Modality-Aware Adaptive-Integration Guided Single-Stream Network for RGB-T Saliency Detection

Yu Pang[a], Hao Wu[b], Chengdong Wu[c], Yuanyuan Tan[a]

[a]School of Artificial Intelligence, Shenyang University of Technology, Shenyang, 110870, China
[b]Faculty of Computer Science, Macquarie University, Sydney, NSW 2109, Australia
[c]Faculty of Robot Science and Engineering, Northeastern University, Shenyang, 110167, China

Abstract

RGB-T saliency detection becomes gradually a hot topic in saliency detection field recently. However, existing works (especially CNN based methods) usually use the two-stream structure to separately extract saliency cues from RGB and thermal infrared images, and then integrate them into the final detection result, this strategy greatly increases parameters scale while multi-modal fusion results are also very sensitive to two modalities' quality. Based on above observation, we develop a novel Modality-Aware Adaptive-Integration Guided Single-Stream Network(MAANet), to detect salient objects from RGB-T image pairs. The feature pyramid network(FPN) is adopted as the basic structure of our MAANet. In order to tactfully fuse two supplementary modalities: *In the encoder*: RGB and thermal infrared images are concatenated into 4-channel input of the encoder structure in the proposed MAANet. *In the decoder*: We propose a novel Modality-Aware Adaptive-Integration based Attention mechanism (MAAM) to enable the decoder to optimally perform the fusion of two modalities, and produce more accurate saliency predictions. *Finally*: A novel coarse-and-refined bidirectional optimization(CRBO) method is proposed to suppress irrelevant background regions of saliency map generated by the decoder. The proposed MAANet could better take both advantages of two modalities and is not sensitive to any one modality compared to previous RGB-T methods, meanwhile, MAANet is also more lightweight than previous works. Extensive experiments demonstrate that the proposed model performs favorably against most state-of-the-art RGB-T methods under different evaluation metrics, even outperforms than most RGB and RGB-D methods. Our codes and results are released at https://github.com/SUTPangYu/MAANet

Keywords: RGB-T salient object detection; Single-stream network; Attention mechanism; Modality-aware strategy; Coarse-and-refined integration;

1. Introduction

Saliency detection aims to estimate the most interesting object of an image, since they have different visual significance with other background regions in subsequent computer vision and image processing tasks, such as semantic recognition [1]-[4], image segmentation [5], visual tracking[6] and image compressing [7].

With the rapid development of deep convolutional neural networks (CNNs), numerous CNN-based methods [8]-[15] are proposed to detect salient objects from RGB images. However, they are powerless in some complicated scenes due to the limitation of RGB data themselves, such as large illumination variation, haze and smog, night scene and so forth. Under these circumstances, only using the information provided by RGB camera is not able to generate accurate saliency predictions. In order to solve above problems, RGB-T(thermal infrared data [1]) salient object detection is becoming gradually a hot topic recently, since thermal infrared cameras can capture infrared radiation emitted by the object whose temperature is above absolute zero. Therefore RGB and thermal infrared data complement each other for handling complex environments, this provides theoretical basis for the development of RGB-T salient object detection.

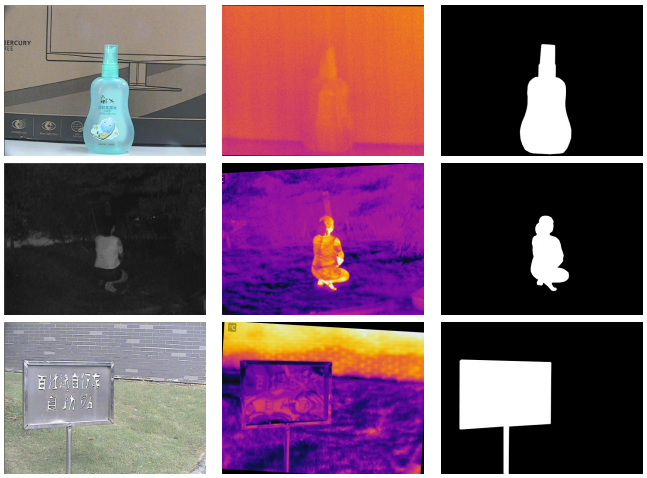
1.1. Problem description

As a new research direction to be developed, there are three unsolved problems in existing RGB-T methods:

Problem1: Two-stream structure is adopted widely in existing works, but it greatly increases the number of parameters in the network, efficiency is thus limited. As an important pre-processing stage in computer vision and image processing tasks, this limitation is fatal to saliency detection. Therefore, how to establish a lightweight RGB-T multi-modal saliency module is an unsolved problem.

*Corresponding author:Yu Pang, pangyu@sut.edu.cn;

[1]In this paper, "T", "Thermal infrared image", "Thermal modality" and "Thermal infrared data" represent the same thing. In order to clearer description, we use them interchangeably for different explanation purposes.



(a) (b) (c)

Figure 1: RGB and thermal infrared images in different types of scenes (a)RGB images (b)Thermal infrared images (c)Ground truth

Problem2: Existing two-stream structures in RGB-T saliency detection field regard two modalities as equal roles and only achieve simple modality interaction. However, there is a great difference between RGB and thermal infrared modalities, they have their own strengths in different types of scenes. Such as the first row of Fig.1, RGB image is more discriminative than thermal infrared image for capturing the contrast between foreground and background. Meanwhile, thermal modality has better performance than RGB modality in the second row. A more extreme case is the third row in Fig.1, both two modalities fail to detect salient object from the given scene, each of two modalities only captures contrast cues in different regions, this case increases the difficulty of modality fusion. In summary, simple multi-modal integration might result that integration results are sensitive to two modalities, i.e., integration results might be powerless when any of two modalities works bad. Therefore, how to explore deeply the relationship between two modalities, to avoid the influence of false saliency cues originated from terrible modality to the final detection result, is also an unsolved problem.

Problem3: FPN(feature pyramid network) is a encoder-decoder structure which is insensitive to object scale of scene. Most CNN-based saliency detection methods adopt FPN as basic structure, they utilize directly the finest high-resolution decoder output as the final prediction. However, this results that some tiny noises cannot be suppressed effectively, since both the first encoder layer and the penultimate decoder layer join the construction of the final decoder output, in which the first encoder layer might bring some tiny noises in some complicated scenes. Although previous works [16] introduce global cues to solve this problem, efforts are still unsatisfactory, since previous remedy strategy only integrates directly the encoder and decoder outputs to solve this problem, which is too simple to improve performance greatly.

1.2. Our solution

To address above problems, we propose a Modality-Aware Adaptive-Integration Guided Single-Stream Network(MAANet), which utilizes the FPN as basic structure.

Firstly: We concatenate RGB and thermal infrared images as 4-channel input of MAANet, and the encoder utilizes VGG16 net trained on ImageNet database to extract powerful features from RGB-T image pairs. Then, the decoder aims at resuming size and reducing channel number for feature maps, to generate the final saliency map. To solve problem1, we only regard thermal infrared image as attention map to guide the optimization of network, since thermal infrared image provides less content information compared to RGB image, but it's more suitable to enhance the network by using its ability capturing the contrast between foreground and background, such as Fig.1. Above single-stream CNN establishment strategy not only takes both advantages of two modalities but also achieves module lightweight, such as Fig.2.

Secondly: Based on the designed single-stream CNN, we propose a novel modality-aware adaptive-integration based attention mechanism (MAAM) between the encoder and the decoder, to solve problem2. As well known, the strategy adopting additional modality as attention map is still not used in RGB-T methods and is only applied to RGB-D methods(reason is analyzed in problem2 description). Thereby we not only construct *thermal-induced attention map* but also construct *mask-guided attention map*. Then our module sequentially learns to build *pixel-wise modality-aware weighted map*, to search "reliable regions" whose saliency values could be accurately inferred by thermal-induced attention map. Based on the pixel-wise modality-aware weighted map, we build modality-aware integration mechanism to achieve pixel-wise integration of thermal-guided and mask-guided attention maps, this operator could better guide the relationship between thermal infrared prior and saliency cues, and can filter the inaccurate cues from thermal modality and make thermal modality better optimize the network. Thereby enhancing the contrast of network output and the whole network is able to produce more discriminative prediction.

Thirdly: To solve problem3, we also develop a coarse-and-refined bidirectional optimization(CRBO) method to further optimize saliency map after obtaining the output of network. Based on observation that the encoder output indicates coarse saliency cues but is insensitive to tiny noises, the CRBO builds the bidirectional optimization between the encoder output and the decoder output, to suppress tiny noises and modify some details for generating more accurate saliency map.

In summary, the contributions of our work are listed as follows:

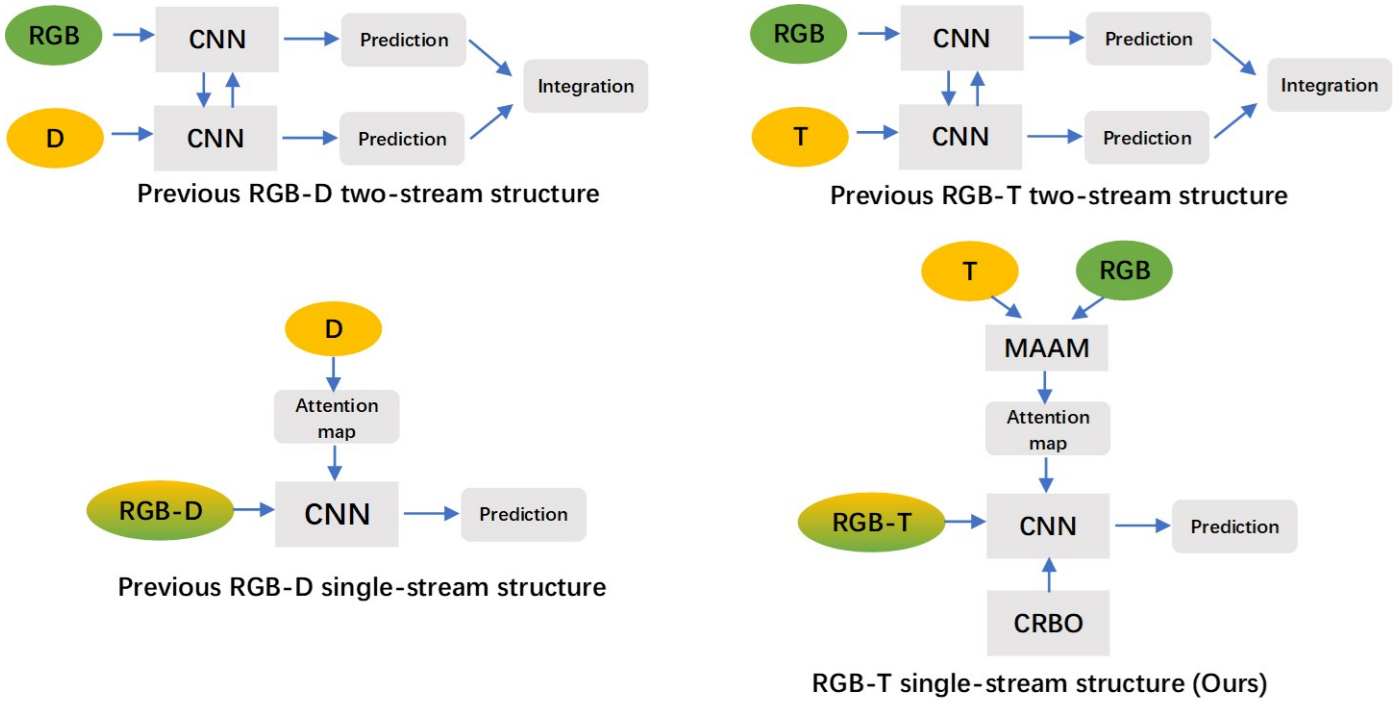


Figure 2: The difference between our proposed method and previous works. Previous RGB-D methods follow single-stream or two-stream structures, and previous RGB-T methods only follow two-stream structure. We are the first time to propose RGB-T single-stream framework while our single-stream has great difference to RGB-D single-stream structure. "RGB", "D" and "T" refer to RGB image, Depth map and Thermal infrared image. "MAAM" and "CRBO" are respectively our proposed modality-aware adaptive-integration based attention mechanism and coarse-and-refined bidirectional optimization.

