RGB-T Salient Object Detection via Fusing Multi-level CNN Features

Qiang Zhang, Nianchang Huang, Lin Yao, Dingwen Zhang,
Caifeng Shan, *Senior Member, IEEE,* and Jungong Han

***Abstract*—RGB-induced salient object detection has recently witnessed substantial progress, which is attributed to the superior feature learning capability of deep convolutional neural networks (CNNs). However, such detections suffer from challenging scenarios characterized by cluttered backgrounds, low-light conditions and variations in illumination. Instead of improving RGB based saliency detection, this paper takes advantage of the complementary benefits of RGB and thermal infrared images. Specifically, we propose a novel end-to-end network for multi-modal salient object detection, which turns the challenge of RGB-T saliency detection to a CNN feature fusion problem. To this end, a backbone network (e.g., VGG-16) is first adopted to extract the coarse features from each RGB or thermal infrared image individually, and then several adjacent-depth feature combination (ADFC) modules are designed to extract multi-level refined features for each single-modal input image, considering that features captured at different depths differ in semantic information and visual details. Subsequently, a multi-branch group fusion (MGF) module is employed to capture the cross-modal features by fusing those features from ADFC modules for a RGB-T image pair at each level. Finally, a joint attention guided bi-directional message passing (JABMP) module undertakes the task of saliency prediction via integrating the multi-level fused features from MGF modules. Experimental results on several public RGB-T salient object detection datasets demonstrate the superiorities of our proposed algorithm over the state-of-the-art approaches, especially under challenging conditions, such as poor illumination, complex background and low contrast.**

***Index Terms*—RGB-T salient object detection, Adjacent-depth feature combination, Multi-branch group fusion, Joint attention guided bi-directional message passing**

I. Introduction

SALIENT object detection aims to identify the most visually distinctive objects or regions in an image, and has attracted lots of attention in recent years. As a preprocessing step, salient object detection plays a critical role in many computer vision tasks, including visual tracking [1], [2], recognition [3], [4], content based image compression [5], [6], image fusion [7], [8] and so on.

While numerous salient object detection methods have been presented [9]–[25], most of them are designed for RGB images only, which may fail to distinguish salient objects from backgrounds when being exposed to challenging conditions, such as poor illumination, complex background, and low contrast. To address such issues, we advocate a multi-modal salient object detection method by fusing RGB and thermal infrared (RGB-T) images considering the popularity of thermal infrared sensors. More specifically, we present an end-to-end RGB-T salient detection model by using the recently developed deep convolutional neural networks (CNNs).

Unlike RGB cameras, thermal infrared cameras are a kind of passive sensors that capture the thermal infrared radiation emitted by all objects with a temperature above absolute zeros, meaning that thermal infrared images are invariant to illumination conditions [26]–[28]. As a result, when applied to salient object detection, thermal images tend to provide additional saliency cues to boost the saliency detection performance. Fig. 1 illustrates the validity of integrating RGB-T images for salient object detection under challenging conditions.

In fact, RGB-T images have shown significant superiorities over RGB images in many computer vision tasks, such as face recognition [29] and video surveillance [26], [30]. Motivated by that, a few works have already exploited RGB-T images to boost the saliency detection performance. For example, Li *et al.* [27] presented a robust salient object detection method based on multi-task manifold ranking with cross-modality consistency. Although the experimental results demonstrated its performance superiority over the traditional RGB-induced saliency detection methods, using low-level hand-crafted features might be a bottleneck for further performance improvement in [27]. In [28], a deep CNNs based RGB-T salient object detection method was developed, in which the saliency map of each modality, i.e., RGB and thermal infrared, was independently induced at the first stage. Afterwards, the saliency map for each RGB or thermal infrared image was first independently induced by using the deep CNNs. Then the saliency maps of these two modalities were fused to derive the final saliency map. However, such a fusion at the saliency map level does not seem to well explore the complementary information/features across RGB and thermal images. In addition, the deep CNNs in [28] were pre-trained for image classification on ImageNet dataset [31], rather than for salient object detection, meaning that the saliency cues might not be well explored.

