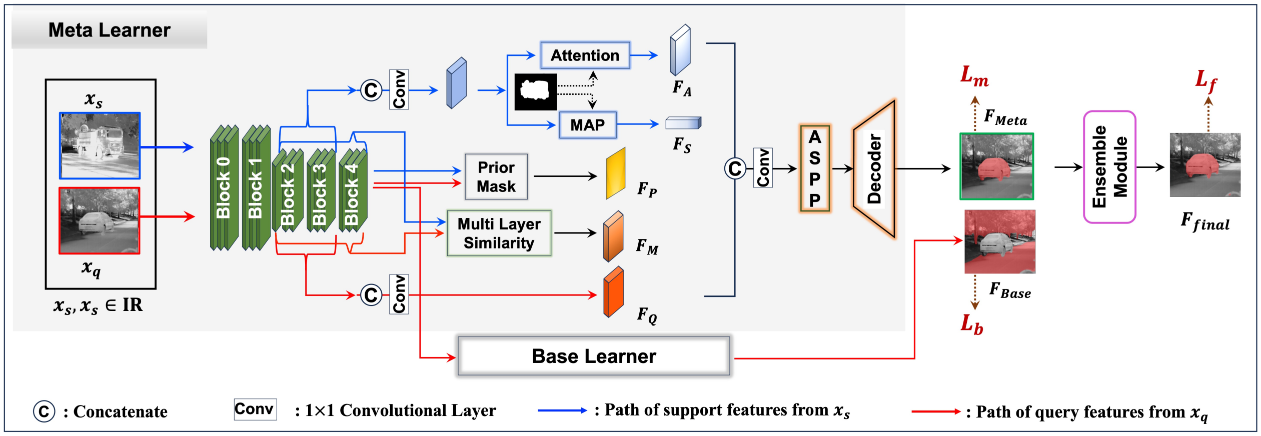

Figure 1. Illustration of the architecture of our baseline (MSANet) [7].

where R_S and R_Q are reshaped tensors of F_S^{low} and F_Q^{low}, which have dimensions $C \times N$ (with $N = H_l \times W_l$). The Frobenius norm is then computed on the difference between these Gram matrices to derive the adjustment factor map, which guides the adjustment process as calculated:

$$F_\psi = Reshape(\|G_S - G_Q\|_F) \in \mathbb{R}^{H_p \times W_p} \qquad (3)$$

where $\| \cdot \|_F$ indicates the Frobenius norm, and $Reshape$ is a function reshaping the input tensor to the size of $H_p \times W_p$, which are the dimensions of the meta and base predictions.

In the adjustment process, the foreground and background of the meta prediction are separately concatenated with the adjustment factor map, followed by a 1×1 convolutional layer, yielding F_{fg_final} and F_{bg_ψ}. Afterward, in the ensemble process, the base prediction map from the base learner and the adjusted background map F_{bg_ψ} are ensembled through concatenation and a convolutional layer, yielding F_{bg_final}. Finally, the final prediction map F_{final} is obtained by concatenating F_{fg_final} and F_{bg_final}.

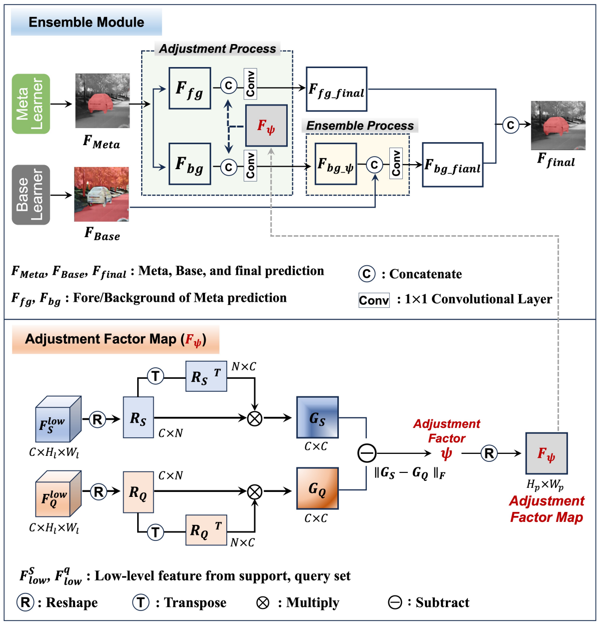

Figure 2. Illustration of a standard ensemble module.

