BMDENet: Bi-Directional Modality Difference Elimination Network for Few-Shot RGB-T Semantic Segmentation

Ying Zhao, Kechen Song, *Member, IEEE*, Yiming Zhang, and Yunhui Yan

***Abstract*—Few-shot semantic segmentation (FSS) aims to segment the target prospects of query images using a few labeled support samples. Compared with the fully-supervised methods, FSS has a greater ability to generalize to unseen classes and reduce the pressure to label large pixel-level datasets. To cope with the complex outdoor lighting environment, we introduce the thermal infrared images (T) to the FSS task. However, the existing RGB-T FSS methods all ignore the differences between various modalities for direct fusion, which may hinder cross-modal information interaction. Also considering the effect of successive downsampling on the results, we propose a bi-directional modality difference elimination network (BMDENet) to boost the segmentation performance. Concretely, the bi-directional modality difference elimination module (BMDEM) reduces the heterogeneity between RGB and thermal images in the prototype space. The residual attention fusion module (RAFM) mines the bimodal features to fully fuse the cross-modal information. In addition, the mainstay and subsidiary enhancement module (MSEM) enhances the fusion features for the existing problem of the advanced model. Extensive experiments on Tokyo Multi-Spectral-4^{i} dataset prove that BMDENet achieves the state-of-the-art on both 1- and 5-shot settings.**

***Index Terms*—Few-shot semantic segmentation, RGB-T FSS, difference elimination, cross-modal.**

I. Introduction

SEMANTIC segmentation [1], as one of the complex tasks of scene understanding, aims to segment a given image into several visually meaningful regions and assign a semantic label to each pixel in the image. It plays an important role in many image analysis tasks, such as robot navigation [2], [3], medical imaging diagnosis [4], [5], sports tactics analysis [6], and other fields [7], [8], [9]. Moreover, with the advanced achievements of supplementary thermal infrared images in various fields [10], [11], this trend has also stimulated the development of RGB-T semantic segmentation. Zhou et al. created a novel MFFENet [12] to explore the cross-modal fusion of RGB and thermal features under multi-stage for RGB-T urban scene understanding. These fully-supervised methods rely on large-scale manually labeled pixel-level images (e.g., PASCAL VOC [13] and COCO [14]) and generalize badly. The emergence of FSS provides a solution.

Shaban et al. [15] are pioneers in formally defining the FSS problem by proposing a classical two-branch network named OSLSM. Prototype-based few-shot segmentation has attracted lots of research attention in the past, such as SG-One [16], CANet [17], PFENet [18], ASGNet [19], etc. But prototypes alone may not be sufficient to represent the features of the support set. They will inevitably lead to information loss. Recently, HSNet [20] explores the direction of establishing element-to-element dense correspondence through 4D convolutions. It gains a huge performance improvement. ASNet [36] improves on HSNet. SS-PFENet [37] and A-MCG [38] enhance self-supervision in one-shot learning.

When facing variable illumination or complex background, the limited and fixed RGB support images can degrade segmentation performance. Therefore, multimodal FSS began to develop, and certain achievements have been made up to now. A dual-stream deep neural network (RDNet) was first proposed by Zhang et al. [21] in 2020. It uses late fusion to connect probabilistic maps generated by RGB and depth for joint prediction. In 2021, V-TFSS designed by Bao et al. [22] uses edge similarity in RGB and thermal images to fuse bimodal information. The latest ADFNet [23] learns class-specific prototypes from RGB and depth channels and utilizes an attention-based fusion module to better fuse two modalities.

However, the above approaches may still suffer from two limitations: 1) The existing multimodal semantic segmentation models ignore the modality differences between RGB and complementary images caused by different imaging mechanisms. 2) The current state-of-the-art FSS model HSNet [20] takes segmentation performance to a new stage, but when the image resolution is not high enough, successive convolution and downsampling lead to a partial loss of detailed information and hamper the learning of the meta-learner.