The work was supported in part by National Natural Science Foundation of China under Grants No. 61773301, 61876140 and 61971004, and by China Postdoctoral Support Scheme for Innovative Talents under Grant No. BX20180236. (Corresponding author: Jungong Han)

Q. Zhang is with Key Laboratory of Electronic Equipment Structure Design, Ministry of Education, Xidian University, Xi'an Shaanxi 710071, China, and also with Center for Complex Systems, School of Mechano-Electronic Engineering, Xidian University, Xi'an Shaanxi 710071, China. Email: qzhang@xidian.edu.cn.

N. Huang, L. Yao, and D. Zhang are with Center for Complex Systems, School of Mechano-Electronic Engineering, Xidian University, Xi'an Shaanxi 710071, China.

C. Shan is with Philips Research, High Tech Campus, 5656 AE, Eindhoven, The Netherlands.

J. Han is with WMG Data Science, University of Warwick, Coventry, CV4 7AL, U. K. Email: jungonghan77@gmail.com

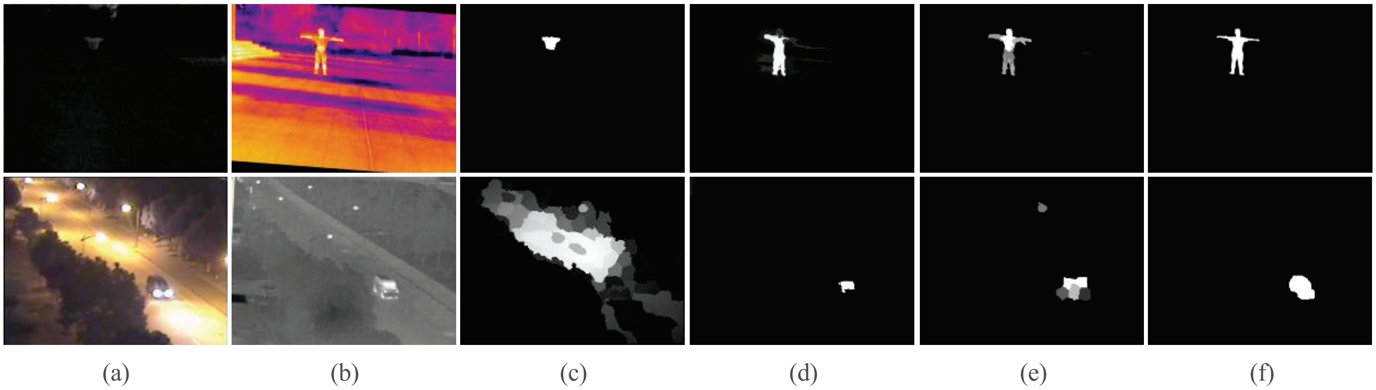


Fig. 1. Illustration of the validity for salient object detection by integrating RGB-T images. (a) RGB images; (b) Thermal infrared images; (c) Saliency maps induced from the RGB images by [16]; (d) Saliency maps induced from the thermal images by [16]; (e) Saliency maps induced from the RGB-T images by [27]; (f) Ground truth. The saliency maps in (c), (d) and (e) clearly demonstrate that thermal infrared images can provide complementary saliency cues for RGB images under challenging scenes with poor illumination (the first row) or complex background (the second row).

In this paper, rather than integrating RGB-T information only at the saliency map level [28], which might be already too late, we propose a novel end-to-end deep neural network for RGB-T salient object detection by fusing multi-modal information at various stages. The proposed method turns the challenging RGB-T salient object detection into a CNN feature fusion problem, which covers the following three subproblems: 1) How to effectively extract the single-modal features from the input RGB or thermal infrared images; 2) How to fuse the extracted multi-modal features in a comprehensive way; and 3) How to infer the final saliency map using the fused features.