B. Proposed Overall Architecture with Auxiliary Information

In the proposed methods, depicted in Fig. 3, two identical meta-learners, one for each domain, generate meta predictions F_{Meta}^{IR} and F_{Meta}^{RGB} respectively, uti-

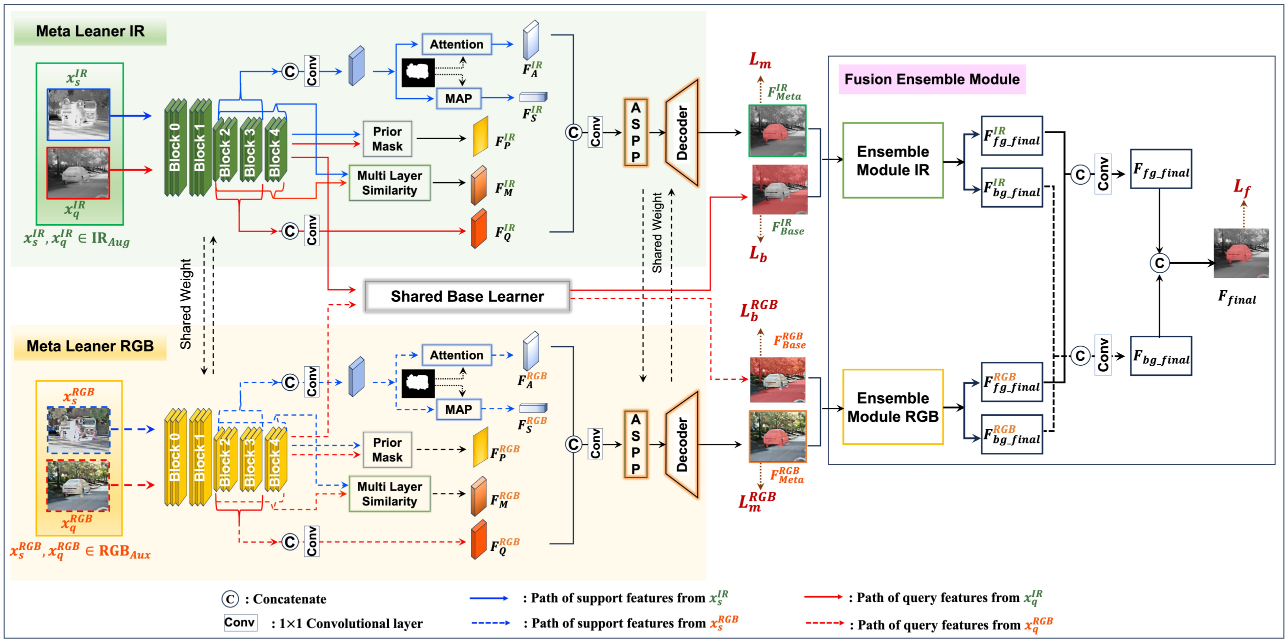

Figure 3. The overall architecture of the proposed methods

Data	Model	ResNet-50												ResNet-101											
		1-shot						5-shot						1-shot						5-shot					
		Fold-0	Fold-1	Fold-2	Fold-3	**MIoU**	**FB-IoU**	Fold-0	Fold-1	Fold-2	Fold-3	**MIoU**	**FB-IoU**	Fold-0	Fold-1	Fold-2	Fold-3	**MIoU**	**FB-IoU**	Fold-0	Fold-1	Fold-2	Fold-3	**MIoU**	**FB-IoU**
SODA [9]	PFENet [13]	32.10	24.50	32.10	32.75	30.36	55.00	37.71	28.16	44.24	39.83	37.49	61.30	35.16	26.78	45.38	38.71	36.51	58.72	41.74	33.39	59.45	49.04	45.91	66.63
	HSNet [10]	33.74	23.97	33.74	35.24	31.67	58.43	40.77	29.79	51.03	41.31	40.73	64.25	36.06	26.07	40.50	40.12	35.69	60.69	43.17	33.84	55.12	47.83	44.99	66.50
	BAM [8]	38.65	35.32	42.80	49.47	41.56	65.94	43.89	39.54	52.42	61.73	49.40	71.77	39.21	36.51	51.98	53.77	45.37	68.26	46.85	40.63	59.28	61.45	52.05	73.68
	VAT [6]	35.40	26.65	38.88	35.12	34.01	59.58	40.51	31.68	48.51	43.12	40.95	63.39	36.63	30.18	45.89	38.71	37.85	62.33	42.97	36.27	56.50	47.86	45.90	67.20
	MSANet [7]	43.58	37.35	47.45	55.34	45.93	70.15	48.15	42.00	61.23	59.39	52.69	74.44	42.39	**38.73**	**53.59**	58.11	48.20	70.72	48.20	**42.72**	64.34	63.88	54.78	75.18
	MSI [11]	32.32	24.98	42.19	37.82	34.33	59.22	37.07	29.68	51.03	45.28	40.76	62.75	31.29	28.60	46.65	40.78	36.83	61.50	36.91	34.17	54.09	48.58	43.44	65.57