1) Comparing to previous RGB-T saliency detection works, we are the first time to propose a single-stream CNN framework which is insensitive to low-quality modality, to achieve high prediction accuracy and module lightweight simultaneously.

2) We propose a modality-aware adaptive-integration based attention mechanism (MAAM). Different from previous works that utilize directly additional modality (thermal infrared image or depth map) as attention map, our model could make thermal modality better optimize network via constructing saliency cues filtering mechanism, to achieve more accurate RGB-T saliency prediction.

3) We develop a coarse-and-refined bidirectional optimization(CRBO) module to further improve the final saliency prediction accuracy by exploring the relationship between the refined high-resolution and the coarse low-resolution outputs effectively.

4) Our module performs much better than other competitors on three challenging RGB-T datasets. Even, our module also achieves top performance on some RGB-D datasets.

2. Related Work

2.1. RGB saliency detection

In the early stage, saliency detection methods usually adopt low-level features, such color, texture, gradient features, etc. Meanwhile, they locate salient objects under the help of prior knowledge, e.g., background prior [17, 18], contrast prior [19], center prior [20], etc. Considering that only utilizing prior knowledge might be powerless, some methods attempt to optimize saliency maps generated by prior knowledge via various models, including graph-based diffusion mechanisms [20, 21], low-rank matrix recovery methods [22], etc. However, all above models rely heavily on heuristic hand-crafted features, model's generalizability is limited.

Recent advances in this field have been obtained by CNN-based methods since hierarchical structure learned by these modules can encode the high-level semantic information from the complicated scenes. Hou *et al.* [8] propose a novel salient object detection network with short connections between layers. Zhao *et al.* [9] design gated units to better fuse the encoder and decoder blocks in feature pyramid network. Wang *et al.* [10] design a pyramid attention structure to better achieve multi-scale saliency cues fusion while further enhance the edge information of saliency map.

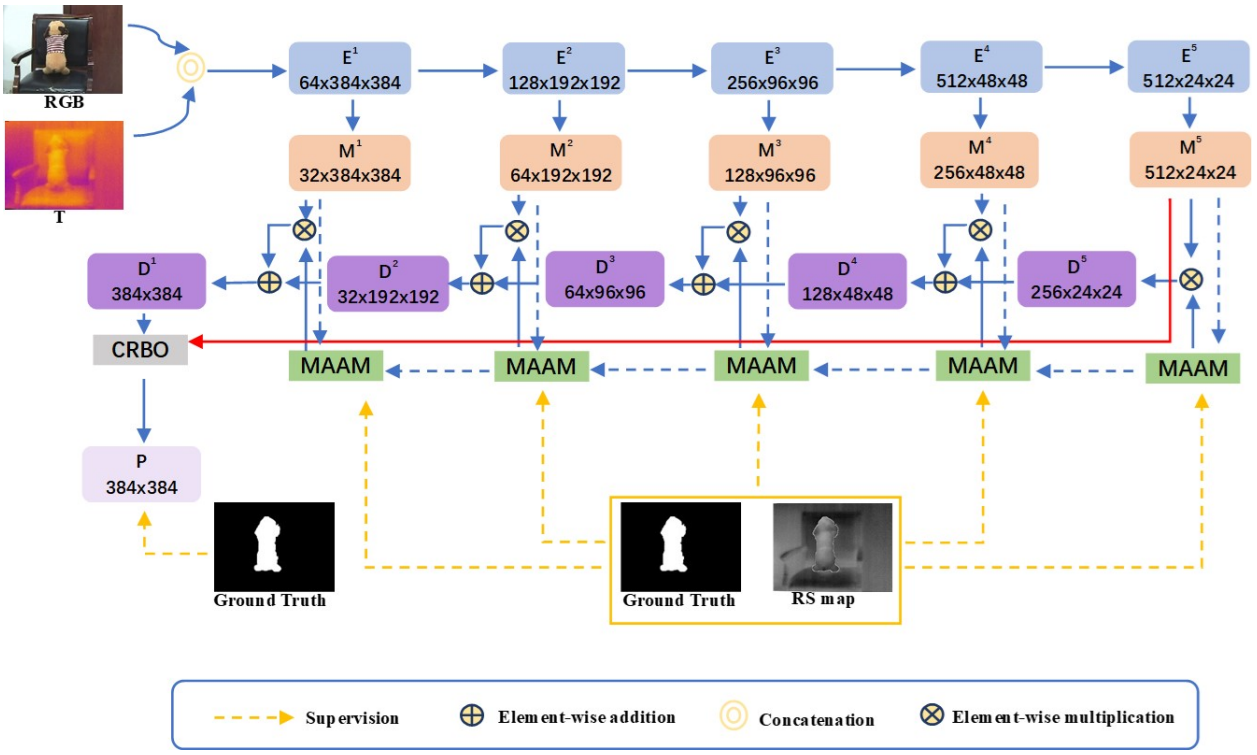


Figure 3: The framework of the proposed MAANet. "RGB" and "T" represents RGB and thermal infrared images, E^i,M^i and D^i are respectively the i-th encoder block, transition block and decode block. "MAAM" corresponds to the modality-aware adaptive-integration based attention mechanism and "CRBO" is the coarse-and-refined bidirectional optimization, they are introduced in Fig.4 and Fig.5. P is the final saliency prediction.

Whether they are traditional methods or deep methods, their structures are designed based on RGB data. However, these above methods might fail to distinguish salient object from the complicated background in some challenging conditions, such as poor illumination, complex background, and low contrast, etc.

2.2. RGB-T saliency detection

With the rapid development of thermal infrared sensors, RGB-T saliency detection task is gradually advocated by taking both advantages of RGB and thermal infrared images. Comparing to RGB camera, thermal infrared cameras are a kind of passive sensors that capture the thermal infrared radiation emitted by all objects with temperature above absolute zeros, meaning that thermal infrared images are invariant to illumination conditions [23]. Therefore, thermal infrared images could provide an additional help for RGB images to boost saliency performance, especially in some complicated scenes.

Unlike RGB methods, as a new research direction, only a few RGB-T modules are proposed currently. Li *et al.* [24] collect a new RGB-T dataset named VT821, and propose a multi-task manifold ranking algorithm to enforce cross-modal consistency. Tu *et al.* [25] present a collaborative graph learning algorithm to take both advantages of two modalities in the graph construction of salient object detection while collect a challenging RGB-T dataset, i.e., VT1000. In [26], Ma *et al.* learn multi-scale deep features from RGB and thermal infrared images while train SVM regressors for RGB-T saliency detection. Zhang *et al.* [27] utilize the pre-trained network to extract the coarse features from RGB and thermal images respectively, novel modules are further developed to fuse them and generate accurate saliency prediction. In [28] and [29], features extracted from RGB and thermal infrared images are fused to enforce saliency consistency based on a novel CNN framework. Zhang *et al.* [30] fuse different levels and different modalities to achieve RGB-T integration for more accurate saliency detection. In [29], Wang *et al.* propose a two-stream network, called cross-guided fusion network, to detect salient objects via an asymmetric cross-modal integration mechanism.

Most CNN-based RGB-T salient object detection modules are designed as two-stream structure. In particular, two independent networks are built based on RGB and thermal images, *the outputs of two networks are fused directly* or *features extracted from two networks are fused, then the fused feature is fed into additional module to generate the final prediction.* However, two modalities are usually fused directly no matter which way is employed in previous works, meaning that the final prediction might be limited if any of two modalities fails to capture the contrast between foreground and background, since they ignore the fact that two modalities have their own strengths in dif-

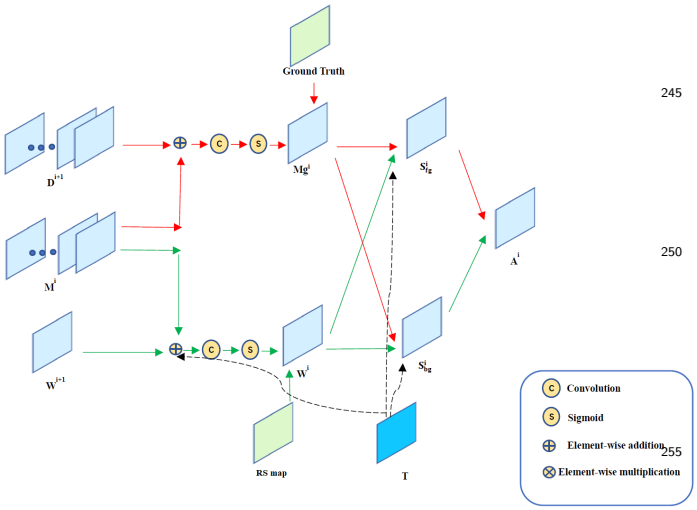


Figure 4: The flow of the proposed MAAM. A^i is the output of MAAM in the i-th decoder layer.

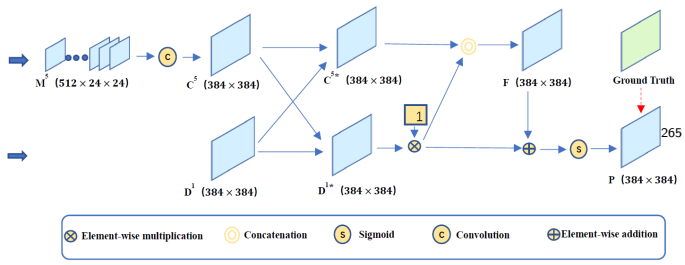


Figure 5: The framework of the proposed CRBO. It's emphasized that "1" in the yellow square box is also set to different values, 1 is the best value in this paper, but other tasks or datasets might be set to other values for obtaining better performance. In order to emphasize its importance, we retain this block for guiding other readers

ferent aspects, i.e, two modalities adapt different types of scenes. Furthermore, two-stream structure also increases greatly parameters number and reduce model efficiency.

In this paper, we deeply analyze the characteristics of the thermal stream itself as well as its internal relationship with the RGB stream, and develop a novel single-stream network, where the thermal modality focuses more on playing a guiding role to enhance its effect in the network, rather than the blind fusion of the two modalities.

3. Proposed method

We adopt the feature pyramid network(FPN) [31] as our basic structure, overall architecture is shown in Fig.3, where the whole framework includes the encoder and the decoder structures. $\mathbf{E}^i$, $\mathbf{M}^i$ and $\mathbf{D}^i$ correspond to the i-th encoder layer, transition layer and decoder layer. Here $i \in \{1, 2, 3, 4, 5\}$ and their output feature maps are denoted as E^i, M^i and D^i. Each transition layer uses a 3×3 convolution operation to optimize the features maps from the corresponding encoder block. Each decoder layer is generated by using a novel modality-aware adaptive-integration based attention mechanism (MAAM) to fuse current transition layer and previous decoder layer, such as Fig.4. The generated D^1 is the single-channel saliency prediction with the same resolution as input RGB-T image pair. Finally, we develop a novel coarse-and-refined bidirectional optimization (CRBO) framework to integrate D^1 and M^5 into the final prediction P, such as Fig.5.