To address the first problem, we adopt a backbone network (i.e., VGG-16 net [32] or Res-Net [33]) to extract the features from each single-modal input image. Afterwards, an adjacent-depth features combination (ADFC) module is employed to capture the multi-level features of single-modal images, considering that different-depth features capture varieties of semantic information and fine visual details.

With respect to the second problem, motivated by the group convolution in [34], a multi-branch group fusion (MGF) module is put in place to fuse the features of RGB-T image pairs, which consists of two branches at each level. One branch contains several paths via group convolutions to reduce the network parameters while the other branch has just one path to capture the wholly cross-modal features. As a result, MGF module is expected to capture the cross-modal features between RGB and thermal infrared images effectively but at a considerably lower computational complexity.

For the third problem, we introduce a joint attention guided bi-directional message passing (JABMP) module for saliency prediction via integrating the multi-level fused features obtained from MGF modules. With the proposed JABMP module, high-level semantic information in deeper layers will be passed to shallower layers, and low-level spatial details contained in shallower layers will also be passed to deeper layers. Accordingly, the cross-level complementary information among the fused features will be well captured by using the proposed module. Moreover, a joint channel-spatial attention (JCSA) block, different from the gate function in [35], is adopted to control the message passing in JABMP module. By using JCSA, some important features with higher channel attention as well as spatial attention will be selected and propagated to the next level, and some superfluous features will be suppressed during the message passing, which will enhance the feature discriminability for the final RGB-T saliency prediction.

In summary, the main contributions of this work are as follows:

1) An end-to-end CNN based RGB-T salient object detection method is proposed, which achieves the state-of-the-art performance on several datasets, including [26], [27]. To the best of our knowledge, it is the first end-to-end CNN established for RGB-T salient object detection.

2) An ADFC module is dedicated to extract each single-modal image features. By using multiple ADFC modules, multi-level features of input images containing rich spatial details as well as semantic information, rather than one specific level of features as in the traditional methods [36]–[38], are extracted for the subsequent fusion and saliency prediction modules.

3) A MGF module, instead of the simple concatenation, is presented to capture the cross-modal complementary information between each RGB-T image pairs and reduce the number of network parameters.

4) In order to effectively capture the cross-level complementary information among the fused features, a JABMP module is employed for the final saliency prediction in the proposed network. Especially, a JCSA block, rather than a gate function as in [35], is adopted to control the message passing in the proposed bi-directional message passing module.

The rest of the paper is organized as follows. Section II briefly reviews some related work, and Section III illustrates the proposed multi-modal salient object detection model in detail. Experimental results and conclusions are given in Section IV and Section V, respectively.

II. Related work

A. RGB salient object detection

Over the past two decades, a considerable number of RGB salient object detection methods have been developed [11]–[25]. Early salient object detection methods utilized low-level hand-crafted features with specific statistical priors, such as color contrast [11], [12], object prior [13], [14], and background prior [15]–[17], to model and approximate human saliency. A complete survey on RGB salient object detection methods is beyond the scope of this paper and we refer the readers to a recent survey paper [18] for details.

Recently, to extract more sophisticated features, tremendous deep learning based saliency detectors have been proposed [19]–[25], [39]–[41], and achieved substantially better performance than those previous methods. For example, Lee *et al.* [23] proposed to first encode low-level distance map and high-level sematic features of deep CNNs to form a new feature vector, and then evaluate saliency by a multi-level fully connected neural network classifier. Hou *et al.* [24] presented a salient object detection method by introducing a series of short connections between shallower and deeper side-output layers. Zhang *et al.* [25] introduced a generic aggregating multi-level convolutional feature framework for salient object detection, which first integrated multi-level feature maps into multiple resolutions and then adaptively learned to combine these feature maps at each resolution to predict the saliency maps. In [39], two pooling based modules, i.e., a global guidance module (GGM) and a feature aggregation module (FAM), aided to improve the performance for salient object detection. A novel recurrent residual refinement network (R3Net) equipped with residual refinement blocks (RRBs) was presented in [40] to detect salient regions from an input image more accurately.