In a word, the main contributions of our works are summarized as follows:

Manuscript received 20 February 2023; revised 14 March 2023; accepted 18 May 2023. Date of publication 22 May 2023; date of current version 27 October 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 51805078; in part by the Fundamental Research Funds for the Central Universities under Grant N2103011; in part by the Central Guidance on Local Science and Technology Development Fund under Grant 2022JH6/100100023; and in part by 111 Project under Grant B16009. This brief was recommended by Associate Editor J. Meng. *(Corresponding authors: Kechen Song; Yunhui Yan.)*

The authors are with the School of Mechanical Engineering and Automation, the National Frontiers Science Center for Industrial Intelligence and Systems Optimization, and the Key Laboratory of Data Analytics and Optimization for Smart Industry, Ministry of Education, Northeastern University, Shenyang 110819, Liaoning, China (e-mail: songkc@me.neu.edu.cn; yanyh@mail.neu.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCSII.2023.3278941.

Digital Object Identifier 10.1109/TCSII.2023.3278941

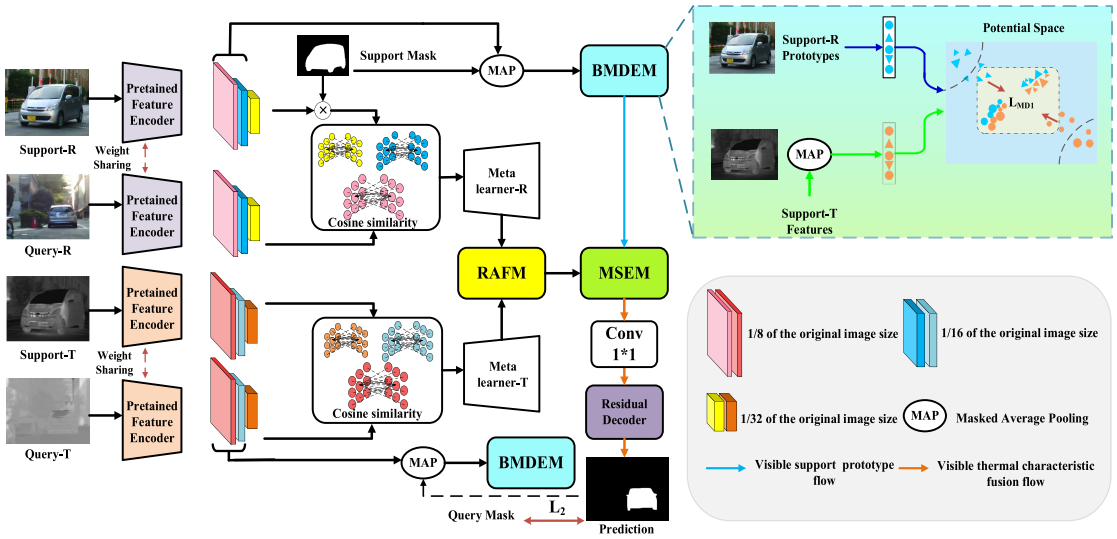

Fig. 1. Architecture of the proposed BMDENet.

1) For the first limitation, we present BMDEM to solve the compatibility problem caused by the differences in imaging mechanisms between different modal images and impose supervised signals on the process of eliminating modal differences.
2) For the second limitation, the proposed MSEM compensates for the lost information from different perspectives effectively.
3) We adopt RAFM to improve the quality of multimodal representation by first highlighting the specific advantages of each modality, and then mining the deep semantic information of multimodal features.
4) Experiments on the Tokyo Multi-Spectral-4^i dataset [22] (mIoU for 1-shot: 38.9%, 5-shot: 41.08%) demonstrate the superiority of the BMDENet over existing methods.

II. Proposed Method

In brief, we create a bi-directional modality difference elimination network (BMDENet) for the RGB-T FSS. We show the overall scheme of the proposed method in Fig. 1.

A. BMDEM

We first use the convolutional neural network ResNet-50 [24] pre-trained on ImageNet [25] as the backbone. And we represent the embedding networks for RGB and thermal images as f_{rgb}, f_{th}. Formally, we compute the rich mapping of intermediate features in a task as $Frgb = frgb(x_j^R)$, $F_{th} = f_{th}(x_j^T)$ and represent them as a dense set of pairs $\{(F_l^s, F_l^q)\}_{l=1}^N$, $\forall F \in RGB \cup T$. Then, we select three pairs and filter out the background information in the support images with the support masks, resulting in $\{(F_l^{'s}, F_l^q)\}_{l=1}^3$.