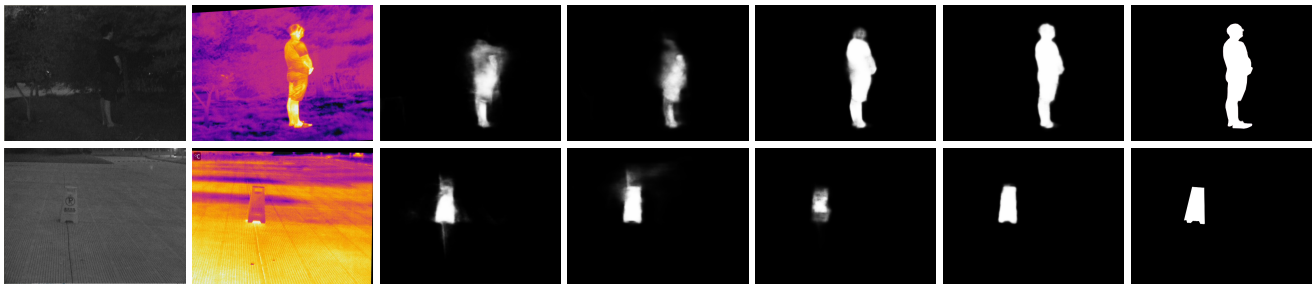
3.1. Encoder network

Given RGB image and corresponding thermal image, we concatenate them as 4-channel input, which is fed into a single-stream network with FCN structure. Here, we use VGG16 [32] where parameters are pre-trained on the ImageNet as encoder network. It's noticed that VGG net trained on ImageNet mainly adapts 3-channel RGB images, we adopt DPR method [33] to initialize the parameters of the first convolutional layer and other layers adopt conventional ImageNet pre-trained parameters, this operator makes two modalities' information can be fused in the input stage. Fig.6 shows the visual results of the proposed method with different outputs: Fig.6(c) are baseline results (i.e., FPN), Fig.6(d) are the results of our proposed MAANet with RGB images as inputs, and Fig.6(e) are the results of our proposed MAANet with the addition of RGB and thermal infrared images as inputs. Fig.6(f) are the results that use the concatenation of RGB and thermal infrared images as the inputs of our MAANet. We find that, using "addition" of RGB and thermal infrared images as input could generate more accurate detection results when thermal infrared image is more discriminative than RGB image (the first row), otherwise, only using RGB image as input is better (the second row). But their results are inferior to the results using the "concatenation" of RGB and thermal infrared images as inputs, since "concatenation" can suppress the feature response of channels generated by the modality that is not sufficiently discriminative.

The outputs of five layers in the encoder network are respectively conv1-2, conv2-2, conv3-3, conv4-3 and conv5-3 in VGG16, their sizes and channel numbers are shown in Fig.3. Five transition layers are generated by using 3×3 convolution operation to process the feature maps from the corresponding encoder layer. The goal is to adjust the channel number of feature maps for all encoder layers and make each encoder's channel number be the same with the corresponding decoder layer. Their sizes and channel numbers are also shown in Fig.3.

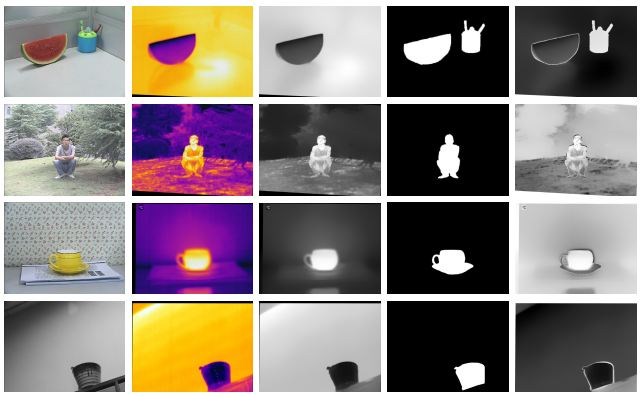
3.2. Modality-aware adaptive-integration based attention mechanism(MAAM) in decoder network

Comparing to RGB-D methods adopting depth map as attention map directly, thermal infrared images' ability capturing the contrast between foreground and background is not as good as that of depth map, this hinders the development of attention mechanisms for RGB-T



(a) (b) (c) (d) (e) (f) (g)

Figure 6: The results of different inputs to the proposed MAANet (a)RGB images (b)Thermal infrared images (c)Baseline (d)MAANet with RGB image as input (e)MAANet with the addition of RGB and thermal infrared images as input (f)MAANet with the concatenation of RGB and thermal infrared images as input (g)Ground Truth



(a) (b) (c) (d) (e)

Figure 7: The visual results of RS maps (a)RGB images (b)Thermal infrared images (c)Thermal-guided attention maps (d)Ground truth (e)RS maps

saliency detection. To solve this problem, we propose a modality-aware adaptive-integration based attention mechanism (MAAM). For the decoder network, the output of the i-th decoder layer D^i is denoted as:

$$D^i = \begin{cases} EDI(D^{i+1}, M^i, A^i) & \text{if} \quad i = 1, 2, 3, 4. \\ EDI(M^i, A^i) & \text{if} \quad i = 5. \end{cases} \tag{1}$$

where A^i denotes the attention map generated by MAAM in the i-th decoder layer, and we have

$$EDI(D^{i+1}, M^i, A^i) = A^i \otimes M^i + D^{i+1}. \tag{2}$$

and

$$EDI(M^i, A^i) = A^i \otimes M^i. \tag{3}$$

where A^i is the modality-aware adaptive-integration based attention map(the result of the i-th MAAM), which could help to enhance the contrast between foreground and background in the i-th decoder output computation. M^i and D^{i+1} are the outputs of the i-th transition layer and the $i+1$-th decoder layer.

$A^i (i = 1, 2, ..., 5)$ in Eq.(1) will play important role in the whole network. *Different from previous attention map construction, our MAAM model adopts a learning mechanism to achieve decision integration of mask-guided attention map and thermal-guided attention map. The contributions of our MAAM are listed as follows: (1) We are the first to propose the concept of thermal-guided attention map. (2) We propose a modality-aware adaptive-integration mechanism to integrate mask-guided attention map and the proposed thermal-guided attention map for generating more discriminative attention map. (3)We also propose a foreground-background dual optimization strategy in the multi-modal integration stage to optimize attention map.*

The flow of the proposed MAAM is shown in Fig.4, now we detail how to compute attention map in the i-th decoder layer:A^i. Before this, we need to build three relative components: thermal-guided attention map T', pixel-wise modality-aware weighted map W^i, mask-guided attention map Mg^i, they are detailed from section3.2.1 to section3.2.3. Finally, a novel foreground-background dual optimization(FBDO) is proposed to integrate them into the final A^i, FBDO is detailed in section3.2.4.

3.2.1. Thermal-guided attention map

We define thermal infrared image to be T, in order to make it guide and optimize the network accurately, we use it as attention map. Considering that there might be tiny noises in raw thermal infrared image due to the limitation of thermal sensors. We firstly gray thermal infrared image T, then use Gaussian filter to smooth it, we regard smoothness result as thermal-guided attention map directly, it's defined to be T'. The visual results of thermal-guided attention map are shown in Fig.7(c), they are the thermal-guided attention maps of the corresponding thermal infrared images (Fig.7(b)), we find that they have fewer tiny noises and are more suitable to guide the network.

However, only suppressing noises in thermal infrared image is insufficient. Thermal infrared image's discriminative ability separating foreground and background is usually not enough, thereby only utilizing thermal-guided attention map T' cannot enhance contrast effectively, we need to learn additional mask-guided attention map from RGB data, i.e., section3.2.2.

3.2.2. Mask-guided attention map

Constructing mask-guided attention map is inspired by previous works [34], i.e., under the supervision of ground truth, generating a saliency prediction as attention map to enhance the contrast between foreground and background in the decoder. Our mask-guided attention map is:

$$Mg^i = \begin{cases} \delta(Conv(M^i + D^{i+1})) & \text{if} \quad i = 1, 2, 3, 4. \\ \delta(Conv(M^i)) & \text{if} \quad i = 5. \end{cases} \tag{4}$$

where Mg^i is supervised by the ground truth, M^i and D^{i+1} refer to the outputs of the i-th transition layer and the i+1-th decoder layer. $Conv()$ and $\delta()$ are convolutional layer and element-wise sigmoid function respectively. It's noticed that, we only use the output of transition layer when constructing the mask-guided attention map in the 5-th decoder layer(i.e., Mg^i when $i = 5$), since there is not upper decoder layer at this time.

Transition layer's parameters originate from VGG16 net trained on ImageNet database which contains numerous RGB images, thereby mask-guided attention map helps to enhance the contrast between foreground and background based on RGB stream, it provides an important supplement for thermal-guided attention map.

3.2.3. Pixel-level modality-aware weighted(PWM) map W^i in the i-th decoder layer

Thermal-guided and mask-guided attention maps focus on enhancing the contrast for the network based on thermal and RGB streams. Thermal-guided attention map provides more accurate contrast information if thermal infrared image identifies salient object more accurately than RGB image in the given scene. Otherwise, we try our best to use mask-guided attention map to guide the network. Based on the above observation, integrating two attention maps is an effective strategy similarly. However, this kind of image-level integration only adapts the situation that at least one modality is effective, such as the first two rows in Fig.1. When both two modalities fail to capture the contrast between foreground and background(Such as the third row of Fig.1), above image-level integration is powerless. In other words, for each pixel m in the given scene, we need to infer which modality could predict its saliency value accurately. This kind of pixel-wise integration is more effective similarly at this time.

Inspired by above observation and analysis, we are the first to propose the concept of pixel-level modality-aware weighted map(PWM map) , it aims to infer which regions' saliency values are correctly predicted by thermal modality. We have the pixel-wise modality-aware weighted map W^i in the i-th layer:

$$W^i = \begin{cases} \delta(Conv(M^i + W^{i+1} + T')) & \text{if} \quad i = 1, 2, 3, 4. \\ \delta(Conv(M^i + T')) & \text{if} \quad i = 5. \end{cases} \tag{5}$$

where M^i and W^{i+1} refer to the outputs of the i-th transition layer and the i + 1-th pixel-wise modality-aware weighted map, T' is thermal-guided attention map. $Conv()$ and $\delta()$ are convolutional layer and element-wise sigmoid function respectively, they ensure that W^i is a single-channel map. W^i is supervised by RS(Loss item is listed in Eq.(12)), here RS is defined to be :

$$RS = \exp(1 - |T' - GT|). \tag{6}$$

where T' and GT correspond to thermal-guided attention map and ground truth. For any pixel m, higher value in the RS indicates that pixel m's saliency value is predicted by thermal modality more accurately, i.e., thermal modality has higher probability to predict accurately the saliency value of pixel m. Via minimizing the third item of Eq.(12), we can learn a reliable PWM map in each layer to infer prediction accuracy of T' and Mg^i for any pixel m(Mg^i is also single-channel map with the same resolution as W^i while we resize T' to the size of W^i in the i-th layer). The introduction of the PWM map W^i can balance effectively the contributions of different attention maps by perceiving thermal modality's quality, and thereby achieving pixel-wise decision-level multi-modal integration.

The visual results of RS are shown in Fig.7. Fig.7(c) and Fig.7(d) are respectively thermal-guided attention maps and ground truth, Fig.7(e) are corresponding RS maps. For the first row, thermal modality only can predict accurately the saliency value of blue pen container, thereby only this region has high value in the corresponding RS map. In contrast, if thermal modality can capture the contrast between foreground and background completely, all pixels' values in the RS map are close to 1 or 0, such as the third and fourth rows. Here, it's emphasized that a good attention map could have positive correlation(the third row) or negative correlation (the fourth row) with the ground truth.

