Bidirectional Alternating Fusion Network for RGB-T Salient Object Detection

Zhengzheng Tu, Danying Lin, Bo Jiang, Le Gu, Kunpeng Wang, and Sulan Zhai(✉)

Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, Anhui University, Hefei 230601, China

zhengzhengahu@163.com, danying_lin@foxmail.com, jiangbo@ahu.edu.cn, awngule@foxmail.com, kp.wang@foxmail.com, 01044@ahu.edu.cn

Abstract. RGB-Thermal Salient Object Detection(SOD) aims to identify common salient regions or objects from both the visible and thermal infrared modalities. Existing methods usually based on the hierarchical interactions within the same modality or between different modalities at the same level. However, this approach may lead to a situation where one modality or one level of features dominates the fusion result during the fusion process, failing to fully utilize the complementary information of the two modalities. Additionally, these methods usually overlooking the potential for the network to extract specific information in each modality. To address these issues, we propose a Bidirectional Alternating Fusion Network (BAFNet) consisting of three modules for RGB-T salient object detection. In particular, we design a Global Information Enhancement Module(GIEM) for improving the information representation of high-level features. Then we propose a novel bidirectional alternating fusion strategy which is applied during decoding, and we design a Multi-modal Multi-level Fusion Module(MMFM) for collaborating mulit-modal mulit-level information. Furthermore, we embed the proposed Modal Erase Module (MEM) into both GIEM and MMFM to extract the inherent specific information in each modality. Our extensive experiments on three public benchmark datasets show that our method achieves outstanding performance compared to state-of-the-art methods.

Keywords: salient object detection · visible and thermal infrared · alternating fusion · specific information

This work was supported by the University Synergy Innovation Program of Anhui Province under Grant No.GXXT-2022-014, in part by National Natural Science Foundation of China under Grant No.62376005.

Z. Lin et al. (Eds.): PRCV 2024, LNCS 15038, pp. 34–48, 2025.
https://doi.org/10.1007/978-981-97-8685-5_3

1 Introduction

RGB-Thermal salient object detection (SOD) aims to accurately identify and highlight visually salient objects or regions by fusing the visible and thermal infrared information. Visible light image contains rich color and texture information, which provides better performance in identifying common objects. Thermal infrared image reflects the heat distribution of objects and has better insight in situations such as low light, night time or obstacle obstruction.

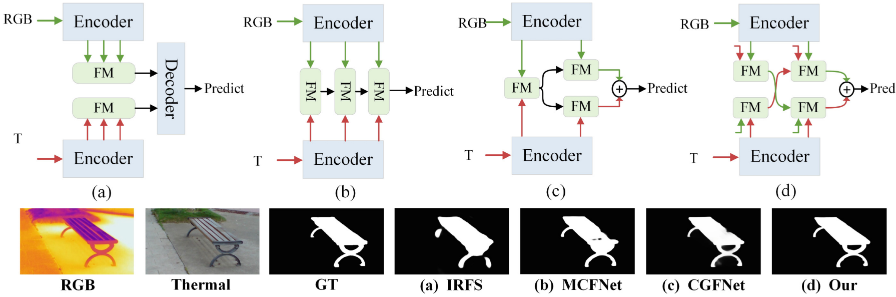

Fig. 1. Existing methods with different modal fusion strategies and 'FM' means Fusion Module. (a)The feature fusion strategy of different levels in same modality. (b)The feature fusion strategy of different modalities in same level. (c)The feature fusion strategy of combining (a) and (b). (d) Our fusion strategy fuses features through alternating collaborations across different modalities and levels.