	Ours (Method3)	**44.01**	**38.48**	**50.60**	**61.19**	**48.57**	**71.81**	**50.92**	**42.79**	**60.09**	**65.36**	**54.79**	**75.98**	**45.38**	38.04	52.51	**60.90**	**49.21**	**72.21**	**51.96**	42.50	**65.64**	**65.74**	**56.46**	**76.14**
SCUTSEG [14]	PFENet [13]	47.15	21.76	38.44	6.74	28.52	63.52	48.43	25.65	40.92	7.51	30.63	65.68	49.16	26.42	37.52	12.75	31.46	65.06	53.75	33.71	39.63	23.39	37.62	67.37
	HSNet [10]	36.72	20.65	27.47	14.98	24.95	61.30	42.24	26.88	32.60	18.01	29.93	64.62	38.46	20.99	28.34	8.66	24.11	61.38	45.07	27.47	34.37	14.24	30.29	64.71
	BAM [8]	47.48	25.49	44.73	1.96	29.92	65.93	50.72	30.84	45.99	5.16	33.18	67.91	50.47	29.61	40.30	9.47	32.46	65.35	52.65	38.72	44.14	**27.53**	40.76	69.04
	VAT [6]	37.28	22.18	34.14	10.2	26.13	61.36	44.54	27.83	36.87	12.45	30.42	63.67	39.11	24.72	33.04	13.55	27.61	62.55	44.44	30.54	38.80	18.56	33.08	65.12
	MSANet [7]	48.35	27.29	46.51	3.97	31.53	66.68	51.69	35.46	48.53	13.20	37.22	68.68	50.38	30.18	44.68	11.83	34.27	66.96	52.48	39.26	47.18	21.75	40.17	68.49
	MSI [11]	39.16	25.15	32.54	7.39	26.06	62.54	43.35	28.27	34.44	7.22	28.32	63.93	40.72	25.40	28.53	7.69	25.59	62.03	44.50	28.78	30.61	6.31	27.55	63.61
	Ours (Method3)	**55.44**	**34.30**	**55.09**	**16.51**	**40.33**	**70.48**	**57.46**	**41.75**	**54.29**	**28.00**	**45.38**	**72.43**	**62.36**	**37.11**	**52.91**	**14.45**	**41.71**	**71.14**	**66.49**	**46.83**	**56.03**	25.90	**48.81**	**73.62**

Table 1. Comparison with SOTA methods on the SODA and SCUTSEG datasets under 1-shot and 5-shot settings, using ResNet-50 and ResNet-101 backbone networks. Entries in **bold** indicate the best performance, while those underlined denote the second best.

lizing shared encoder, ASPP, and decoder modules. A shared base learner produces predictions F^{IR}_{Base} and F^{RGB}_{Base}.