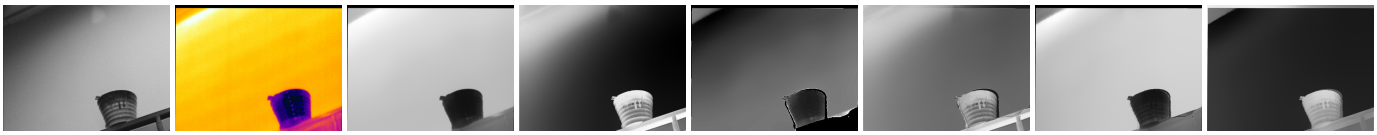
3.2.4. MAAM attention map A^i computation based on a novel foreground-background dual optimization strategy(FBDO)

After obtaining T', $W^i(i = 1, 2, ..., 5)$ and $Mg^i(i = 1, 2, ..., 5)$, we propose a foreground-background dual optimization(FBDO) strategy to achieve the multi-modal integration of T' and Mg^i to obtain A^i:

$$A^i = S^i_{fg} \otimes (1 - S^i_{bg}). \tag{7}$$

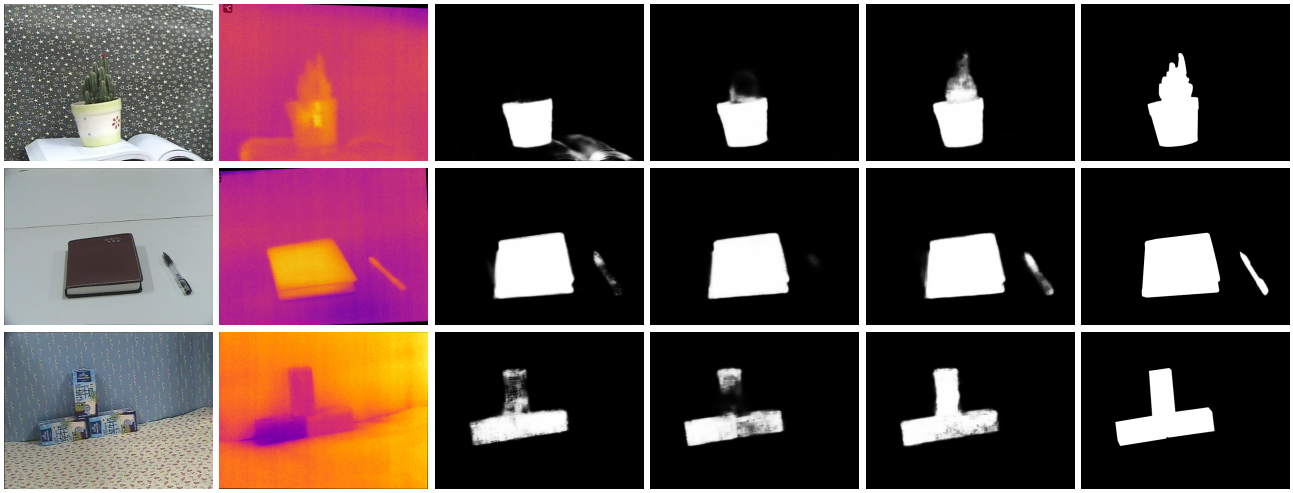
where

$$S^i_{fg} = W^i \otimes T' + (1 - W^i) \otimes Mg^i. \tag{8}$$



(a) (b) (c) (d) (e) (f) (g) (h)

Figure 8: The visual result of foreground-background dual optimization, we take the 3-th($i = 3$) decoder layer as an example. (a)RGB image (b)Thermal infrared image (c)Thermal-guided attention map T' (d)Mask-guided attention map Mg^i (e)Pixel-wise modality-aware weighted map W^i (f)Foreground attention map S^i_{fg} (g)Background attention map S^i_{bg} (h)MAAM attention map A^i



(a) (b) (c) (d) (e) (f)

Figure 9: The comparison between the proposed MAAM and other attention mechanisms (a)RGB images (b)Thermal infrared images (c)MAANet without any attention mechanism (d)MAANet using mask-guided attention mechanism (e)MAANet (f)Ground Truth

and

$$S^i_{bg} = (1 - W^i) \otimes T' + W^i \otimes (1 - Mg^i). \quad (9)$$

where A^i is determined by S^i_{fg} and S^i_{bg}. Firstly, S^i_{fg} is called foreground attention map and computed in Eq.(8). For pixel m, its value in S^i_{fg} is mainly determined by its value in T' if we infer that thermal modality predicts the saliency value of pixel m more accurately(i.e., its value in W^i is higher). Otherwise, we mainly use its value in Mg^i. Secondly, S^i_{bg} is called background attention map and computed in Eq.(9), similarly, we still use W^i as balancing item to balance T' and $1 - Mg^i$ to extract background cue, it's emphasized that $1 - Mg^i$ aims to extract background cue since Mg^i is supervised by the ground truth.

Constructing the FBDO is based on the following analysis: S^i_{fg} itself has been enough to extract saliency cues, we could regard S^i_{fg} as A^i directly. However, above operator is based on the assumption that there is positive correlation between thermal modality and ground truth, e.g., the second and the third rows of Fig.7. But, attention map which has negative correlation with ground truth also could be regarded as a good attention map, e.g., the last row of Fig.7. We take it as an example to explain the contribution of the FBDO in Fig.8. Thermal-guided attention map(Fig.8(c)) confuses foreground and background, which results that all pixel's values in the RS map are close to 0(Fig.8(e)), mask-guided attention map based on RGB data will play a major role in the S^i_{fg} construction, thereby Mg^i(Fig.8(d)) and S^i_{fg}(Fig.8(f)) are very similar. However, Mg^i is very powerless and not a good attention map at this time. In contrast, thermal-guided attention map still could provide contrast information even if there is negative correlation between it and ground truth. Based on above analysis, we establish Eq.(9) to extract background cues, which could take full advantage of thermal modality. Since the integration of S^i_{fg} and S^i_{bg} is more effective than single S^i_{fg}. In other words, the fusion of Eq.(8) and Eq.(9) unleashes thermal modality's ability capturing contrast information and makes thermal modality better guide the network, so that generating more discriminative attention map A^i.

After A^i computation, we are able to obtain D^i according to Eq.(1). Until $i = 1$, D^i is a single-channel map

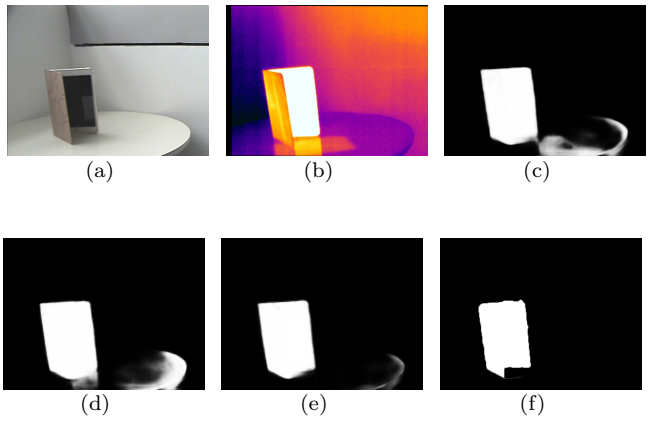


Figure 10: The visual result of the contribution of CRBO (a)RGB image (b)Thermal infrared image (c)Decoder output D^1 (d)The addition result of the coarse localization and the refined output, i.e., $D^1 + C^5$ (e)CRBO (f)Ground Truth

with size 384×384. Under the supervision of the ground truth, it's regarded as an effective saliency prediction.

Fig.9 shows the visual results of the comparison between the proposed MAAM and other attention mechanisms. Fig.9(c) are the results without any attention mechanism, i.e., there is not A^i in Eq.(2) and Eq.(3). Fig.9(d) are the results using mask-guided attention mechanism, i.e., using Mg^i to replace A^i in Eq.(2) and Eq.(3). We find that their results are inferior to MAANet with MAAM. Such as the first row, green flower shares similar features with surrounding background in RGB image, thermal modality could provide additional contrast information, the proposed MAAM could make thermal modality guide the network at this time. In contrast, green flower is difficult to be found if not utilizing attention mechanism or utilizing mask-guided attention mechanism, since both these two strategies ignore the important role of thermal modality. Meanwhile, two modalities have their own strengths in different regions for the third RGB-T image pair. Our proposed MAAM can detect salient object more completely compared to other two modules, since our proposed MAAM aims to achieve pixel-wise multi-modal integration, which is more effective than image-level integration in this situation.

3.3. Coarse-and-refined bidirectional optimization(CRBO)

The output of the decoder (i.e., D^1) is determined by the previous decode layer output D^2 and current transition layer output M^1. Considering that M^1 originates from E^1 which is inevitable to contain some tiny background noises (e.g., redundant edges, tiny spot) in some complicated scenes, high-resolution output D^1 also suppresses background noises hardly, these tiny noises might be regarded as small-scale objects. Therefore, adding low-resolution output as supplement is very necessary, since the coarse localization cannot segment clear object boundary but provides the basis for top-down high-resolution output. Some works attempt to fuse directly encoder output (the coarse localization) and decoder output (the refined localization) into the final prediction, like [16]. However, this simple fusion is difficult to achieve great performance improvement.

In order to solve above problem, we propose a novel coarse-and-refined bidirectional optimization(CRBO) strategy. The flow of the proposed CRBO is shown in Fig.5, its establishment is based on two observations: For one thing, the refined high-resolution output could better capture the edge information of salient object but is sensitive to some tiny background noises. For the other, the coarse low-resolution output cannot provide accurate details of salient object, but gives a coarse object localization from global distribution cue and is insensitive to tiny background noises. To this end, the CRBO aims at achieving the co-optimization between the coarse and the refined saliency predictions, the specific calculation process of the CRBO is listed as follows:

Firstly, on the basis of the refined output D^1, we extract the fifth transition layer output M^5 and then use 3×3 convolution operator to process it for generating the coarse output C^5 which is also a single-channel map with size 384×384.

Secondly, based on their own strengths(D^1 and C^5), we respectively optimize them:

$$D^{1*} = D^1 - \exp(-\beta C^5). \tag{10}$$

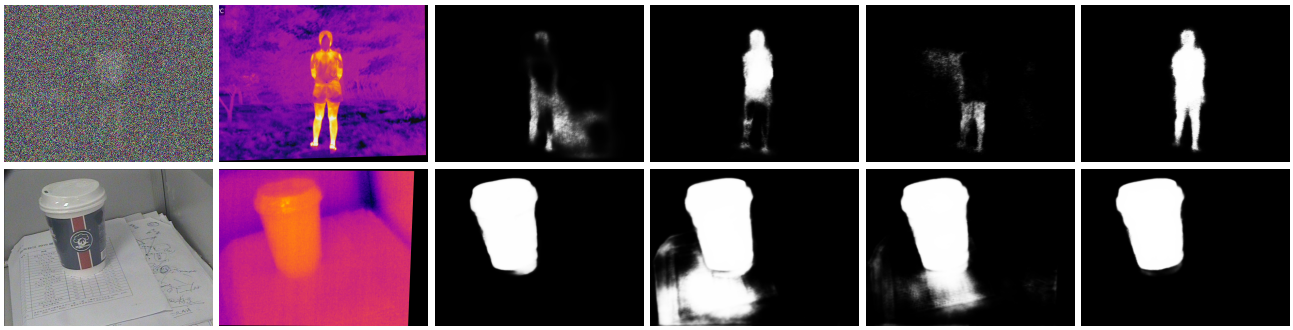
and

$$C^{5*} = C^5 \otimes D^1 + C^5. \tag{11}$$

As shown in Fig.5, Eq.(10) and Eq.(11) are computed simultaneously. The goal of Eq.(10) is to use the coarse output C^5's ability locating coarsely salient object to suppress some tiny noises which are far from salient object in D^1, parameter $\beta = 6$ here. Comparing to the simple $D^{1*} = C^5 + D^1$, Eq.(10)'s computation way could make the optimized D^{1*} not change too much from D^1, and C^5 only modifies some noises. This operator avoids the effect of the coarse localization to object structure. Then the goal of Eq.(11) is to use D^1's ability segmenting salient object accurately to help the coarse output C^5 to obtain clearer edge information. Undoubtedly, Eq.(11) is more effective to achieve this goal than $C^{5*} = D^1 + C^5$, since "multiplication" is superior to "addition" in segmentation process.

Thirdly, we fuse D^{1*} and C^{5*} to generate the final prediction P. The final prediction P is learned under the supervision of the ground truth.

The contribution of the CRBO is shown in Fig.10, we find that the decoder output D^1 (Fig.10(c)) has numerous background noises mistaken as foreground. Although integrating directly encoder and decoder outputs (Fig.10(d)) achieves performance improvement slightly, but numerous background noises are still not suppressed. In contrast, utilizing the CRBO could generate more accurate prediction P (Fig.10(e)), there are almost not noises at this time.