SOD has found extensive applications in various computer vision tasks, including semantic segmentation [1–4], RGB-T tracking [5,6], and object recognition [7]. Existing multi-modal methods [8–10] can accurately detect salient objects by continuously improving fusion strategies, which can be divided into three types, as shown in Fig. 1. The first feature fusion strategy (illustrated in Fig. 1(a)) [11,12] incorporates features extracted from the same modality at different levels to obtain the multi-scale features, and then feeds the fused features from the two modalities into the decoder to generate the saliency prediction map. The second strategy (illustrated in Fig. 1 (b)) [14–16], the most popular multi-modal fusion approach, gradually fuses features from two modalities at the same level via top-down decoding. The third strategy (illustrated in Fig. 1 (c)) [18,19] combines both (a) and (b) fusion methods, fusing different modalities at the same level and handling multi-level features in a bidirectional manner. Although the fusion strategies mentioned above have greatly boosted salient object detection performance, they still suffer from two key issues. The first issue is as shown in IRFS [12] and MCFNet [16] of Fig. 1, in the two fusion strategies, one of the modalities or one of the feature levels will dominate the results. The former increases modality variability, failing to leverage complementary information between modalities. The latter leads to bias between global context and local information, increasing noise interference. The second issue is as shown in CGFNet [19] of Fig. 1, this strategy equally delivers the fusion of multi-modal features of the higher level to the single-modal features of the lower level. Although it can establish interactions between multi-modalities and multi-levels, the detailed information loss due to ignoring the low-level information of the other modality during low-level multi-modal feature fusion. Additionally, existing fusion strategies mostly focus solely on available modal information, neglecting specific information the network can mine. Based on the above problems, we propose a novel fusion strategy (illustrated in Fig. 1(d)). For the initial fusion module(illustrated in Fig. 1(FM)), we input multi-modal features of corresponding sizes, and the obtained fusion features are alternately sent to the lower-level fusion module. In this way, we allow each modality to alternatively dominate the fusion result, ensuring that we can leverage the information from each modality to maintain inter-modality balance during the fusion process. In this paper, we propose a Bidirectional Alternating Fusion Network(BAFNet) for RGB-T salient object detection. Specifically, we design a new alternative fusion strategy and deploy it into the decoding process, which aims to address the imbalance of modal information due to the under-utilization of auxiliary modalities. Notably, we design a Global Information Enhancement Module (GIEM) to localize salient objects by leveraging the global and semantic information of high-level features. Then, we design a Multi-modal Multi-level Fusion Module(MMFM) and embed it into the alternating decoding process to achieve mutual interaction of multi-modal as well as multi-level features. In addition, we design the Modal Erasure Module (MEM), inspired by TCLNet [21] to replace the frame information with modal information against the existing methods that only focus on the acquired modal information. To fully use the specific information from different modalities and different levels, we embed MEM into both GIEM and MMFM to generate the modality-specific information, which will be served as the auxiliary modal information to improve the performance of the network.

The key contributions of our BAFNet can be summarized as follows:

- To fully exploit the complementary information between the two modalities, we propose a RGB-T salient object detection method with a new multi-modal fusion strategy, called a Bidirectional Alternating Fusion Network(BAFNet).
- We devise a Global Information Enhancement Module(GIEM) to mine the ability of high-level features for global representation of salient regions. And we design a Multi-modal Multi-level Fusion Module(MMFM) and embed it into each level of features to realize multi-modal and multi-level interactive learning. Additionally, to extract the inherent specific information in each modality, we embed our proposed Modal Erase Module (MEM) into both GIEM and MMFM.
- Extensive experiments on three public RGB-T SOD datasets demonstrate the outstanding performance of our method in comparison to state-of-the-art RGB-T SOD methods.

2 Related Work

2.1 Salient Object Detection

In recent years, CNN-based methods extract multi-level features to produce saliency maps and achieve good prediction results. Zhao *et al.* [22] design a basic gated network that allows available information from the encoder to be effectively transmitted to the decoder. Tian *et al.* [23] investigate distributional uncertainty in salient object detection and explore class-aware distribution gap to effectively model the discrepancy between training and testing distributions. Pang *et al.* [24] integrate features from adjacent levels by proposing the aggregation interaction module to avoid introducing too much noise. Wu *et al.* [25] certify the existence of a point-labeled dataset, and saliency models trained on this dataset can achieve equivalent performance compared with those trained on the densely annotated dataset. Liu *et al.* [26] investigate the effectiveness of pooling techniques in salient object detection, with a particular focus on their applicability to mobile devices. Ma *et al.* [27] design a framework to effectively address the challenge of scale variation in salient object detection.

2.2 RGB-T Salient Object Detection

Thermal infrared imaging technology operates by detecting the thermal radiation emitted by objects, allowing it to identify their location even in challenging conditions such as nighttime, low-light environments, and areas with smoke or haze *etc*. By utilizing the information from visible and thermal infrared images, the researchers propose RGB-T SOD, and achieve amazing performance beyond RGB SOD. Early RGB-T SOD works [28,29] exploit the complementary cues between RGB and thermal images for SOD. Lee *et al.* [13] design a feature fusion module to provide rich contextual information for the network by fusing the multi-scale encoder features of two modalities respectively in a parallel manner. Zhou *et al.* [17] use channel attention and spatial attention to build a cross-modality fusion module, fusing features of matching sizes from both modalities. They then employ a multi-level consistent module to integrate complementary information from various levels. Tu *et al.* [20] design a new dual-decoder to model the interaction among two modalities, multi-level features, and global context for learning three complementary information. However, most of the fusion strategies in the above methods fail to adequately establish interactive learning between multi-modality and multi-level. In addition, they overlook the extraction and utilization of specific information in each modality.