The proposed IR-RGB fusion ensemble module comprises two identical ensemble modules for each domain. Each ensemble module integrates the meta prediction and base prediction with their own adjustment factor map. The ensembled foreground prediction maps from the IR and RGB domains are then merged with 1×1 convolutional layers, following the same process for background predictions. The proposed fusion ensemble module complements results from each domain to produce the final foreground/background probability maps and F_{final}.

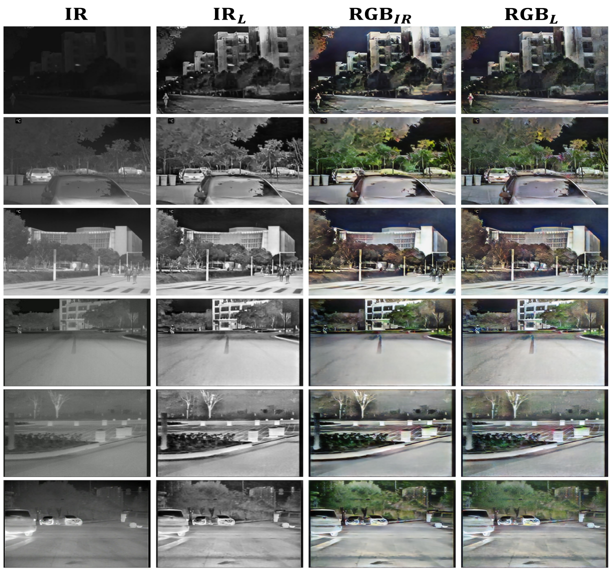

Figure 4. The results of I2I translations using SynDiff. The first three rows display samples from the SODA dataset, while the fourth row onwards exhibits samples from the SCUTSEG dataset.

C. Qualitative Evaluation of the Adversarial Generative Diffusion Models

We demonstrate qualitative results of the generated lightness and RGB images on two different IR datasets, showcasing the effectiveness of our approach in generating realistic and visually appealing results.

Generated Lightness Data for Data Augmentation. The second column in Fig. 4 presents examples of generated lightness domain images IR$_L$. These images exhibit enhanced contrast compared to the original IR datasets, retaining the essential properties and characteristics. IR$_L$ images provide valuable data augmentation, adding diversity and variations to the training data without extra annotations.

Generated RGB Data for Auxiliary Information. The third and fourth columns in Fig. 4 depict generated RGB images. The IR and IR$_L$ are converted into RGB$_{IR}$ and RGB$_L$, which enrich the channel information. While certain categories like trees, skies, and roads translate clearly, others like cars may lack clear color distinction. Despite potential color ambiguities, these images maintain distinct object contours and contain valuable channel information.

Note that our fusion model is designed to distill information from synthetic RGB images to improve segmentation, even if the synthetic images are not highly realistic as in conventional image-to-image (I2I) translation problems. Thus, the goodness of the synthetic RGB images is not critical to our developments, and as such was not a focus. In addition, the goodness of the synthetic RGB images may be evaluated by using RGB-IR segmentation methods designed for paired data by substituting true RGB with synthesized RGB data. However, this approach was not pursued due to the scarcity of RGB-T datasets with annotations suitable for FSS settings. Although some datasets containing RGB-T pairs with labels do exist (e.g., PST900 [12]), they are not suitable for FSS tasks as they contain only four categories. While Multi-Spectral-4^i, derived from the MFNet [4] RGB-T dataset, has been utilized for FSS tasks [1,16], we were unable to leverage it due to the unavailability of publicly released code. Furthermore, since our datasets do not contain true IR-RGB pairs, both evaluations were not possible.

D. Detailed Comparison with SOTA Methods.

Implementation Details. SOTA models were originally designed for RGB datasets and utilize pre-trained backbone networks on ImageNet [3]. For a fair comparison, all models were implemented using the same backbone networks, ResNet-50 and ResNet-101, pre-trained on our IR datasets following the same training protocol as the encoder of the base learner [8]. Consistent training augmentation, optimization strategies, and evaluation procedures were applied across all models, with the exception of learning rates and batch sizes, which were optimized for maximal performance on each individual network.