However, most of these studies focus on the RGB salient object detection. Under some challenging conditions, such as poor illumination, complex background or low contrast, these models may fail to distinguish salient objects from backgrounds, as shown in Fig. 1 (c).

B. RGB-D salient object detection

As a departure from RGB images, depth images provide affluent spatial structures and 3D information for salient objects and backgrounds, which benefit the salient object detection [42]. Therefore, RGB-D salient object detection has attracted much attention in recent years. So far, a variety of RGB-D salient object detection models have been presented to boost the performance of saliency detection [42]–[48]. For examples, Chen *et al.* [42] presented a progressively complementarity-aware fusion network for RGB-D salient object detection by adding the cross-level complementarity in the process of cross-modal fusion. In [45], two saliency maps were first pre-deduced from the source RGB and depth images via a two-streamed CNN , respectively. Then a switch map was generated by using a saliency fusion module to adaptively fuse the two saliency maps. In [44], the depth information was first enhanced by utilizing contrast prior into a CNNs based architecture. Then the enhanced depth cues were integrated with RGB features for salient object detection by using a fluid pyramid integration.

C. RGB-T salient object detection

Recently, considering their complementary benefits, a few works also attempted to exploit RGB-T images to boost the saliency detection performance. For example, Li *et al.* [27] proposed a robust multi-task manifold ranking based RGB-T salient object detection method with cross-modality consistency. Ma *et al.* [28] presented an adaptive RGB-T saliency detection method by learning multiscale deep CNN features and SVM regressors. In [49], a novel collaborative graph learning algorithm was presented for RGB-T image saliency detection. Specifically, superpixels were taken as graph nodes, and hierarchical deep features were collaboratively used to jointly learn the graph affinity and node saliency in a unified optimization framework.

Although CNN based RGB-T salient object detection algorithms are not well investigated yet, a large number of deep neural networks with RGB-T inputs have been presented for some other computer vision or image processing tasks, such as pedestrian detection [36]–[38], image fusion [50], object tracking [51]–[53]. For example, Wagner *et al.* [37] presented an RGB-T pedestrian detection method by fusing information with CNNs, where information from the RGB and thermal infrared images was integrated via an early-fusion and a late-fusion based CNN architecture. In addition to early-fusion (also called low-level fusion in [38]) and late-fusion (also called high-level fusion in [38]), another two CNN architectures for information fusion, i.e., middle-level fusion and confidence-level fusion, were explored for RGB-T pedestrian detection. Their experimental results revealed that the middle-level fusion provides the best performance among the four fusion models on RGB-T pedestrian detection.

III. Proposed RGB-T salient object detection model with multi-level CNN feature fusion

In this section, we will discuss the proposed RGB-T salient object detection model in detail. Fig. 2 shows the diagram of the proposed network, which is composed of three components: single-modal image feature extraction, multi-modal image feature fusion, and saliency map prediction. These will be described in detail in the following subsections.

A. Multi-level feature extraction for each single-modal image using ADFC modules

The RGB-T salient object detection network may be incorporated with any basic network, such as VGG-16 net [32] and Res-Net [33]. Here, we employ the VGG-16 net as the backbone network to carry out the feature extraction from RGB and thermal infrared branches, which is well known for its elegance, simplicity, and good generalization. For saliency detection, we make two modifications on the VGG-16 net i.e.,removal of all the fully-connected layers and skip of the