3 The Proposed Approach

3.1 Overview

As shown in Fig. 2 (A), we propose a Bidirectional Alternating Fusion Network(BAFNet) which employs a bidirectional alternating fusion approach in the

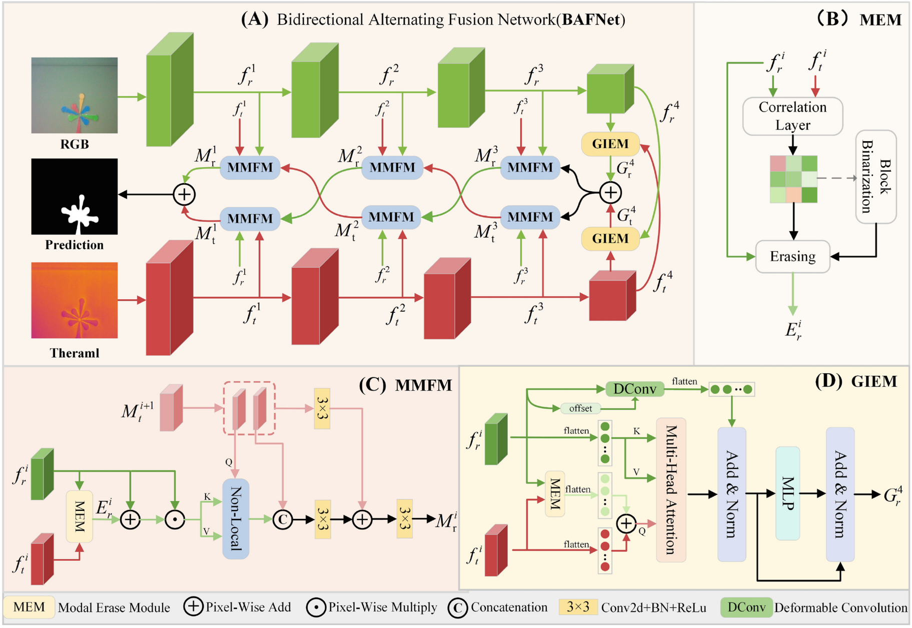

Fig. 2. The architecture of our proposed BAFNet comprises several modules: the Global Information Enhancement Module(GIEM), which is responsible for enhancing both semantic and positional information extracted from salient regions within high-level features. And the Multi-modal Multi-level Fusion Module(MMFM) for modeling interactions between multi-modal as well as multi-level features. Furthermore, the Modal Erase Module (MEM) is incorporated to effectively mine the specific information in each modality.

decoding stage to accurately generate saliency maps. In particular, we utilize SwinB [45] as the backbone network to extract multi-level features from visible and thermal infrared images denoted as $\{f_r^i\}_{i=1}^4$ and $\{f_t^i\}_{i=1}^4$, respectively.

We design a Global Information Enhancement Module(GIEM) comprising multi-head attention [32] and deformable convolutions [31] to capture the dependency between global features and enhance the sensitivity of the network to salient object locations. Next, we devise a Multi-modal Multi-level Fusion Module(MMFM) and embed it into the decoding process in a cascaded manner, integrating both multi-modal and multi-level features to learn their interaction. Additionally, we propose a Modal Erase Module(MEM) and insert it into both GIEM and MMFM to generate the modality-specific information as auxiliary information. Further elaboration on these modules will be provided in the subsequent sections.

3.2 Modal Erase Module

Existing methods primarily concentrate on fusing available modal information, overlooking the potential for the network to extract specific information in each modality. Hou *et al.* [21] devised a temporal saliency module to encourage the network to mine new information from the current frame by erasing salient information from previous frames within the current frame. Inspired by Hou [21], we introduce a Modal Erase Module (MEM), as depicted in Fig. 2 (B). Specifically, we erase information information from one modality, encouraging the network to extract the specific information within that modality for salient object detection. At first, we calculate the high corresponding region of one modality in the other modality, where w denotes the feature mapping space and R^i denotes the correlation matrix of the two modalities. The calculation follows the procedure shown below:

$$R_1^i = (f_r^i)^T(w^T f_t^i), \quad R_2^i = (f_t^i)^T(w^T f_r^i), \tag{1}$$

and the we binary the correlation matrix R^i to generate the binary mask M to identify the regions which will be erased, and subsequently utilize the $softmax$ activation function to filter out the erased regions from it, resulting in a new matrix denoted as G^i. In this way G^i and M^i can denote the regions to be kept and the regions to be erased , respectively. Following [21] based on G^i and M^i, we can obtain the erased features $\left\{E_r^i\right\}_{i=1}^4$ and $\left\{E_t^i\right\}_{i=1}^4$ as modality-specific information. The calculation follows the procedure shown below:

$$G_1^i = softmax(R_1^i) \odot Binary(R_1^i), \quad G_2^i = softmax(R_2^i) \odot Binary(R_2^i), \tag{2}$$

$$\left\{E_r^i\right\} = Conv(f_r^i \times G_1^i) + f_r^i \odot Binary(R_1^i), \tag{3}$$

$$\left\{E_t^i\right\} = Conv(f_t^i \times G_2^i) + f_t^i \odot Binary(R_2^i), \tag{4}$$

where $Conv$ denotes the convolution operation with 3×3 convolution kernel, × denotes the matrix multiplication, $\odot$ represents pixel-level multiplication, and i takes the value from 1,2,3,4.