As shown in Fig. 1, we adopt the same bi-directional modality difference elimination module for the support and query set. Therefore, we will next take the process of the support set as an example. Then, we apply the channel connection of the second (l=1) and third-level (l=2) features and use a 3×3 convolutional layer to obtain feature maps F_{rgb}^{supp}, F_{th}^{supp}, respectively. A masked average pooling (MAP) [16] is used on F_{rgb}^{supp} and F_{th}^{supp} to obtain the representative RGB and thermal prototypes (V_{rgb}^s, V_{th}^s) in the support set. The i-th element v_i of V is calculated by the following equation:

$$v_i = \frac{\sum_{x,y}^{w,h} F_{x,y}^{\text{supp}} * [y_{x,y} = c]}{\sum_{x,y}^{w,h} [y_{x,y} = c]}, \tag{1}$$

where $y(c)$ indicates the ground mask, (x, y) indexes the position in the image, [*] is an indicator function. After converting two modalities to the same prototype space, we supervise the proximity of the two prototypes to each other by the following losses, which are carried out simultaneously in both directions.

$$L_{MD} = L_1(V_{rgb}^s, V_{th}^s) + L_1(V_{rgb}^q, V_{th}^q), \tag{2}$$

where $L_1(*)$ denotes the loss of Euclidean distance.

Since the true mask of the query set is not available, to implement the bi-directional modality difference elimination module of the query set, we construct a history bank that is continuously updated during the learning process. For each query image in both the training and validation phases, we get the corresponding prediction mask from the history bank according to the index and use it as the real mask.

B. RAFM

The existing bimodal FSS methods V-TFSS [22] and ADFNet [23] both interact or fuse immediately after extracting the bimodal features. They both belong to the shallow early fusion. And [22] has proved that early fusion has better performance than late fusion. Unlike the above methods mentioned, we find that deep early fusion [26] can achieve better results than shallow early fusion for dense comparison models, as compared in Table I. The architecture of the RAFM is shown in Fig. 2.

The RAFM consists of two components. The first is the feature interaction component and the second is the feature aggregation component. In the first component, we

TABLE I
QUANTITATIVE COMPARISON RESULTS OF DIFFERENT METHODS. "*" DENOTES OUR BASELINE

Methods	1-shot						5-shot					
	Fold-0	Fold-1	Fold-2	Fold-3	mIoU	FBIoU	Fold-0	Fold-1	Fold-2	Fold-3	mIoU	FBIoU
HSNet^{shallow}[20]	23.89	5.15	22.65	22.51	18.57	52.29	27.25	7.33	22.93	24.59	20.53	54.02
HSNet^{deep*}[20]	41.09	17.86	36.64	40.19	33.95	63.10	**48.15**	23.14	42.16	45.89	39.84	66.21
CANet [17]	27.59	9.36	28.53	33.78	24.82	57.01	31.97	10.58	30.06	35.29	26.98	57.24
PGNet [34]	24.87	6.11	29.88	39.37	25.15	57.45	31.18	5.98	29.26	43.34	27.44	58.85
PFENet [18]	32.22	6.25	26.76	33.64	24.71	57.85	34.08	13.93	27.07	41.55	29.15	60.20
ASGNet [19]	33.46	5.40	30.26	34.74	25.97	59.16	36.43	10.89	31.77	44.77	30.97	62.05
V-TFSS [22]	29.37	7.62	27.66	42.49	26.79	58.64	27.00	8.14	29.88	47.85	28.22	59.23
ASNet [36]	40.86	17.61	36.61	45.38	35.12	63.82	44.46	**24.04**	**44.67**	49.74	40.73	66.44
Ours	**45.17**	**20.75**	**42.43**	**47.26**	**38.90**	**65.29**	45.60	23.88	44.51	**50.33**	**41.08**	**66.52**

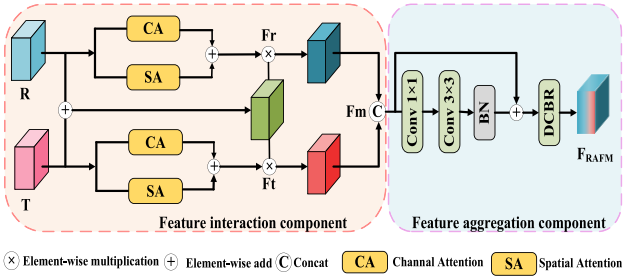

Fig. 2. The residual attention fusion module (RAFM).