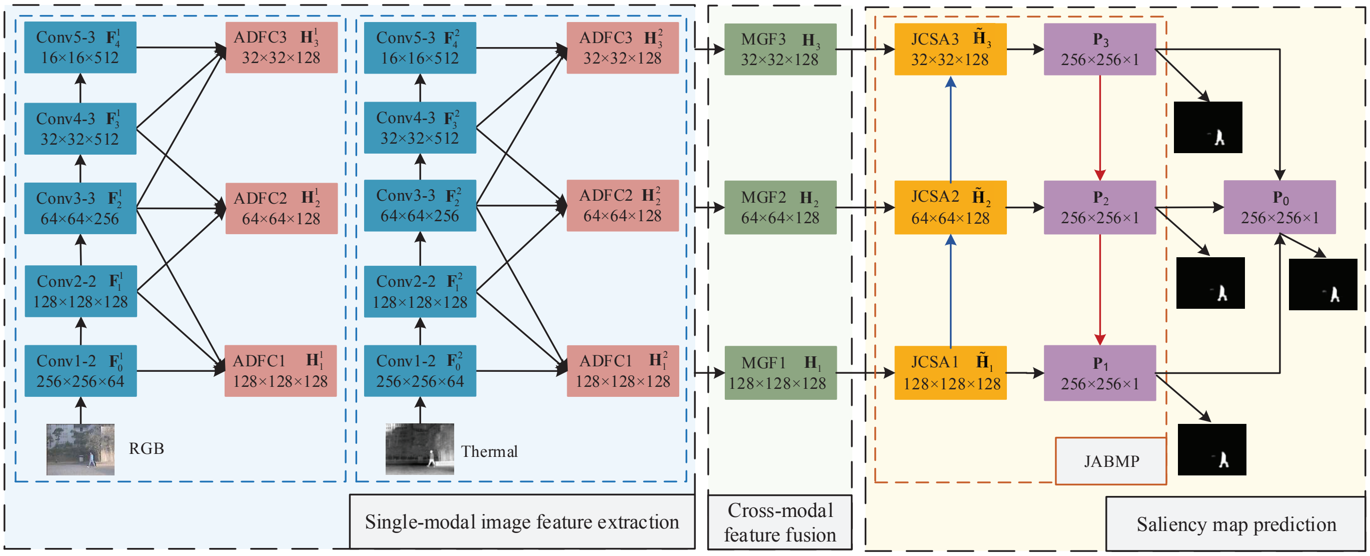


Fig. 2. The overall architecture of our proposed RGB-T salient object detection. Each colorful box is considered as a feature block. The solid arrows between blocks indicate the information streams. RGB-T input images are assumed to have been well registered in advance and have been rescaled to the fixed sizes (e.g., 256 × 256 in this paper). The RGB image and thermal infrared image are first fed into the backbone network, i.e., VGG-16 net, respectively. Based on that, multi-level features are further generated for each RGB or thermal infrared image by using the proposed ADFC modules. Then the features from ADFC modules that correspond to the same level in the two branches are fused by using MGF module. After that, JABMP module is performed on the fused multi-level features to obtain the final saliency map, where a JCSA block is adopted to control the message passing.

pool5 layer to maintain more spatial information for the input image. The modified VGG-16 net includes five convolutional blocks.

After the RGB image or thermal infrared image is fed into the backbone network, features at different levels/depths are extracted for each single-modal input image, which capture various semantic information and visual details. Shallower-level features contain more visual details but lack some semantic information, while deeper levels of features carry more semantic information but are limited when it comes to details. Therefore, features from different levels in the backbone networks are complementary to each other.

In this work, we propose an adjacent-depth feature combination (ADFC) module to integrate the multi-level features of single-modal images. We select some middle-level features from the backbone networks for each single-modal input image by using multiple proposed ADFC modules to obtain multi-level features and reduce the burden of network parameters. More specifically, we first extract five layers of RGB or thermal infrared image features from different depths of VGG-16 net: conv1-2 (containing 64 feature maps of size 256×256, denoted by $\mathbf{F}_0^n$) , conv2-2(128 feature maps of size 128×128, denoted by $\mathbf{F}_1^n$), conv3-3(256 feature maps of size 64×64, denoted by $\mathbf{F}_2^n$), conv4-3 (256 feature maps of size 32×32, denoted by $\mathbf{F}_3^n$) and conv5-3 (512 feature maps of size 16×16, denoted by $\mathbf{F}_4^n$). Here, n=1 or 2 denotes RGB or thermal infrared image, respectively.