3.3 Global Information Enhancement Module

The high-level features contain rich semantic information which can be used to localize semantic content such as location, shape, *etc.* Therefore, we design a Global Information Enhancement Module (GIEM) to enhance the ability to understand global features, as illustrated in Fig. 2 (D). Specifically, we embed the MEM into GIEM to obtain the modality-specific information $E_m^4 (m \in [r, t])$, serving as additional information. Subsequently, we perform summation fusion with $f_n^4 (n \in [t, r])$ to obtain the enriched mulit-modal fusion feature denoted as f_{tr}^4. For fully utilizing the complementary information of different modalities, we feed f_{tr}^4 as $Q \in R^{HW \times C}$, and f_m^4 as $K \in R^{HW \times C}$ and $V \in R^{HW \times C}$along with the multi-head attention. In addition, with the purpose of improving the sensitivity of the network to the position of salient objects. We introduce learnable offsets into deformable convolutions to adjust the shape and position of the convolution kernel, enabling better adaptation to object shapes and positions. This process is detailed in the following equation:

$$MSA(Q,K,V) = Softmax(QK^T/\sqrt{d_k})V, \tag{5}$$

$$M_g = MSA(f_n^4 + MEM(f_m^4, f_n^4), f_m^4, f_m^4), \tag{6}$$

$$G_m = LN(MLP(LN(M_g + DConv(f_m^4))) + M_g), \tag{7}$$

where MSA denotes the multi-head attention, LN is the layer of normalization, MLP is the Multi-Mayer Perceptron [33], and $DConv$ represents the deformable convolutions.

3.4 Multi-modal Multi-level Fusion Module

We propose a novel fusion strategy by alternately allowing different modalities to take a dominant role, thereby preventing one modality information from being overlooked by the other modality or level due to modalities imbalance during the decoding process. Additionally, we introduce a Multi-modal Multi-level Fusion Module(MMFM) embedded into the decoding process, which fully models the relationships between different modalities and scale features. The detailed framework of the module as illustrated in Fig. 2 (C). In particular, we embed the MEM into the MMFM to generate the modality-specific information $E_m^i(m \in [r,t])$, and directly add E_m^i with f_m^i, then multiply it with f_m^i to extract the common salient regions denoted as $\tilde{f_m}^i$. Besides, we achieve interactive learning between different modalities and levels by alternately allowing the higher-level features of the other modality to guide the lower-level features for fusion. Specifically, the high-level features of the other modality $M_n^{i+1}(n \in [t,r])$ are split by channel dimension and 2× up-sampled, ensuring that the channels and resolution scales of the high-level features which is consistent with those of the low-level features. The obtained features are denoted as M_1^i and M_2^i, respectively.

$$\tilde{f_m^i} = (MEM(f_m^i, f_n^i) + f_m^i) \cdot f_m^i, \tag{8}$$

$$[M_1, M_2] = Split(UP(M_n^{i+1})), \tag{9}$$

where MEM denotes Modal Erase Module, $Split$ indicates the division operation and UP denotes a 2× up-sample operation.

On the one side, with the aim of leveraging the information of the high-level features with different channel dimensions, M_1^i is sent as Q and f_r^i as K and V to the Non-Local [46] attention mechanism, followed by the collocation of the channel dimensions with M_2^i. On the other side, for obtaining the complete information of the high-level features, we add the high-level features after convolution with the current features to obtain the final fused feature denoted as $M_m^i(m \in [r,t])$. The calculation process can be demonstrated as follows:

$$N(Q,K,V) = Softmax(Q^TK)V + V, \tag{10}$$

$$M_m^i = C([N(M_1, f_m^{i+1}), f_m^{i+1}, M_2]) + C(UP(M_n^{i+1})), \tag{11}$$

where N denotes the attention operation as shown in $Eq.11$, C represents the convolution operation with 3×3 convolution kernel, followed by a batch normalization and $ReLu$ activation function [34], $[*]$ denotes the concatenation operation of the channel dimensions, and i takes the value of $i = \{3, 2, 1\}$.

Finally, we directly add up the alternating fusion results obtained from MMFM to obtain the final feature for predicting the saliency map.

3.5 Loss Function

In this paper, we denote the saliency prediction map and the ground-truth as P and G, respectively. Following [16, 20], we introduce the binary cross-entropy loss function [35], the smoothness loss function, and the dice function [36] for optimizing the overall parameters of the network during the training process to generate accurately saliency prediction maps. The total loss functions are as follows:

$$L = L_{bce}(P, G) + L_{smooth}(P, P) + L_{dice}(P, G). \tag{12}$$

4 Experiments

4.1 Experimental Setup

Dataset: In our approach, we utilize the RGB-T SOD dataset, consisting of 2500 pairs of images sourced from VT5000-Train [30], as the training set. Subsequently, we employ three public datasets including VT5000-Test, VT1000 [28], and VT821 [37], respectively, as the testing set to comprehensively evaluate the effectiveness of our method.

Implementation Details: Our method is implemented using PyTorch and trained on a single GeForce RTX 3090 GPU with a batch size of 4 for 100 epochs. We set the learning rate to 1e-5 initially, and adjust it to 1e-4 after 40 epochs. Both during training and testing phases, we resize all images into 384×384.