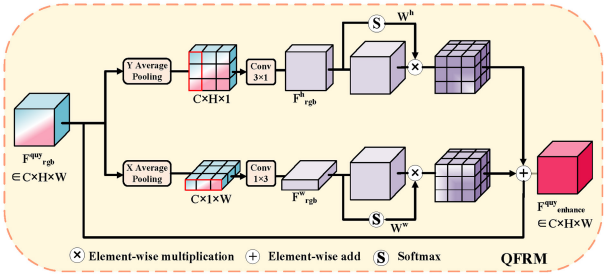

Fig. 3. The proposed query feature enhancement module (QFRM).

employ parallel channel attention (CA) and spatial attention (SA) [27] to highlight the respective uniqueness of the modalities. Instead of merging the multi-layer features extracted by the encoder at once, we first use the RGB and thermal modal features (R, T) that have undergone similarity computation and meta-learner, element sum the two, and then feed the RGB and thermal modal features into the first component separately to obtain the enhanced feature representation (F_r, F_t). Thereafter, the vectors in both branches are fused and reweighted by performing element-by-element multiplication, channel concatenation, and convolution.

$$F_r = SA(R) + CA(R). \tag{3}$$

$$F_t = SA(T) + CA(T). \tag{4}$$

$$F_m = Conv1 \times 1(Concat(F_r \otimes (R \oplus T), F_t \otimes (R \oplus T)). \tag{5}$$

In the feature aggregation component, we further fuse features effectively by residual learning to explore deeper complementary semantic relationships, as follows:

$$F'_m = BN(Conv3 \times 3(Conv1 \times 1(F_m))). \tag{6}$$

$$F_{RAFM} = DCBR(RELU(F'_m) + F'_m). \tag{7}$$

These features first reduce the number of channels through convolution blocks $Conv1 \times 1$, followed by a 3×3 convolution and nonlinear function ($ReLU$). To expand the receptive field and extract subtle local fusion features, the dilated convolution (k=3, r=1) [28] is followed by a batch normalization layer and $ReLU$ ($DCBR$) respectively.

C. MSEM

It can be divided into two independent sub-modules: the mainstay enhancement module (MEM) and the query feature enhancement module (QFRM). The architecture of the QFRM is shown in Fig. 3.

In particular, the MEM emphasizes the main part of the object by obtaining the global feature vector in two ways. 1)V^s_{rgb}. 2) Referring to PFENet [18], we generate a prior mask to better match the target region in the query image.

The QFRM concentrates on mining local information in the horizontal and vertical regions of the image. The shape of the general convolutional kernel is standard square, which greatly limits the learning ability of the model, since most of the targets in the road scenes have irregular shapes other than the main part, i.e., the subsidiary parts. DPCN [29] generates region-matching mappings using windows of different shapes. Different from [29], we propose to learn the contextual details of the subsidiary parts by reweighting the asymmetric convolution.

To reduce the effect between two directions and lower the computational effort simultaneously, we first apply the average pooling to change the height and width of the query features F^{quy}_{rgb} to 1.

$$F^w_{rgb}(b, c, 1, y) = \frac{1}{H} \sum_{x=0}^{H-1} F^{quy}_{rgb}(b, c, x, y). \tag{8}$$

$$F^h_{rgb}(b, c, x, 1) = \frac{1}{W} \sum_{y=0}^{W-1} F^{quy}_{rgb}(b, c, x, y). \tag{9}$$

Afterward, the initial shape is recovered by bilinear interpolation. A *softmax* layer is applied to compute the corresponding spatial weight map, denoted as W^w, $W^h \in R^{2\times H\times W}$. Finally, we add the original feature maps (F^w_{rgb}, F^h_{rgb}) according to weights (W^w, W^h) to obtain the fused feature map $F^{quy}_{enhance}$. The residual connection is used to facilitate network

training, denoted as:

$$(W^w||W^h) = softmax[Conv(F^w_{rgb}||F^h_{rgb})], \tag{10}$$

$$F^{quy}_{enhance} = W^w * F^w_{rgb} + W^h * F^h_{rgb} + F^{quy}_{rgb}, \tag{11}$$

where *soft* max is used to obtain the final weight vector. *Conv* indicates the convolution of the 1×3 or 3×1 shape. In this way, the QFRM achieves the role of establishing local correlations to enhance the subsidiary parts of the query features. Finally, we connect the outputs of the MEM and QFRM along the channel direction and then input them into the decoder to obtain the final segmentation mask $M_{final} \in R^{2\times Hq\times Wq}$.