(a) (b) (c) (d) (e) (f)

Figure 11: The visual results of the MAANet with/without PWM. (a)RGB images (b)Thermal infrared images (c)MAANet using mask-guided attention map (d)MAANet using thermal-guided attention map (e) MAANet using the addition of two attention maps (f)MAANet using MAAM

Table 1: The experimental results of different inputs to the proposed MAANet

	Baseline			RGB			RGB+T(addition)			RGB+T(Concatenation)		
	VT821	VT1000	VT5000	VT821	VT1000	VT5000	VT821	VT1000	VT5000	VT821	VT1000	VT5000
$\mathcal{M}$	0.0383	0.0253	0.0375	0.0357	0.0253	0.0368	0.0359	0.0227	0.0353	**0.0315**	**0.0179**	**0.0295**
F_β^{max}	0.8405	0.9248	0.8724	0.8323	0.9279	0.8728	0.8297	0.9337	0.8760	**0.8725**	**0.9426**	**0.8984**
F_β^{mean}	0.7607	0.8526	0.7824	0.7668	0.8605	0.7945	0.7578	0.8690	0.8032	**0.8049**	**0.8823**	**0.8400**
F_β^w	0.7557	0.8717	0.7900	0.7654	0.8727	0.7943	0.7663	0.8871	0.8063	**0.8130**	**0.9057**	**0.8352**
S_m	0.8602	0.9226	0.8785	0.8658	0.9225	0.8786	0.8664	0.9293	0.8847	**0.8839**	**0.9405**	**0.8998**
E_m	0.8817	0.9259	0.8971	0.8896	0.9349	0.9049	0.8763	0.9372	0.9095	**0.9066**	**0.9459**	**0.9265**

3.4. Loss function

The total loss function of our scheme is composed of the final prediction loss, mask-guided map loss and pixel-wise modality-aware weighted map loss. We assume that GT denotes supervision from the ground truth. The total loss $\mathcal{L}$ is defined as:

$$\mathcal{L} = BCE(P, GT) + \sum_{i=1}^{5} BCE(Mg^i, GT) + \sum_{i=1}^{5} MSE(W^i, RS). \quad (12)$$

where the first item is saliency prediction loss, P refers to the final prediction. The second item is mask-guided attention map loss, Mg^i is the mask-guided attention map in the i-th decoder layer. The third item is pixel-wise modality-aware weighted (PWM) map loss, W^i is PWM map in the i-th decoder layer, RS map is introduced in section 3.2.3 and it could be obtained via Eq.(6). In addition, the first two items adopt binary cross-entropy (BCE) loss to compute distance, since they are supervised by the ground truth which is binary supervision and BCE is mainly used to evaluate binary classification. The third item adopts mean square error (MSE) loss as metric, since RS is not binary and all values in RS are continuous, and MSE tends to evaluate the fitting degree of the model on the given data.

4. Experiments

We compare the proposed method with other state-of-the-art methods on three benchmark RGB-T datasets, including VT821 [24], VT1000 [23] and VT5000 [35]. All of them contain various indoor and outdoor scenes as well as corresponding thermal infrared images, like human, scenery, buildings, etc. Some scenes are full of low illumination and smog, even including night scenes. All RGB-T image pairs are labeled by manual annotation.

For fairness, six evaluation metrics are used in our experiments. In order to take full advantages of precision and recall, we firstly report three main metrics: F_β^{max}, F_β^{mean} and F_β^w . Mean absolute error ($\mathcal{M}$) aims to measure the mean difference between the prediction and the ground truth. S-measure(S_m) score evaluates performance by integrating two kinds of similarities. E-measure(E_m) can jointly capture image level information and local pixel pairing information for fairer evaluation.

4.1. Implementation details

We start with the backbone VGG-16 net in our proposed model, convolutional layers' weights initialization strategy and convolutional layers selection strategy are introduced in methodology section, it's inspired by [34]. As well known, there are two parts in the VT5000, including VT5000-training (2500 images) and VT5000-test (2500 images). Here, we use the VT5000-training as our training

Table 2: The experimental results of the proposed MAANet using MAAM and other attention mechanisms

	Baseline			MA			TA			MGA			MAAM		
	VT821	VT1000	VT5000	VT821	VT1000	VT5000	VT821	VT1000	VT5000	VT821	VT1000	VT5000	VT821	VT1000	VT5000
$\mathcal{M}$	0.0383	0.0253	0.0375	0.0370	0.0202	0.0338	0.0368	0.0213	0.0343	0.0342	0.0185	0.0310	**0.0315**	**0.0179**	**0.0295**
F_β^{max}	0.8405	0.9248	0.8724	0.8567	0.9409	0.8850	0.8585	0.9362	0.8841	0.8633	0.9423	0.8936	**0.8725**	**0.9426**	**0.8984**
F_β^{mean}	0.7607	0.8526	0.7824	0.7806	0.8776	0.8032	0.7853	0.8717	0.8082	0.7900	**0.8908**	0.8270	**0.8049**	0.8823	**0.8400**
F_β^w	0.7557	0.8717	0.7900	0.7762	0.8941	0.8098	0.7778	0.8919	0.8107	0.7971	0.9035	0.8284	**0.8130**	**0.9057**	**0.8352**
S_m	0.8602	0.9226	0.8785	0.8706	0.9359	0.8899	0.8712	0.9325	0.8872	0.8805	0.9394	0.8973	**0.8839**	**0.9405**	**0.8998**
E_m	0.8817	0.9259	0.8971	0.8947	0.9453	0.9075	0.8977	0.9383	0.9123	0.9048	**0.9535**	0.9245	**0.9066**	0.9459	**0.9265**

Table 3: The contribution of CRBO to the proposed MAANet

	Baseline			C5+D1			D1			P		
	VT821	VT1000	VT5000	VT821	VT1000	VT5000	VT821	VT1000	VT5000	VT821	VT1000	VT5000
$\mathcal{M}$	0.0383	0.0253	0.0375	0.0366	0.0196	0.0325	0.0360	0.0188	0.0315	**0.0315**	**0.0179**	**0.0295**
F_β^{max}	0.8405	0.9248	0.8724	0.8606	0.9412	0.8922	0.8691	**0.9433**	0.8918	**0.8725**	0.9426	**0.8984**
F_β^{mean}	0.7607	0.8526	0.7824	0.7914	0.8828	0.8145	0.8030	**0.8888**	0.8233	**0.8049**	0.8823	**0.8400**
F_β^w	0.7557	0.8717	0.7900	0.7838	0.8973	0.8162	0.7964	0.9013	0.8225	**0.8130**	**0.9057**	**0.8352**
S_m	0.8602	0.9226	0.8785	0.8750	0.9387	0.8952	0.8785	0.9384	0.8952	**0.8839**	**0.9405**	**0.8998**
E_m	0.8817	0.9259	0.8971	0.8984	0.9462	0.9142	0.9044	0.9514	0.9219	**0.9066**	**0.9459**	**0.9265**

Table 4: The results of the proposed MAANet based on different encoders

Methods		VGG16	VGG19	ResNet	VGG16-SA
VT821	$\mathcal{M}$	**0.0315**	0.0332	0.0383	0.0319
	F_β^{max}	**0.8725**	0.8680	0.8642	0.8711
	F_β^{mean}	**0.8049**	0.7998	0.7979	0.8044
	F_β^w	0.8130	0.8101	0.8003	**0.8141**
	S_m	0.8839	0.8832	0.8783	**0.8843**
	E_m	**0.9066**	0.9003	0.9001	0.9014
VT1000	$\mathcal{M}$	**0.0179**	0.0188	0.0190	0.0182
	F_β^{max}	**0.9426**	0.9423	0.9414	0.9423
	F_β^{mean}	**0.8823**	0.8818	0.8803	0.8820
	F_β^w	0.9057	0.9048	0.9042	**0.9059**
	S_m	**0.9405**	0.9386	0.9376	0.9402
	E_m	**0.9459**	0.9433	0.9389	0.9458
VT5000	$\mathcal{M}$	**0.0295**	0.0308	0.0313	0.0303
	F_β^{max}	**0.8984**	0.8978	0.8902	0.8977
	F_β^{mean}	**0.8400**	0.8365	0.8351	0.8363
	F_β^w	0.8352	**0.8354**	0.8280	0.8346
	S_m	**0.8998**	0.8882	0.8877	0.8990
	E_m	**0.9265**	0.9209	0.9201	0.9207

data to train the proposed MAANet. VT5000-test is used for testing, and all samples in the VT821 and VT1000 are also used for testing. Our model is implemented based on the Pytorch toolbox and trained on a PC with TITAN Xp GPU for 40 epochs with mini-batch size 4. Both input RGB image and thermal infrared image are resized to 384 × 384. For optimizing the network, we adopt the stochastic gradient descent (SGD) with a momentum of 0.9 and a weight decay of 0.0005. In order to promote learning ability, our learning rate is set to 0.001 and later use the "poly" policy [36] with the power of 0.9.

4.2. Ablation study

Our ablation experiments contain four parts, validating respectively the effectiveness of RGB-T fusion, MAAM, CRBO and encoder selection.

4.2.1. The validation of the effectiveness of RGB-T fusion in the encoder network

We fed different inputs into the proposed MAANet, the results of MAANet with different inputs are shown in Table.1. RGB and thermal infrared images are concatenated as 4-channel input of the encoder network in our method, it's represented by "RGB+T(concatenation)" in Table.1. We also evaluate other fusion strategies: "RGB" denotes that only three-channel RGB image is fed into the encoder network of our method and "RGB+T(addition)" means that we use the element-wise addition result of RGB and thermal infrared images as the input of the encoder network. Detection results based on different inputs are shown in Table.1, it's observed that concatenation input makes MAANet achieve better results than other strategies. Since four channels are parallel in concatenation operator, thereby feature computation of the color channels will not be affected if the quality of thermal infrared image is very poor.

4.2.2. The validation of the effectiveness of modality-aware adaptive-fusion based attention mechanism (MAAM)

We propose a novel modality-aware adaptive-integration based attention mechanism (MAAM) to optimize the network. We also utilize other attention mechanisms to replace the proposed MAAM, results are shown in Table.2. "MA" represents that there is not any attention mechanism in our method, i.e., we don't utilize A^i in Eq.(2) and Eq.(3). Comparing to "MAAM" that the proposed MAAM is utilized to enhance the contrast of network output, its performance is worse than "MAAM". Since introducing attention map into each layer could provide an important indicator to better capture the contrast between foreground and background. "MGA" and "TA" represent that mask-guided attention map and thermal-guided attention map are respectively utilized as the final attention map, i.e., Mg^i and T' respectively replace A^i in Eq.(2) and Eq.(3). We observe that both "MGA" and "TA"

Table 5: The comparison results of all methods on three RGB-T saliency detection datasets, red words represent the best values