Evaluation Metrics: We employ three quantitative metrics which are widely used in RGB-T SOD: the weighted F-measure (wF), the Mean Absolute Error (MAE), and the S-measure (Sm) to evaluate the performance of our method.

4.2 Comparison Experiments:

We compare the performance of our proposed BAFNet with 12 SOTA methods on three datasets. The following methods for comparison are included: ADF [30], MIDD [20], CSRNet [14], CGFNet [19], SwinNet [38], OSRNet [39], TNet [40], DCNet [41], LSNet [42], HRTransNet [43], MCFNet [16], and CAVER [44].

Quantitative Comparison: As depicted in Table 1, our proposed BAFNet outperforms existing SOTA methods across the three public datasets, achieving superior performance in all three evaluation metrics. We denote optimal performance in red and sub-optimal performance in blue. In Table 1, it can be observed that our method shows comparable improvement over existing SOTA methods on three datasets, especially in the wF metrics where the improvement is most significant. In particular, comparing to the sub-optimal HRTransNet, our method achieves higher performance on average by 0.6%, 1.4% and 14.4% respectively, across the three public datasets. In addition, in comparison to SwinNet, a method that also utilizes SwinB as the backbone network, we achieve better performance by 0.6%, 3.9%, and 23.9%, respectively.

Table 1. COMPARISON OF PERFORMANCE WITH 12 METHODS ON THREE COMMONLY USED TEST DATASETS, WHICH WE COMPARED QUANTITATIVELY USING THREE METRICS. THE BEST RESULTS ARE EMPHASIZED IN RED AND SUB-OPTIMAL ARE INDICATED IN BLUE FONT.

Methods	VT821			V1000			V5000		
	Sm↑	wF↑	MAE↓	Sm↑	wF↑	MAE↓	Sm↑	wF↑	MAE↓
ADF(20)	0.810	0.627	0.716	0.910	0.804	0.034	0.864	0.722	0.048
MIDD(21)	0.871	0.760	0.045	0.915	0.856	0.027	0.868	0.763	0.043
CSRNet(21)	0.884	0.821	0.038	0.918	0.878	0.024	0.868	0.796	0.042
CGFNet(21)	0.881	0.829	0.038	0.923	0.900	0.023	0.883	0.831	0.035
SwinNet(21)	0.904	0.818	0.030	0.938	0.894	0.018	0.912	0.846	0.026
OSRNet(22)	0.875	0.801	0.043	0.926	0.891	0.022	0.875	0.807	0.040
TNet(22)	0.899	0.841	0.030	0.929	0.895	0.021	0.895	0.840	0.033
DCNet(22)	0.877	0.822	0.033	0.923	0.902	0.021	0.871	0.819	0.035
LSNet(23)	0.878	0.809	0.033	0.925	0.887	0.023	0.872	0.819	0.035
HRTransNet(23)	0.906	0.849	0.026	0.938	0.913	0.017	0.912	0.870	0.025
MCFNet(23)	0.891	0.835	0.029	0.932	0.906	0.019	0.887	0.836	0.033
CAVER(23)	0.891	0.835	0.033	0.936	0.909	0.017	0.892	0.835	0.032
Ours	0.911	0.866	0.024	0.944	0.925	0.014	0.919	0.884	0.022

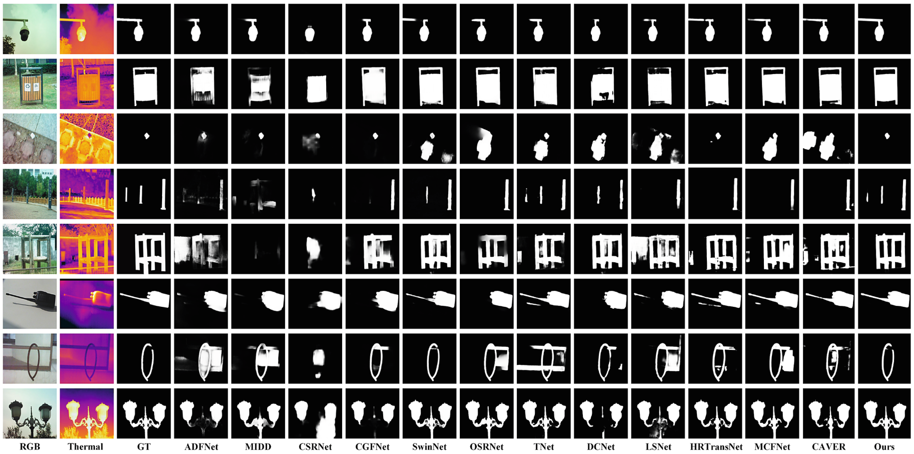

Fig. 3. Visual comparison of our method with 12 state-of-the-art methods on 8 challenging samples.