D. Loss Function

The final prediction of BMDENet generates the second loss L_2. L_2 is the mean cross entropy loss (CE) between the prediction mask M_{final} and corresponding ground truth mask over all pixel locations.

$$L_2 = -\frac{1}{h\times w}\sum_{x=1}^{h}\sum_{y=1}^{w}\left(M_{final}\cdot log(y_q)\right). \tag{12}$$

The final total training loss L is the weighted sum of L_{MD} and L_2:

$$L = \alpha L_{MD} + \beta L_2, \tag{13}$$

where α, β are adjustable loss weights. The α and β in Eqn. (13) are set to 1.0 by default.

III. Experiments

A. Dataset and Implementation Details

To evaluate our method, we utilize the Tokyo Multi-Spectral-4^i dataset [22], which is part of the Tokyo Multi-Spectral images and annotations [30]. In total, 16 classes are included. The model is optimized using Adam [31] for 200 epochs. The learning rate is initialized as 0.00005 with batch size 2. For data augmentation, we use random horizontal flipping and scaling with scale ∈ {1, 1.5}. We set spatial size of both support and query images to 200×200. Our framework is constructed on PyTorch1.8 with an NVIDIA 3060. We employ mean intersection over union (mIoU) and foreground-background IoU (FBIoU) as the evaluation metrics, following prior FSS approaches [32], [33], [34], [35]. The dataset and code are available at: https://github.com/VDT-2048/BMDENet.

B. Comparative Experiments

***Quantitative Evaluation*:** In comparison experiments, we compare our BMDENet with the classical FSS methods under 1-shot and 5-shot settings. As shown in Table II, the deep early fusion method yields better results than shallow early fusion. This may be because for many-to-one frameworks like V-TFSS [22] and ADFNet [23], shallow early fusion can help them to generate better prototypes and get more accurate prospects when matching with query features later. On the contrary, for many-to-many models like HSNet [20] and ASNet [36], shallow early fusion destroys the dense correlation between the support set and the query set, and thus the information may be incomplete or erroneous when the meta-learner propagates it in a top-down manner later on. In addition, our model significantly outperforms previous methods on both in 1- and 5-shot setting, reaching the state-of-the-art on the dataset [22].

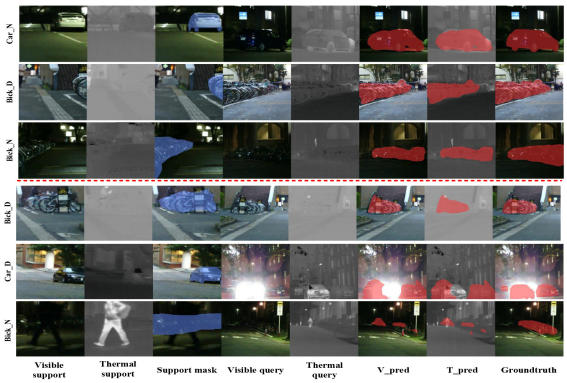

Fig. 4. Segmentation results on unseen classes under 1-shot setting.

TABLE II
Quantitative Comparison Results of Proposed Modules

Baseline	BMDEM	RAFM	MSEM	4^0	4^1	4^2	4^3	mIoU	FBIoU
√				41.09	17.86	36.64	40.19	**33.95**	63.10
√	√			43.70	18.79	35.46	43.83	**35.45**	63.52
√		√		43.42	18.83	36.05	45.05	**35.84**	63.87
√			√	40.47	19.47	41.84	46.84	**37.16**	64.45
√		√	√	41.62	17.23	43.52	47.46	**37.46**	64.54
√	√	√		41.74	20.05	41.14	44.61	**36.89**	64.52
√	√		√	41.07	17.49	42.20	46.87	**36.91**	64.23
√	√	√	√	45.17	20.75	42.43	47.26	**38.90**	65.29

Qualitative Evaluation: To exemplify the performance of the model, we report some qualitative results on the 1-shot setting. 1) As shown in Fig. 4, BMDENet does a good job of recognizing objects in the query image, even though the objects are in challenging scenes, such as low illumination, darkness, or complex backgrounds. This strongly demonstrates that BMDENet can well capture the complementary information of the two modalities and mitigate changes in object scale and position (e.g., bick). 2) Below the red line, we find that the model can recognize the target well when the light is sufficient, even if the T-image does not obviously capture the information; the opposite will fail. The model also fails when in a dark environment and there is no significant temperature difference between the object and the environment. Overall, BMDENet is still not sufficient to handle difficult samples.