Methods		**MAANet**	**MGFL**	**MTMR**	**SDGL**	**M3S-NIR**	**CRA**	**MIED**	**ADFNet**	**APNet**	**ECFFNet**	**FMCF**
Type		RGB-T	RGB-T	RGB-T	RGB-T	RGB-T	RGB-T	RGB-T	RGB-T	RGB-T	RGB-T	RGB-T
VT821	$\mathcal{M}$	**0.0315**	0.0714	0.1084	0.0850	0.1398	0.1010	0.0498	0.0766	0.0342	0.0345	0.0810
	F_β^{max}	**0.8725**	0.7793	0.7471	0.7793	0.7807	0.7727	0.8302	0.8041	0.8650	0.8645	0.7057
	F_β^{mean}	0.8049	0.7260	0.6629	0.7311	0.7350	0.7287	0.7608	0.7165	0.8162	0.8099	0.6385
	F_β^{w}	0.8130	0.6437	0.4621	0.5826	0.4075	0.4621	0.7388	0.6267	0.7920	0.8006	0.6134
	S_m	**0.8839**	0.7816	0.7251	0.7648	0.7233	0.6975	0.8431	0.8101	0.8671	0.8760	0.7605
	E_m	**0.9066**	0.8418	0.8152	0.8473	0.8613	0.8482	0.8744	0.8445	0.9115	0.9084	0.8024
VT1000	$\mathcal{M}$	**0.0179**	0.0664	0.1194	0.0896	0.1454	0.1167	0.0297	0.0339	0.0214	0.0215	0.0371
	F_β^{max}	**0.9426**	0.8297	0.7549	0.8066	0.7692	0.7659	0.9152	0.9230	0.9301	0.9301	0.8692
	F_β^{mean}	0.8823	0.8011	0.7148	0.7645	0.7173	0.7300	0.8526	0.8468	0.8826	0.8764	0.8189
	F_β^{w}	0.9057	0.7355	0.4854	0.6524	0.4628	0.4806	0.8587	0.8043	0.8834	0.8847	0.8104
	S_m	**0.9405**	0.8195	0.7054	0.7863	0.7258	0.6827	0.9124	0.9094	0.9205	0.9226	0.8734
	E_m	**0.9459**	0.8833	0.8360	0.8568	0.8280	0.8324	0.9286	0.9224	0.9401	0.9379	0.9150
VT5000	$\mathcal{M}$	**0.0295**	0.0847	0.1143	0.0886	0.1680	0.1091	0.0499	0.0482	0.0347	0.0377	0.0558
	F_β^{max}	**0.8984**	0.7253	0.6623	0.7374	0.6440	0.6646	0.8357	0.8633	0.8747	0.8715	0.7908
	F_β^{mean}	0.8400	0.6610	0.5952	0.6721	0.5751	0.6210	0.7608	0.7783	0.8197	0.8066	0.7289
	F_β^{w}	0.8352	0.5897	0.3968	0.5585	0.3271	0.3882	0.7520	0.7218	0.8064	0.8015	0.7057
	S_m	**0.8998**	0.7498	0.6793	0.7493	0.6520	0.6517	0.8515	0.8633	0.8739	0.8727	0.8136
	E_m	**0.9265**	0.8167	0.7952	0.8239	0.7820	0.8104	0.8800	0.8914	0.9177	0.9118	0.8659
Methods		**SSNet**	**MMCI**	**AFNet**	**PDNet**	**TANet**	**S2MA**	**JL-DCF**	**DBFI**	**C3A**	**CPD**	**MSCLDL**
Type		RGB-D	RGB-D	RGB-D	RGB-D	RGB-D	RGB-D	RGB-D	RGB-T	RGB-T	RGB	RGB
VT821	$\mathcal{M}$	0.0531	0.0887	0.0688	0.0567	0.0525	0.0983	0.0758	0.0324	0.0767	0.0573	0.0929
	F_β^{max}	0.8102	0.7229	0.7384	0.7773	0.7952	0.8042	0.8436	0.8631	0.8057	0.8008	0.7980
	F_β^{mean}	0.7382	0.6148	0.6610	0.7119	0.7153	0.6877	0.7263	**0.8552**	0.8074	0.7105	0.7292
	F_β^{w}	0.7279	0.5689	0.6240	0.6799	0.6825	0.6945	0.7198	**0.8552**	0.6554	0.6820	0.5520
	S_m	0.8356	0.7584	0.7786	0.8095	0.8161	0.8112	0.8388	0.8820	0.8200	0.8267	0.7487
	E_m	0.8554	0.7856	0.8208	0.8584	0.8571	0.8204	0.8487	0.9040	0.8449	0.8442	0.7797
VT1000	$\mathcal{M}$	0.0266	0.0404	0.0329	0.0327	0.0306	0.0302	0.0300	0.0179	0.0809	0.0316	0.0833
	F_β^{max}	0.9190	0.8755	0.8873	0.8950	0.8968	0.9219	0.9223	0.9240	0.8802	0.9034	0.8653
	F_β^{mean}	0.8474	0.7982	0.8376	0.8360	0.8365	0.8242	0.8288	**0.9140**	0.8048	0.8338	0.8029
	F_β^{w}	0.8667	0.7904	0.8285	0.8337	0.8375	0.8577	0.8458	**0.9143**	0.7278	0.8338	0.6447
	S_m	0.9149	0.8792	0.8889	0.8970	0.8987	0.9180	0.9128	0.9272	0.8878	0.9054	0.8052
	E_m	0.9260	0.8969	0.9225	0.9212	0.9236	0.9124	0.9145	0.9441	0.9004	0.9177	0.8058
VT5000	$\mathcal{M}$	0.0433	0.0565	0.0504	0.0474	0.0474	0.0536	0.0503	0.0320	0.0881	0.0502	0.0905
	F_β^{max}	0.8412	0.7970	0.8181	0.8302	0.8248	0.8328	0.8504	0.8764	0.7905	0.8262	0.7553
	F_β^{mean}	0.7795	0.7074	0.7481	0.7607	0.7507	0.7367	0.7385	**0.8692**	0.6790	0.7395	0.6737
	F_β^{w}	0.7664	0.6787	0.7264	0.7438	0.7364	0.7382	0.7447	**0.8681**	0.5965	0.7267	0.5457
	S_m	0.8581	0.8187	0.8314	0.8439	0.8413	0.8523	0.8607	0.8862	0.8365	0.8459	0.7575
	E_m	0.8918	0.8579	0.8792	0.8834	0.8834	0.8658	0.8638	0.9234	0.8860	0.8705	0.7716

boost performance compared to "MA", but both they are inferior to "MAAM". Since "MAAM" could further take both advantages of "MGA" and "TA" by our decision-level multi-modal integration. Comparing to "TA" that uses directly thermal infrared image as attention map, "MAAM" adopts decision-level integration strategy(i.e., constructing PWM map) to integrate RGB and thermal modalities, and makes attention map better play guiding role for the network, this strategy is more effective than the conventional RGB-D strategy. Also, we give visual comparison results in Fig.11 to demonstrate of the contribution the PWM map. Two modalities have their own strengths in different regions, any one of mask-guided attention map and thermal-guided attention map fails to enhance the decoder effectively. Integrating directly two attention maps as the final attention map might bring more noises, such as Fig.11(e). In contrast, introducing PWM map can learn pixel-wise weights of two components to produce more accurate saliency predictions, such as Fig.11(f). In summary, results in Table.2 and Fig.11 demonstrate adequately the superiority and effectiveness of the proposed MAAM.

4.2.3. The validation of the effectiveness of coarse-and-refined bidirectional optimization(CRBO)

We propose a coarse-and-refined bidirectional optimization(CRBO) to optimize the decoder output, also, we design other optimization strategies to compare with the proposed CRBO, results are shown Table.3. "P" represents that we use the proposed CRBO to optimize the decoder output D^1 to the final prediction P. It's seen that the CRBO could boost D^1 to more favorable results. In contrast, "$C^5 + D^1$" indicates the the direct element-wise addition of C^5(the convolution result of M^5) and D^1, its various evaluation metrics are inferior to "P". The advantage of "P" to "$C^5 + D^1$ " demonstrates adequately that the proposed CBRO could better fuse the coarse localization and the refined localization. To the best of our knowledge, this is due to the fact that both D^1 and C^5 are single-channel maps, simple integration might not produce too much performance improvement.

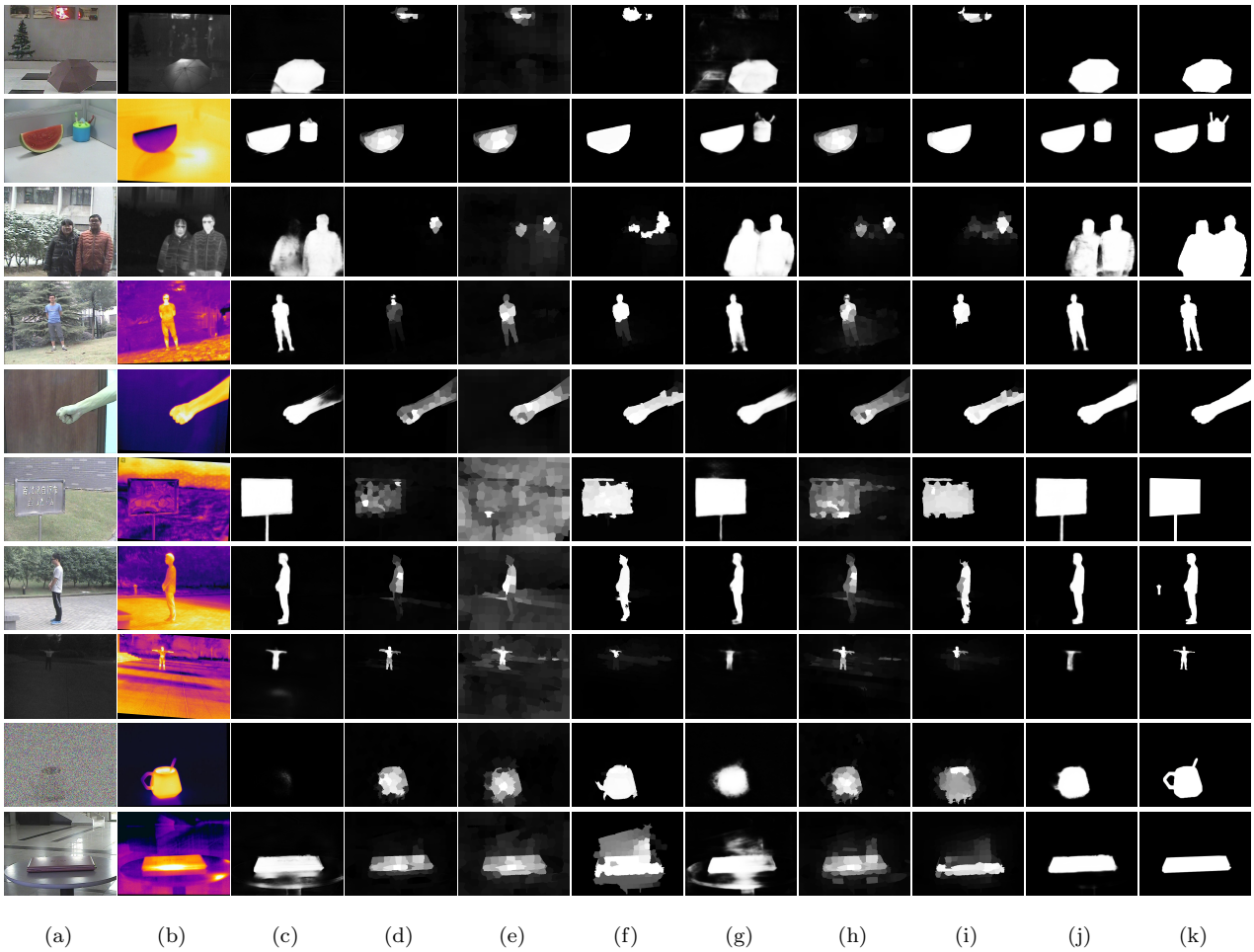


Figure 12: Visual comparison results on some images selecting from VT821, VT1000 and VT5000. (a) RGB images (b) Thermal infrared images (c) ADF (d) CRA (e) M3S-NIR (f) MGFL (g) MIED (h) MTMR (i) SDGL (j) MAANet (k) Ground truth

4.2.4. The validation of the effectiveness of VGG16 encoder

Our MAANet adopts VGG16 net as our backbone to extract powerful features from RGB-T images. Actually, other pre-trained backbones are also considered here, e.g., VGG19 net and ResNet. We show the comparison results in Table.4, the model's performance with VGG16 net as encoder is superior to others. Considering the classic performance of the self-attention mechanism [33] in computer vision, we also use its classic version to optimize each layer when using VGG16 as backbone and regard the optimized features as the input of transition layers. Results are also shown in Table.4, i.e., "VGG16-SA", but we find that its overall performance are still inferior to "VGG16". To the best of our analysis, this is due to the fact that the exploration of the non-local relationship between pixels relies on the encoder features trained on RGB image database(i.e., ImageNet) and ignores the important role of thermal stream, this misleads feature representation of RGB-T image. Therefore, we adopt VGG16 net as the final backbone in our MAANet.

4.3. Comparison experiments

We compare the proposed method with state-of-the-art methods on three RGB-T datasets, the compared RGB-T methods include:DBIF[37],C3A[38], MGFL [39], MTMR[24], SDGL[23], M3S-NIR[40], CRA[41], MIED[42], ADFNet[35], APNet[43], ECFENet[28] and FMCF[30]; The compared RGB-D methods include: SSNet[34], MMCI[44], AFNet[45], PDNet[46], TANet[47], S2MA[48] and JL-DCF[49]; The compared RGB methods include: CPD[50] and MSCLDL[17]. It's noticed that, for several RGB-D methods, we also train them on the VT5000 RGB-T dataset.