Vision Comparison: In Fig. 3, we select 8 challenging samples for visual comparison against existing methods. From the 2nd and 5th rows of the figure, we can observe that the complex structure of the salient objects poses a significant challenge for the methods. Not only does our BAFNet accurately localize the object region matter,but it also extracts the clear boundary contour. Existing methods seem to struggle with these complexly constructed objects, while our method can produce precise saliency maps. It suggests that our method effectively leverages the high-level fusion features from one modality to guide the low-level detailed information of the other modality, enabling accurate localization of salient objects. Additionally, detailed contour information is supplemented through alternating decoding processes.

Table 2. OUR ABLATION EXPERIMENTS ON THREE MODULES ON VT821, VT1000 AND VT5000. THE BEST RESULTS ARE SHOWN IN **BOLD**.

Components	VT821			VT1000			VT5000		
	Sm↑	wF↑	MAE↓	Sm↑	wF↑	MAE↓	Sm↑	wF↑	MAE↓
w/o Strategy	0.906	0.850	0.026	0.942	0.921	0.015	0.913	0.873	0.024
w/o MEM	0.900	0.842	0.028	0.939	0.914	0.015	0.909	0.867	0.024
w/o MMFM	0.905	0.849	0.028	0.941	0.920	0.015	0.910	0.872	0.024
w/o GIEM	0.909	0.856	0.025	0.942	0.921	0.015	0.914	0.878	0.023
Ous	**0.911**	**0.866**	**0.024**	**0.944**	**0.925**	**0.014**	**0.919**	**0.884**	**0.022**

4.3 Ablation Study:

Effectiveness of Alternating Fusion Strategy: To assess the effectiveness of each module in our method, we conduct ablation experiments as detailed in Table 2. We replace the alternating fusion strategy with the sequential fusion strategy to verify the effectiveness of our proposed feature fusion strategy. Comparing the results in the 1st and 5th rows of the table, we observed an average drop of 0.4%,1.7% and 8.1% in performance, respectively. It demonstrates that we can leverage the complementary information between the two modalities by alternatively allowing the higher-level features of each modality to guide the lower-level features of the other modality.

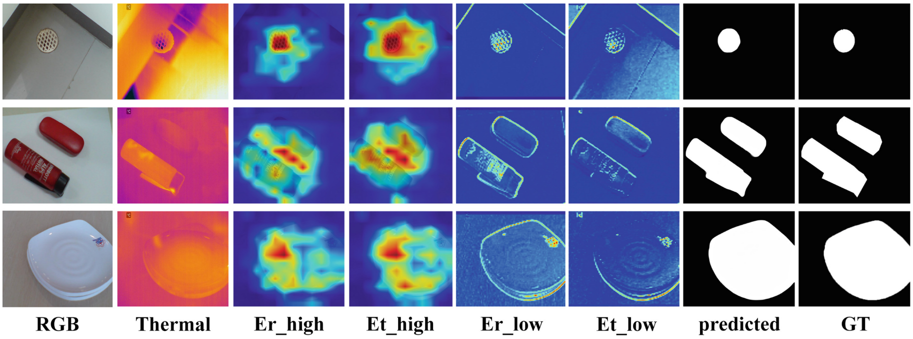

Fig. 4. Feature visualization of both high level and low level features within the Modal Erase Module(MEM)

Effectiveness of MEM: Comparing the 2nd and 5th rows of the table, we confirm that encouraging the network to mine the modality-specific information can help to improve the detection performance. It can be seen that the removal of MEM dramatically drops the performance by 0.8%, 1.7% and 10.9%. In addition, we also visualize the characteristics of the MEM to see what auxiliary information it provide for the network, as shown in Fig. 4. Where the 3rd and 4th columns indicate the high-level erased RGB modal and Thermal modal features, respectively, it can be seen that the high-level features with richer semantic information which can be used to locate the salient regions. And the 5th and 6th columns represent the low-level modal features, respectively, which can be used to introduce the boundary information of the salient regions. Finally, we can see that our prediction maps can accurately segment the salient objects, which demonstrates that by introducing additional modal information, we are able to improve the performance of the network.

Effectiveness of MMFM: Comparing the 3rd and 5th rows of the table, we remove the MMFM which results in a performance drop of 0.6%, 1.1%, and

10.9%, respectively. It indicates that enabling interactive learning between different modalities and levels via one modality of higher-level features to guide the fusion of lower-level features contributes to improving performance.

Effectiveness of GIEM: By directly summing the fusion of the last level of features to replace GIEM for verifying its performance, and comparing the 4th and 5th rows of the table, we can observe a performance drop of 0.3%, 0.6% and 6.8% respectively. This indicates that introducing multi-head attention and deformable convolutions allows the network to model the dependency of different positions in the two modalities.