C. Ablation Experiments

To demonstrate the effectiveness of the BMDENet, we perform ablation experiments on the 1-shot setting of the dataset [22]. There are three modules: BMDEM, RAFM, and MSEM. The ablation experiments help us to identify which module contributes significantly. The bimodal HSNet [20] with deep early fusion is the baseline. Comparing the third and sixth rows, we found that the fusion performance enhancement with only RAFM is not as high as the combined BMDEM and RAFM. This also suggests that BMDEM facilitates the full interaction and fusion of subsequent multimodal information

by reducing heterogeneity. Meanwhile, we obtain the best performance when combining BMDEM, RAFM, and MSEM, proving the effectiveness of all modules designed.

IV. Conclusion

In brief, we raise a bi-directional modality difference elimination network (BMDENet) for the RGB-T FSS task of road scenes. Specifically, we develop the BMDEM, which reduces differences by transforming visible and thermal features into the same space without modality interaction, and ensures specificity while reducing heterogeneity. To improve the robustness of the model to illumination changes, we introduce the RAFM to take full advantage of RGB and thermal properties for better information exchange and complementarity. In addition, the MSEM is designed to fill in the details and enhance the fused features using global representation and local correlation. Furthermore, we perform comprehensive experiments to demonstrate the advantages of our model. Our next step will consider how to weigh the relationship between the two modalities and add other sensors to further improve the results.

References

[1] K. Dang, C. Zhou, Z. Tu, M. Hoy, J. Dauwels, and J. Yuan, "Actor-action semantic segmentation with region masks," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2018, pp. 1–12.

[2] J. Wang, B. Ma, and K. Yan, "Mobile robot circumnavigating an unknown target using only range rate measurement," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 69, no. 2, pp. 509–513, Feb. 2022.

[3] G. Park, D. Im, D. Han, and H.-J. Yoo, "A 1.15 TOPS/W energy-efficient capsule network accelerator for real-time 3D point cloud segmentation in mobile environment," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 67, no. 9, pp. 1594–1598, Sep. 2020.

[4] R. Perfetti, E. Ricci, D. Casali, and G. Costantini, "Cellular neural networks with virtual template expansion for retinal vessel segmentation," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 54, no. 2, pp. 141–145, Feb. 2007.

[5] O. Ali, H. Ali, S. A. A. Shah, and A. Shahzad, "Implementation of a modified U-Net for medical image segmentation on edge devices," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 69, no. 11, pp. 4593–4597, Nov. 2022.

[6] C.-N. Chen and W.-T. Chu, "How it flies and why it flies? Volleyball trajectory segmentation and classification," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 68, no. 5, pp. 1591–1595, May 2021.

[7] Q. Chen et al., "A 67.5μJ/prediction accelerator for spiking neural networks in image segmentation," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 69, no. 2, pp. 574–578, Feb. 2022.

[8] L. He et al., "A multibit delta–sigma modulator with double noise-shaped segmentation," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 62, no. 3, pp. 241–245, Mar. 2015.

[9] Z. Tu et al., "Fusing disparate object signatures for salient object detection in video," *Pattern Recognit.*, vol. 72, pp. 285–299, Dec. 2017.

[10] K. Song, L. Huang, A. Gong, and Y. Yan, "Multiple graph affinity interactive network and a variable illumination dataset for RGBT image salient object detection," *IEEE Trans. Circuits Syst. Video Technol.*, early access, Dec. 29, 2022, doi: 10.1109/TCSVT.2022.3233131.

[11] K. Song, J. Wang, Y. Bao, L. Huang, and Y. Yan, "A novel visible-depth-thermal image dataset of salient object detection for robotic visual perception," *IEEE/ASME Trans. Mechatronics*, early access, Oct. 27, 2022, doi: 10.1109/TMECH.2022.3215909.