The comparison results are shown in Table.5. The proposed MAANet achieves the best overall performance on three datasets, only DBIF is superior slightly to our MAANet in a few metrics and DBIF is also a classic RGB-T saliency detection method which constructs bidirectional fusion strategy to achieve RGB-T integration, but ignoring the modality-aware mechanism makes bad modality easily mislead its final fusion results, therefore our MAANet still achieves the best overall performance. Among all RGB-T methods, MGFL, MTMR SDGL, M3S-NIR and CRA are

Table 6: The comparison results between the proposed MAANet and other RGB-D methods on three RGB-D datasets

Methods		SSNet	TANet	CPFP	MAANet
NLPR	$\mathcal{M}$	0.0281	0.0413	0.0383	**0.0145**
	F_β^{max}	0.9167	0.8765	0.8842	**0.9511**
	F_β^{mean}	0.8704	0.7960	0.8189	**0.9109**
	F_β^{w}	0.8621	0.7801	0.8077	**0.9078**
	S_m	0.9152	0.8862	0.8843	**0.9448**
	E_m	0.9495	0.9166	0.9201	**0.9612**
RGBD135	$\mathcal{M}$	0.0231	0.0465	0.0383	**0.0138**
	F_β^{max}	0.9280	0.8532	0.8821	**0.9668**
	F_β^{mean}	0.8993	0.7956	0.8292	**0.9300**
	F_β^{w}	0.8773	0.7404	0.7871	**0.9099**
	S_m	0.9241	0.8588	0.8725	**0.9532**
	E_m	0.9681	0.9192	0.9275	**0.9771**
NJUD	$\mathcal{M}$	0.0451	0.0613	0.0532	**0.0299**
	F_β^{max}	0.9107	0.8888	0.8902	**0.9502**
	F_β^{mean}	0.8714	0.8445	0.8371	**0.9132**
	F_β^{w}	0.8572	0.8054	0.8281	**0.8996**
	S_m	0.8991	0.8782	0.8780	**0.9307**
	E_m	0.9221	0.9093	0.9000	**0.9446**

traditional low-level features based methods while others are deep learning based methods. No matter which kind of strategies, performance are inferior obviously to MAANet. Since our proposed MAANet is the first to achieve RGB-T integration by using decision-level integration strategy to fully leverage the advantages of two modalities. Meanwhile, we also achieve better result than several RGB-D methods on three RGB-T datasets, it's emphasized that SSNet is also single-stream based CNN, but its performance is far from the proposed MAANet. To the best of our knowledge, SSNet regards directly additional modality (depth map) as attention map, and there is not the deep analysis about two modalities' own strengths, this is vital for RGB-T strategy. Finally, we also find that several RGB methods' results are worse than our method and other multi-modal methods, this also demonstrates the necessity of multi-modal integration in some extreme scenes. In summary, the comparison results of Table.5 validate adequately the superiority and effectiveness of the proposed MAANet.

4.4. Visual comparison

We select some example images from three RGB-T datasets and show the visual comparison results between the proposed MAANet and other RGB-T methods on these images in Fig.12. Two modalities have their own strengths in different types of scenes, e.g., RGB images are more discriminative than thermal infrared images in the first three rows, and thermal infrared images has greater contrast in the fourth, the fifth and the last two rows. We find that most RGB-T methods are difficult to identify salient objects accurately in all above scenes. This is due to the fact that existing RGB-T methods are usually based on two-stream network, i.e., integrating two modalities' detection results, thereby integration noises will be inevitable

(a) (b) (c) (d)

Figure 13: The visual results of the proposed MAANet in some RGB-D image pairs. (a)RGB images (b)Depth maps (c)MAANet (d)Ground truth

when any one modality works bad. In contrast, our proposed MAANet can take both advantages of two modalities by constructing decision-level integration strategy to suppress bad one of two modalities. Also, it's surprised to find that our proposed MAANet could generate clearer edges of salient objects while there are not tiny background noises in saliency maps generated by MAANet, since the proposed CRBO can solve above problems by integrating effectively the coarse and the refined saliency cues. Different from above scenes, for the sixth to the eighth rows, both two modalities are difficult to separate foreground and background completely. For any of these scenes, each modality only can capture the contrast between foreground and background in some regions. We find that MAANet still produces accurate detection results compared to other methods in these scenes, since our method is pixel-wise RGB-T integration which considers the advantages of two modalities in different regions. In summary, visual comparison results show that the proposed MAANet could generate outstanding detection results in different types of scenes.

4.5. Runtime and model size analysis

Based on the VT5000-test, model size of the trained MAANet is only 107 MB, which is very lightweight compared to other two-stream multi-modal methods with VGG net as basic structure. We compare it and several classic multi-modal methods, including SSNet[34], ADFNet [35], MIED [42], TANet[47] and PoolNet [51], results are shown in Table.7. ADF's model size is 320 MB, MIED's model size is 200 MB and TANet's model size is 929 MB, only SSNet is close to our method in model size. As an important pre-processing stage, lightweight module is very vital in saliency detection field. In addition, the average speed of the proposed MAANet is 32(FPS), the method with similar average speed is SSNet, a single-stream RGB-D method, also attains 32(FPS). However, our method's performance is superior obviously to it on the premise of same average speed, other methods' average speeds are also slower than our method, e.g., TANet's average speed is only 14(FPS). Above comparison and analysis illustrate

Table 7: The comparison results of module size

Module	MAANet	SSNet	ADFNet	MIED	TANet	PoolNet
Size(MB)	107	107	320	200	929	200.3

adequately that our method has attained top performance in module lightweight and runtime.

4.6. Application to the RGB-D task

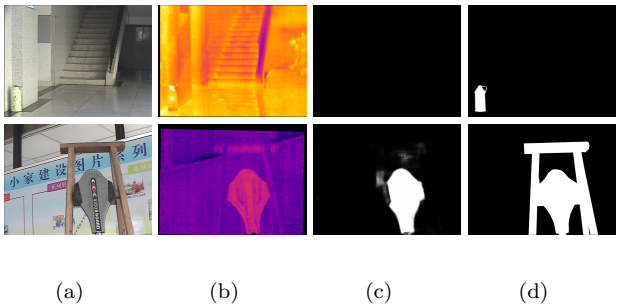
In order to emphasize the effectiveness and robustness of the proposed MAANet. We also compare it and several RGB-D methods on three RGB-D datasets, including NLPR [52], RGBD135 [53] and NJUD [54], the compared methods include SSNet [34], TANet [34] and CPFP [55]. For fairness, we adopt the training strategy of SSNet, i.e., we randomly select 1400 samples from the NJUD dataset and 650 samples from the NLPR dataset for training. Their remaining images and RGBD135 dataset are used for testing. For TANet and CPFP, we directly download corresponding saliency maps to compare.

Comparison results are shown in Table.6. We find that the proposed MAANet produces better results than several RGB-D methods in all evaluation metrics. To the best of our knowledge, the reason why our RGB-T method is still effective in several RGB-D datasets is: The constructed decision-level integration mechanism in our MAANet adopts learning mechanism to make additional modality (thermal or depth) to better guide the network, this operator is not limited by specific modality.

We also show the visual results of the proposed MAANet to some RGB-D image pairs, which are shown in Fig.13. For the first row, RGB image performs better than depth map, our detection result is not influenced by bad depth map, since our filtering mechanism hinders some false depth information. For the second row, two modalities have their own strengths in some regions, it's surprised to find that the proposed MAANet could take both advantages of two modalities and produce better detection results, this benefits from the pixel-wise decision-level integration strategy in the MAAM.

4.7. Limitation

Although our proposed MAANet achieves competitive performance in most cases, performance is still limited when both RGB and thermal infrared images cannot indicate clear semantic information. Such as Fig.14, for the first row, deeper semantic information needs to be explored from RGB image if we want to detect salient object(kettle), also, thermal infrared image fails to describe the contrast information in this time, thereby salient object is not marked by our MAANet. For the second row, both two modalities cannot capture effectively the contrast between foreground and background, this results that a part of salient object is mistaken as background in the final detection result. In summary, we think that our future works concentrate on exploring more powerful semantic information from the complicated RGB-T images, e.g., applying more complex mathematical models[56]-[58] to our works.



(a) (b) (c) (d)

Figure 14: Failure cases. (a)RGB images (b)Thermal infrared images (c)Saliency maps obtained by MAANet (d)Ground truth

5. Conclusion

This paper presents a novel CNN framework for RGB-T saliency detection, called MAANet, which is a encoder-decoder FCN structure. Based on the 4-channel input of RGB-T image pair, we firstly use VGG16 as backbone to extract powerful features. Then, a novel modality-aware adaptive-integration based attention mechanism (MAAM) between the encoder and the decoder is proposed, we fed MAAM into each decoder layer for generating more discriminative saliency maps. Finally, we also develop a coarse-and-refined bidirectional optimization (CRBO) framework to further suppress background noises and highlight salient objects. Experimental results have demonstrated the effectiveness and superiority of the proposed MAANet. Even, our MAANet also adapts various RGB-D datasets. In the future works, we concentrate on promoting the ability extracting powerful features from the complicated RGB-T images. In addition, the applications of the proposed MAAM to other multi-modal tasks also will be future research direction.

6. Acknowledgement

This work was supported in part by the National Natural Science Foundation of China under Grant nos. 62406202, the Natural Science Foundation of Liaoning Province under Grant nos.2024-BS-098, the Young Teacher Training Fund of Shenyang University of Technology 200005847, the Natural Science Foundation of Liaoning Province 2021-KF-12-01 and the Foundation of National Key Laboratory OEIP-O-202005, and .

References

[1] Y.Zhou, H.Su, T.Wang, Q.Hu. Onet: Twin U-Net architecture for unsupervised binary semantic segmentation in radar and remote sensing images, IEEE Trans. Image Process., 34, 2161-2172(2025) DOI: 10.1109/TIP.2025.3530816

[2] X.Xu, H.Liu, T.Zhang, H.Xiong, W.Yu. PreCM: The padding-based rotation equivariant convolution mode for semantic segmentation. IEEE Trans. Image Porcess, 34, 2781-2795(2025) DOI: 10.1109/TIP.2025.3558425

[3] Z.Xie, W.Zhang, B.Sheng, P.Li, C.L.P.Chen. BaGFN:Broad attentive graph fusion network for high-order feature interactions. IEEE Trans. Neural Networks Learn. Syst., 34(8), 4499-4513(2023) DOI: 10.1109/TNNLS.2021.3116209

[4] Y.Qin, N.Zhao, J.Yang, S.Pan, B.Sheng, R.W.H.Lau. UrbanEvolver:Function-aware urban layout regeneration. Int J Comput Vis., (2024) DOI:10.1007/s11263-024-02030-w

[5] K.Wang, X.Zhang, Y.Lu, W.Zhang, S.Huang, D.Yang. GSAL: Geometrics structure adversarial learning for robust medical image segmentation, Pattern Recognit., 140, 109596(2023) DOI:10.1016/j.patcog.2023.109596

[6] F.Meng, X.Gong, Y.Zhang. SiamRank: A siamse based visual tracking network with ranking strategy, Pattern Recognit., 141, 109630(2023) DOI:10.1016/j.patcog.2023.109630

[7] D.Liu, P.An, R.Ma, W.Zhan, X.Huang, A.Yahya. Content-based light field image compression method with Gaussion process regression, IEEE Trans. Multimedia., 22(4), 846-859(2019) DOI: 10.1109/TMM.2019.2934426

[8] Q.Hou, M.Cheng, X.Hu, A.Borji, Z.Tu, P.Torr. Deeply supervised salient object detection with short connections, IEEE Trans. Pattern Anal. Mach. Intell., 41(4), 815-828 (2019) DOI: 10.1109/TPAMI.2018.2815688

[9] X.Zhao, Y.Pang, L.Zhang, H.Lu, L.Zhang. Suppress and balance: A simple gated network for salient object detection, in Proc. Eur. Comput. Vis., pp.35-51(2020) DOI:10.1007/978-3-030-58536-5-3

[10] W.Wang, S.Zhao, J.Shen, S.C.Hoi, A.Borji. Salient object detection with pyramid attention and salient edges, in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp.1448-1457(2019) DOI: 10.1109/CVPR.2019.00154

[11] Q.Wang, Y.Liu, Z.Xiong, Y.Yuan. Hybrid Feature Aligned Network for Salient Object Detection in Optical Remote Sensing Imagery. IEEE Trans Geosc. Remot Sens., 60, 5624915(2022) DOI: 10.1109/TGRS.2022.3181062