5 Conclusion

In this paper, we propose a Bidirectional Alternating Fusion Network(BAFNet) for RGB-T salient object detection. In particular, we design a Global Information Enhancement Module(GIEM) to enhance both semantic and positional information extracted from salient regions within high-level features. And then we design a fusion strategy by continuous alternating decoding, and embed a Multi-modal Multi-level Fusion Module(MMFM) into decoding process to collaborate mulitmodal mulit-level information. Furthermore, the Modal Erase Module(MEM) is incorporated to effectively mine the specific information in each modality. Extensive experimental results demonstrate the superior performance of our method compared to state-of-the-art methods.

References

1. Yang, E., Zhou, W., Qian X.: MGCNet: multilevel gated collaborative network for RGB-D semantic segmentation of indoor scene. IEEE Signal Process. Lett. **29**, 2567–2571 (2022)
2. Xu, J., Xiong, Z.: PIDNet: a real-time semantic segmentation network inspired by PID controllers. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 19529–19539
3. Ying, X., Chuah., M.C.: UCTNet: uncertainty-aware cross-modal transformer network for indoor RGB-D semantic segmentation. In: European Conference on Computer Vision, vol. 13690. Springer, Heidelberg (2022). ISBN:978-3-031-20055-7
4. Xinyi, W., Yuan, X.: RGB-D road segmentation based on geometric prior information. In: Pattern Recognition and Computer Vision: 6th Chinese Conference, PRCV,: Xiamen, China, 13–15 Oct 2023, Proceedings, Part I. Springer, Heidelberg, pp. 434–445 (2023). https://doi.org/10.1007/978-981-99-8429-935
5. Xiao, Y., Yang, M.: Attribute-based progressive fusion network for RGBT tracking. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, no. 3, pp. 2831–2838. https://doi.org/10.1609/aaai.v36i3.20187
6. Tang, Z., Xu, T.: Exploring fusion strategies for accurate RGBT visual object tracking. Inf. Fusion **99**, 101881 (2023). ISSN:1566-2535
7. Loghmani, M.R., Robbiano, L.: Unsupervised domain adaptation through intermodal rotation for RGB-D object recognition. IEEE Robot. Autom. Lett. **5**(4), 6631–6638 (2020). Oct

8. Song, Z., Qin, P.: EdgeFusion: infrared and visible image fusion algorithm in low light. In: Pattern Recognition and Computer Vision: 6th Chinese Conference, PRCV,: Xiamen, China, 13–15 Oct 2023, Proceedings, pp. 259–270. Part I. Springer-Verlag, Berlin, Heidelberg (2023)
9. Jiang, S., Xu, Y.: Multi-scale fusion for RGB-D indoor semantic segmentation. Sci. Rep. **20305**, 2045–2322 (2022)
10. Zhang, T., Li, H.: MGT: modality-guided transformer for infrared and visible image fusion. In: Pattern Recognition and Computer Vision: 6th Chinese Conference, PRCV,: Xiamen, China, 13–15 Oct 2023, Proceedings, Part I. Springer, Heidelberg, pp. 321–332 (2023). https://doi.org/10.1007/978-981-99-8429-926
11. Wang, C., Xu, C.: Cross-modal pattern-propagation for RGB-T tracking. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, pp. 7062–7071 (2020)
12. Wang, D., Liu, J.: An interactively reinforced paradigm for joint infrared-visible image fusion and saliency object detection. Inf. Fusion **98**, 101828 (2023). (Elsevier)
13. Lee, M., Park, C.: SPSN: superpixel prototype sampling network for rgb-d salient object detection. In: Computer Vision-ECCV: 17th European Conference, Tel Aviv, Israel, 23–27 Oct 2022, Proceedings, pp. 630–647. Part XXIX. Springer-Verlag, Berlin, Heidelberg (2022)
14. Fushuo, H., Xuegui, Z.: Efficient context-guided stacked refinement network for RGB-T salient object detection. IEEE Trans. Circuits Syst. Video Technol. **32**(5), 3111–3124 (2022). May
15. Wang, Y., Dong, F.: Interactive context-aware network for RGB-T salient object detection. Multimed. Tools Appl. 1–22 (2024). (Springer)
16. Ma, S., Song, K.: Modal complementary fusion network for RGB-T salient object detection. Appl. Intell. **53**(8), 9038–9055 (2023). (Springer)
17. Wujie, Z., Qinling, G.: ECFFNet: effective and consistent feature fusion network for RGB-T salient object detection. IEEE Trans. Circuits Syst. Video Technol. **32**(3), 1224–1235 (2022). March
18. Guibiao, L., Wei, G.: Cross-collaborative fusion-encoder network for robust RGB-thermal salient object detection. IEEE Trans. Circuits Syst. Video Technol. **32**(11), 7646–7661 (2022). Nov.
19. Wang, J., Song, K.: CGFNet: cross-guided fusion network for RGB-T salient object detection. IEEE Trans. Circuits Syst. Video Technol. **32**(5), 2949–2961 (2022). May
20. Tu, Z., Li, Z.: Multi-interactive dual-decoder for RGB-thermal salient object detection. IEEE Trans. Image Process. **30**, 5678–5691 (2021). https://doi.org/10.1109/TIP.2021.3087412
21. Hou, R., Chang, H.: Temporal complementary learning for video person re-identification. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 Aug 2020, Proceedings, Part XXV 16, pp. 388–405. Springer (2020)
22. Zhao, X., Pang, Y.: Suppress and balance: a simple gated network for salient object detection. In: Computer Vision-ECCV,: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, pp. 35–51. Part II. Springer-Verlag, Berlin, Heidelberg (2020)
23. Tian, X., Zhang, J.: Modeling the distributional uncertainty for salient object detection models. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 2023, pp. 19660–19670 (2023)
24. Pang, Y., Zhao, X.: Multi-scale interactive network for salient object detection. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2020, pp. 9410–9419 (2020)