[12] W. Zhou, X. Lin, J. Lei, L. Yu, and J.-N. Hwang, "MFFENet: Multiscale feature fusion and enhancement network for RGB thermal urban road scene parsing," *IEEE Trans. Multimedia*, vol. 24, pp. 2526–2538, 2021, doi: 10.1109/TMM.2021.308618.

[13] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, 2010.

[14] T. Y. Lin et al., "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.

[15] A. Shaban, S. Bansal, Z. Liu, I. Essa, and B. Boots, "One-shot learning for semantic segmentation," 2017, *arXiv:1709.03410*.

[16] X. Zhang, Y. Wei, Y. Yang, and T. S. Huang, "SG-one: Similarity guidance network for one-shot semantic segmentation," *IEEE Trans. Cybern.*, vol. 50, no. 9, pp. 3855–3865, Sep. 2020.

[17] C. Zhang, G. Lin, F. Liu, R. Yao, and C. Shen, "CANet: Class-agnostic segmentation networks with iterative refinement and attentive few-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5217–5226.

[18] Z. Tian, H. Zhao, M. Shu, Z. Yang, R. Li, and J. Jia, "Prior guided feature enrichment network for few-shot segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 2, pp. 1050–1065, Feb. 2022.

[19] G. Li, V. Jampani, L. Sevilla-Lara, D. Sun, J. Kim, and J. Kim, "Adaptive prototype learning and allocation for few-shot segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 8334–8343.

[20] J. Min, D. Kang, and M. Cho, "Hypercorrelation squeeze for few-shot segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 6941–6952.

[21] Y. Zhang, D. Sidibé, O. Morel, and F. Meriaudeau, "Incorporating depth information into few-shot semantic segmentation," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, 2021, pp. 3582–3588.

[22] Y. Bao, K. Song, J. Wang, L. Huang, H. Dong, and Y. Yan, "Visible and thermal images fusion architecture for few-shot semantic segmentation," *J. Vis. Commun. Image Rep.*, vol. 80, Oct. 2021, Art. no. 103306.

[23] C. Zhang, J. Jiao, W. Xu, N. Li, M. Pang, and J. Dong, "ADFNet: Attention-based fusion network for few-shot RGB-D semantic segmentation," in *Proc. 14th Int. Conf. Mach. Learn. Comput. (ICMLC)*, 2022, pp. 91–96.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[25] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.

[26] Y. Li, J. Zhang, Y. Cheng, K. Huang, and T. Tan, "Semantics-guided multi-level RGB-D feature fusion for indoor semantic segmentation," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, 2017, pp. 1262–1266.

[27] J. Fu et al., "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3146–3154.

[28] F. Yu, and V. Koltun, "Multi-scale context aggregation by dilated convolutions," 2015, *arXiv:1511.07122*.

[29] J. Liu, Y. Bao, G. S. Xie, H. Xiong, J. J. Sonke, and E. Gavves, "Dynamic prototype convolution network for few-shot semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 11553–11562.

[30] Q. Ha, K. Watanabe, T. Karasawa, Y. Ushiku, and T. Harada, "MFNet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, 2017, pp. 5108–5115.

[31] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[32] Y. Liu, X. Zhang, S. Zhang, and X. He, "Part-aware prototype network for few-shot semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 142–158.

[33] G. Shi, Y. Wu, S. Palaiahnakote, U. Pal, and T. Lu, "ARNet: Active-reference network for few-shot image semantic segmentation," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, 2021, pp. 1–6.

[34] C. Zhang, G. Lin, F. Liu, J. Guo, Q. Wu, and R. Yao, "Pyramid graph networks with connection attentions for region-based one-shot semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 9587–9595.

[35] G. S. Xie, J. Liu, H. Xiong, and L. Shao, "Scaleaware graph neural network for few-shot semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 5475–5484.

[36] D. Kang, and M. Cho, "Integrative few-shot learning for classification and segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 9979–9990.

[37] Y. Li, G. W. P. Data, Y. Fu, Y. Hu, and V. A. Prisacariu, "Few-shot semantic segmentation with self-supervision from pseudo-classes," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2021, pp. 1–13.

[38] T. Hu, P. Yang, C. Zhang, G. Yu, Y. Mu, and C. G. Snoek, "Attention-based multi-context guiding for few-shot semantic segmentation," in *Proc. AAAI*, vol. 33, 2019, pp. 8441–8448.