[12] Y.Liu, Z.Xiong, Y.Yuan, Q.Wang. Distilling knowledge from super-resolution for efficient remote sensing salient object detection. IEEE Trans Geosc. Remot Sens., 61, 5609116(2023) DOI: 10.1109/TGRS.2023.3267271

[13] Y.Liu, Z.Xiong, Y.Yuan, Q.Wang. Transcending pixels: boosting saliency detection via scene understanding from aerial imagery. IEEE Trans Geosc. Remot Sens., 61, 5616416(2023) DOI: 10.1109/TGRS.2023.3298661

[14] Y.Liu, Y.Yuan, Q.Wang. Uncertainty-aware graph reasoning with global collaborative learning for remote sensing salient object detection. IEEE Geosc Remot Sens. Lett., 20, 6008105(2023) DOI: 10.1109/LGRS.2023.3299245

[15] H.Guan, J.Lin, R.Lau. A contrastive-learning framework for unsupervised salient object detection. IEEE Trans. Image Process., 34, 2487-2498(2025) DOI: 10.1109/TIP.2025.3558674

[16] K.Fu, D.Fan, G.Ji, Q.Zhao, J.Shen, C.Zhu. Siamese network for RGB-D salient object detection and beyond, IEEE Trans. Pattern Anal. Mach. Intell., 44(9), 5541-5559(2022) DOI: 10.1109/TPAMI.2021.3073689

[17] Y.Pang, C.Wu, H.Wu, X.Yu. Unsupervised multi-subclass saliency classification for salient object detection, IEEE Trans. Multimedia., 25, 2189-2202(2023) DOI: 10.1109/TMM.2022.3144070

[18] Y.Pang, C.Wu, H.Wu, X.Yu. Over-sampling strategy-based class-imbalanced salient object detection and its application in underwater scene, The Visual Comput., 39, 1959-1974(2023) DOI:10.1007/s00371-022-02458-6

[19] M.Cheng, N.J.Miltra, X.Huang, P.H.S.Torr, S.Hu. Global contrast based salient region detection, IEEE Trans. Pattern Anal. Mach. Intell., 37(3), 569-582(2015) DOI: 10.1109/TPAMI.2014.2345401

[20] Y.Pang, X.Yu, Y.Wu, C.Wu, Y.Jiang. Bagging-based saliency distribution learning for visual saliency detection, Signal Process.-Image Commun., 87, 115928(2020) DOI:10.1016/j.image.2020.115928

[21] F.Huang, Q.Jinqing, H.Lu, L.Zhang, X.Ruan. Salient object detection via multiple instance learning, IEEE Trans. Image Process., 8(7), 1911-1922(2017) DOI: 10.1109/TIP.2017.2669878

[22] Q.Zhang, Z.Huo, Y.Liu, Y.Pan, C.Shan, J.Han. Salient object detection employing a local tree-structured low-rank representation and foreground consistency, Pattern Recognit., 92, 119-134(2019) DOI:10.1016/j.patcog.2019.03.023

[23] Z.Tu, T.Xia, C.Li, X.Wang, Y.Ma, J.Tang. RGB-T image saliency detection via collaborative graph learning, IEEE Trans. Multimedia., 22(1), 160-173(2020) DOI: 10.1109/TMM.2019.2924578

[24] C.Li, G.Wang, Y.Ma, A.Zheng, B.Luo, J.Tang. RGB-T saliency detection benchmark: Dataset, baselines, analysis and a novel approach, in Proc. Chin Conf. Image. Graph Technol., 52, 131-142(2018) DOI:10.1007/978-981-13-1702-6-36

[25] Z.Tu, T.Xia, C.Li, X.Wang, Y.Ma, J.Tang. RGB-T image saliency detection via collaborative graph learning, IEEE Trans. Multimedia., 22(1), 160-173(2020) DOI: 10.1109/TMM.2019.2924578

[26] Y.Ma, D.Sun, Q.Meng, Z.Ding, C.Li. Learning multiscale deep features and SVM regressors for adaptive RGB-T saliency detection, in Proc. Int. Symp. Comput. Intell. Design., pp.389-392(2017) DOI: 10.1109/ISCID.2017.92

[27] Q.Zhang, T.Xiao, N.Huang, D.Zhang, J.Han. Revisiting feature fusion for RGB-T salient object detection, IEEE Trans. Circuits Syst. Video Technol., 31(5), 1804-1818(2021) DOI: 10.1109/TCSVT.2020.3014663

[28] W.Zhou, Q.Guo, J.Lei, L.Yu, J-N.Hwang. ECFFNet: Effective and consistent feature fusion network for RGB-T salient object detection, IEEE Trans. Circuits Syst. Video Technol., 32(3), 1224-1235(2022) DOI: 10.1109/TCSVT.2021.3077058

[29] J.Wang, K.Song, Y.Bao, L.Huang, Y.Yan. CGFNet: Cross-guided fusion network for RGB-T salient object detection. IEEE Trans. Circuits Syst. Video Technol., 32(5), 2949-2961(2022) DOI: 10.1109/TCSVT.2021.3099120

[30] Q.Zhang, N.Huang, L.Yao, D.Zhang, C.Shan, J.Han. RGB-T salient object detection via fusing multi-level CNN features, IEEE Trans. Image Process., 29, 3321-3335(2020) DOI: 10.1109/TIP.2019.2959253

[31] T.Lin, P.Dollar, R.Girshick, K.He, B.Hariharan, S.Belongie. Feature pyramid networks for object detection, in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp.2117-2125(2017) DOI: 10.1109/CVPR.2017.106

[32] K.Simonyan, A. Zisserman. Very deep convolutional networks for large-scale image recognition, in Proc. Int. Conf. Learn. Representat., pp.1-14(2015) arXiv:1409.1556

[33] A.Vaswani, N.Shazeer, N.Parmar, J.Uszkoreit, L.Jones, A N.Gomez, L.Kaiser, I.Polosukhin. Attention is all you need. in Proc. Neur Info. Process Sys., pp.1-11(2017) arXiv:1706.03762

[34] X.Zhao, L.Zhang, Y.Pang, H.Lu, L.Zhang. A single stream network for robust and real-time RGB-D salient object detection, in Proc. Eur. Comput. Vis., pp.646-662(2020) DOI:10.1007/978-3-030-58542-6-39

[35] Z.Tu, Y.Ma, Z.Li, C.Li, J.Xu, Y.Liu. RGBT salient object detection: A large-scale dataset and benchmark, arXiv preprint (2020) arXiv:2007.03262

[36] W.Liu, A.Rabinovich, A.C.Berg. Parsenet: Looking wider to see better. arXiv preprint, 9(2015) arXiv:156.04579

[37] Z.Xie, F.Shao, G.Chen, H.Chen, Q.Jiang, X.Meng, Y-S.Ho. Cross-modality double bidirectional interaction and fusion network for RGB-T salient object detection, IEEE Trans. Circuits Syst. Video Technol., 33(8), 4149-4163(2023) DOI: 10.1109/TCSVT.2023.3241196

[38] Y.Pang, H.Wu, C.Wu. Cross-modal co-feedback cellular automata for RGB-T saliency detection. Pattern Recognit., 135, 109138(2023) DOI:10.1016/j.patcog.2022.109138

[39] L.Huang, K.Song, J.Wang, M.Niu, Y.Yan. Multi-graph fusion and learning for RGBT image saliency fusion, IEEE Trans. Circuits Syst. Video Technol., 32(3), 1366-1377(2022) DOI: 10.1109/TCSVT.2021.3069812

[40] Z.Tu, T.Xia, C.Li, Y.Lu, J.Tang. M3S-NIR: Multi-modal multi-scale noise insensitive ranking for RGB-T saliency detection, in Proc. IEEE Conf. Multimedia inf Process. Retr., pp.141-146(2019) DOI:10.1109/mipr.2019.00032

[41] J.Tang, D.Fan, X.Wang, Z.Tu, C.Li. RGBT salient object detection: benchmark and a novel cooperative ranking approach, IEEE Trans. Circuits Syst. Video Technol., 30(12), 4421-4433(2020) DOI: 10.1109/TCSVT.2019.2951621

[42] Z.Tu, Z.Li, C.Li, Y.Lang, J.Tang. Multi-interactive Encoder-decoder Network for RGBT Salient Object Detection, arXiv preprint (2020) arXiv:2005.02315

[43] W.Zhou, Y.Zhu, J.Lei, J.Wan, L.Yu. APNet: Adversarial-learning-assistance and perceived importance fusion network for all-day RGB-T salient object detection, IEEE Trans. Emerg Top. Comput Intell., 6(4), 957-968(2022) DOI:10.1109/tetci.2021.3118043

[44] H.Chen, Y.Li, D.Su. Multi-modal fusion network with multi-scale multi-path and cross-modal interactions for RGB-D salient object detection, Pattern Recognit., 86, 376-385(2019) DOI:10.1016/j.patcog.2018.08.007

[45] N.Wang, X.Gong. Adaptive fusion for RGB-D salient objec detection, IEEE ACCESS., 7, 55277-55284(2019) DOI: 10.1109/ACCESS.2019.2913107

[46] C.Zhu, X.Cai, K.Huang, T.Li, G.Li. PDNet: Prior-model guided depth-enhanced network for salient object detection, in Proc. IEEE Int. Conf. Multimedia Expo., pp.199-204(2019) DOI:10.1109/icme.2019.00042

[47] H.Chen, Y.Li. Three-stream attention-aware network for RGB-D salient object detection, IEEE Trans. Image Process., 28(6), 2825-2835(2019) DOI: 10.1109/tip.2019.2891104

[48] N.Liu, N.Zhang, J.Han. Learning selective self-mutual attention for RGB-D saliency detection, in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp.13756-13765(2020) DOI:10.1109/cvpr42600.2020.01377

[49] K.Fu, D.Fan, G.Ji, Q.Zhao. JL-DCF: Joint learning and densely-cooperative fusion framework for RGB-D salient object detection, in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp.3049-3059(2020) DOI:10.1109/cvpr42600.2020.00312

[50] Z.Wu, L.Su, Q.Huang. Cascaded partial decoder for fast and accurate salient object detection, in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp.3902-3911(2019) DOI:10.1109/cvpr.2019.00403

[51] J.Liu, Q.Hou, M.Cheng, J.Feng, J.Jiang. A simple pooling-based design for real-time salient object detection, in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp.3912-3921(2019) DOI:10.1109/cvpr.2019.00404

[52] H.Peng, B.Li, W.Xiong, W.Hu, R.Ji. RGBD salient object detection: A benchmark and algorithms, in Proc. Eur. Comput. Vis., pp.92-109(2014) DOI:10.1007/978-3-319-10578-9-7

[53] Y.Cheng, H.Fu, X.Wei, J.Xiao, X.Cao. Depth enhanced saliency detection method, in Proc. Int Conf. Int. Multimedia.Comput Serv., 23(2014) DOI:10.1145/2632856.2632866

[54] R.Ju, L.Ge, W.Geng, T.Ren, G.Wu. Depth saliency based on anisotropic center-surround difference, in Proc Int Conf. Image Process., pp.1115-1119(2014) DOI:10.1109/icip.2014.7025222

[55] J.Zhao, Y.Cao, D.Fan, M.Cheng, X.Li, L.Zhang. Contrast prior and fluid pyramid integration for RGBD salient object detection, in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp.3927-3936(2019) DOI:10.1109/cvpr.2019.00405

[56] L.Guo, C.Li, N.Qiao, J.Zhao. Convergence analysis of positive solution for Caputo-Hadamard fractional differential equation, Nonlinear Anal. Model Control., 30(3), 212-230(2025) DOI:10.15388/namc.2025.30.38509

[57] L.Guo, H.Liu, C.Li, J.Zhao, J.Xu. Existence of positive solutions for singular p-Laplacian Hadamard fractional differential equations with the derivative term contained in the nonlinear term., 28(3), 491-515(2023) DOI:10.15388/namc.2023.28.31728

[58] L.Guo, Y.Wang, C.Li, J.Cai, B.Zhang. Solvability for a higher-order Hadamard fractional differential model with a sign-changing nonlinearity dependent on the parameter. J.Appl Anal Comput., 14(5), 2762-2776 (2024) DOI:10.11948/20230389