Results and Analysis. Tab. 1 presents further results associated with Table 3 of the main text, detailing each fold for SOTA methods and our proposed approach on SODA and SCUTSEG datasets. Our proposed method with ResNet-101 achieves the highest mIoU and FB-IoU scores across all folds when compared to SOTA models in both datasets. Although our method yields the highest average scores of four folds, certain individual folds (folds 1 and 2 in the 1-shot setting, fold 1 in the 5-shot setting for SODA, and fold 3 in the 5-shot setting for SCUTSEG) exhibit second-place performance.

E. Ablation Study

To effectively train the encoder, we utilize the proposed datasets IR$_{Aug}$ and RGB$_{Aux}$ during the base learner stage as detailed in the main text. In this ablation study, we compare the base learner's predictions from different encoders to assess their impact. Tab. 2 presents the mean mIoU scores of base learner predictions across four folds for the test set. For both ResNet-50 and ResNet-101 as encoders, the results indicate enhanced mIoU scores with augmented and auxiliary data in both datasets, with all proposed methods. This improvement signifies improved capturing of features from both the support and query sets.

	SODA		SCUTSEG	
Method	**ResNet-50**	**ResNet-101**	**ResNet-50**	**ResNet-101**
Baseline	54.09	56.20	43.26	45.01
Method 1	55.67	**57.32**	**47.65**	48.41
Method 2	55.66	57.00	46.79	48.39
Method 3	**56.10**	57.18	47.35	**50.81**

Table 2. The mean mIoU of the base learner's predictions across four folds for the test set.

References

[1] Yanqi Bao, Kechen Song, Jie Wang, Liming Huang, Hongwen Dong, and Yunhui Yan. Visible and thermal images fusion architecture for few-shot semantic segmentation. *Journal of Visual Communication and Image Representation*, 80:103306, 2021. 4

[2] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062*, 2014. 1

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 4

[4] Qishen Ha, Kohei Watanabe, Takumi Karasawa, Yoshitaka Ushiku, and Tatsuya Harada. Mfnet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5108–5115. IEEE, 2017. 4

[5] 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*, pages 770–778, 2016. 1

[6] Sunghwan Hong, Seokju Cho, Jisu Nam, Stephen Lin, and Seungryong Kim. Cost aggregation with 4d convolutional swin transformer for few-shot segmentation. In *European Conference on Computer Vision*, pages 108–126. Springer, 2022. 3

[7] Ehtesham Iqbal, Sirojbek Safarov, and Seongdeok Bang. Msanet: Multi-similarity and attention guidance for boosting few-shot segmentation. *arXiv preprint arXiv:2206.09667*, 2022. 1, 3

[8] 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, 4

[9] Chenglong Li, Wei Xia, Yan Yan, Bin Luo, and Jin Tang. Segmenting objects in day and night: Edge-conditioned cnn for thermal image semantic segmentation. *IEEE Transactions on Neural Networks and Learning Systems*, 32(7):3069–3082, 2020. 3

[10] Juhong Min, Dahyun Kang, and Minsu Cho. Hypercorrelation squeeze for few-shot segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6941–6952, 2021. 3

[11] Seonghyeon Moon, Samuel S Sohn, Honglu Zhou, Sejong Yoon, Vladimir Pavlovic, Muhammad Haris Khan, and Mubbasir Kapadia. Msi: Maximize support-set information for few-shot segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19266–19276, 2023. 3

[12] Shreyas S Shivakumar, Neil Rodrigues, Alex Zhou, Ian D Miller, Vijay Kumar, and Camillo J Taylor. Pst900: Rgb-thermal calibration, dataset and segmentation network. In *2020 IEEE international conference on robotics and automation (ICRA)*, pages 9441–9447. IEEE, 2020. 4

[13] 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*, 44(2):1050–1065, 2020. 3

[14] Haitao Xiong, Wenjie Cai, and Qiong Liu. Mcnet: Multi-level correction network for thermal image semantic segmentation of nighttime driving scene. *Infrared Physics & Technology*, 113:103628, 2021. 3

[15] 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. 1

[16] Ying Zhao, Kechen Song, Yiming Zhang, and Yunhui Yan. Bmdenet: Bi-directional modality difference elimination network for few-shot rgb-t semantic segmentation. *IEEE Transactions on Circuits and Systems II: Express Briefs*, 2023. 4