25. Wu, Z., Wang, L.: Pixel is all you need: adversarial trajectory-ensemble active learning for salient object detection. In: AAAI Conference on Artificial Intelligence, vol. 37, no. 3, pp. 2883–2891 (2023)
26. Liu, J.-J., Hou, Q.: PoolNet+: exploring the potential of pooling for salient object detection. IEEE Trans. Pattern Anal. Mach. Intell. **45**(1), 887–904 (1 Jan 2023)
27. Ma, M., Xia, C.: Boosting broader receptive fields for salient object detection. IEEE Trans. Image Process. **32**, 1026–1038 (2023). https://doi.org/10.1109/TIP.2022.3232209
28. Tu, Z., Xia, T.: RGB-T image saliency detection via collaborative graph learning. IEEE Trans. Multimed. **22**(1), 160–173 (Jan 2020). https://doi.org/10.1109/TMM.2019.2924578
29. Gao, W., Liao, G.: Unified information fusion network for multi-modal RGB-D and RGB-T salient object detection. IEEE Trans. Circuits Syst. Video Technol. 32(4), 2091–2106 (April 2022)
30. Tu, Z., Ma, Y.: RGBT salient object detection: a large-scale dataset and benchmark. IEEE Trans. Multimed. **25**, pp. 4163–4176 (2020). https://doi.org/10.1109/TMM.2022.3171688
31. Dai, J., Qi, H.: Deformable convolutional networks. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 764–773
32. Vaswani, A., Shazeer, N.: Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17). Curran Associates Inc., Red Hook, NY, USA, pp. 6000–6010
33. Tolstikhin, I.O., Houlsby, N.: Mlp-mixer: an all-mlp architecture for vision. Adv. Neural Inf. Process. Syst. **34**, 24261–24272 (2021)
34. Nair, V., Hinton, G.E.: Rectified linear units improve restricted Boltzmann machines. In: Proceedings of the 27th International Conference on Machine Learning (ICML-10), pp. 807–814 (2010)
35. Godard, C, Mac Aodha, O.: Unsupervised monocular depth estimation with left-right consistency. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6602–6611
36. Milletari, F.: V-net: fully convolutional neural networks for volumetric medical image segmentation. In: 2016 Fourth International Conference on 3D Vision (3DV), Stanford, CA, USA, 2016, pp. 565–571 (2016)
37. Tang, J., Fan, D.: RGBT Salient Object Detection: Benchmark and A Novel Cooperative Ranking Approach, vol. 30, no. 12, pp. 4421–4433 (2020)
38. Liu, Z., Tan, Y.: SwinNet: swin transformer drives edge-aware RGB-D and RGB-T salient object detection. IEEE Trans. Circuits Syst. Video Technol. **32**(7), 4486–4497 (2022)
39. Huo, F., Zhu, X.: Real-time one-stream semantic-guided refinement network for RGB-thermal salient object detection. IEEE Trans. Instrum. Meas. **71**, 1–12 (2022)
40. Cong, R., Zhang, K.: Does thermal really always matter for RGB-T salient object detection? IEEE Trans. Multimed. **25**, 6971–6982 (2023)
41. Tu, Z., Li, Z.: Weakly alignment-free RGBT salient object detection with deep correlation network. IEEE Trans. Image Process. **31**, 3752–3764 (2022). https://doi.org/10.1109/TIP.2022.3176540
42. Zhou, W., Zhu, Y.: LSNet: lightweight spatial boosting network for detecting salient objects in RGB-thermal images. IEEE Trans. Image Process. **32**, 1329–1340 (2023)
43. Tang, B, Liu, Z.: HRTransNet: HRFormer-driven two-modality salient object detection. IEEE Trans. Circuits Syst. Video Technol. **33**(2), 728–742 (2023)

44. Pang, Y., Zhao, X.: CAVER: cross-modal view-mixed transformer for bi-modal salient object detection. IEEE Trans. Image Process. **32**, 892–904 (2023). https://doi.org/10.1109/TIP.2023.3234702
45. Liu, Z, Lin, Y.: Swin transformer: hierarchical vision transformer using shifted windows. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 9992–10002 (2021). https://doi.org/10.1109/ICCV48922.2021.00986
46. Wang, X., Girshick, R.: Non-local neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7794–7803 (2018)