Given the five layers of features, three levels of new features $\{\mathbf{H}_d^n | d = 1, 2, 3; n = 1, 2\}$ for each single-modal RGB or thermal infrared image are obtained by using the proposed ADFC modules, where d denotes the level number. The three levels of features $\{\mathbf{H}_d^n | d = 1, 2, 3; n = 1, 2\}$ contain the same number of channels but vary in spatial resolutions. And each level feature $\mathbf{H}_d^n$ is constructed by the following three layers of features from the VGG-16 net, i.e., $\mathbf{F}_{d-1}^n$, $\mathbf{F}_d^n$ and $\mathbf{F}_{d+1}^n$.

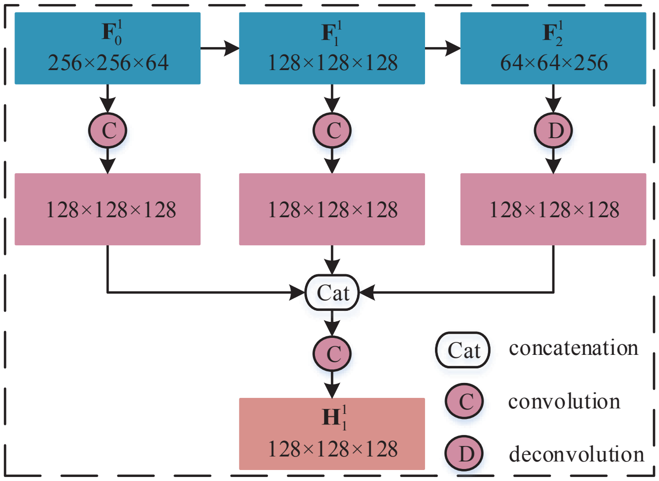


Fig. 3. Generation of the first level features $\mathbf{H}_1^1$ for RGB image by using ADFC module.

Fig. 3 illustrates the details of ADFC module for generating the first level of RGB features $\mathbf{H}_1^1$. Likewise, other levels of features can be generated in a similar way. As shown in Fig. 3, each ADFC module contains three convolutional blocks and one deconvolutional block. More specifically, for the d-th level, a 3×3 convolutional layer $\mathrm{C}\big(*, \theta_d^{n,1}, 2\big)$,[1] a 1×1 convolutional layer $\mathrm{C}\big(*, \theta_d^{n,2}, 1\big)$, and a 2×2 deconvolutional layer $\mathrm{D}\big(*, \gamma_d^n, 1/2\big)$ are first performed on $\mathbf{F}_{d-1}^n$, $\mathbf{F}_d^n$ and $\mathbf{F}_{d+1}^n$ respectively, to ensure the outputs of the three layers have the same number of channels (i.e., 128) and the same spatial

[1] The symbol $\mathrm{C}\big(*,\theta,k\ \big)$ denotes a convolutional layer with pixel stride k and network parameters θ. The same symbol is used in the rest of the paper.

resolutions (e.g., 128×128 for *d*=1). Then, the features from the three layers are concatenated. Finally, a 1×1 convolutional layer $\mathrm{C}(*, \theta_d^{n,3}, 1)$ is performed on the concatenated features to obtain the *d-th* level of features $\mathbf{H}_d^n$. Mathematically, ADFC module can be expressed by

$$\mathbf{H}_d^n = \phi\big(\mathrm{C}\big(\mathrm{Cat}\big(\phi\big(\mathrm{C}\big(\mathbf{F}_{d-1}^n; \theta_d^{n,1}, 2\big)\big), \phi\big(\mathrm{C}\big(\mathbf{F}_d^n; \theta_d^{n,2}, 1\big)\big), \phi\big(\mathrm{D}\big(\mathbf{F}_{d+1}^n; \gamma_d^n, 1/2\big)\big)\big); \theta_d^{n,3}, 1\big)\big), \quad (1)$$

where $\mathrm{Cat}(\cdot)$ denotes the cross-channel concatenation, and $\phi(\cdot)$ is a ReLU activation function [54].

As discussed above, the *d*-th level of constructed features $\mathbf{H}_d^n$ contains the features $\mathbf{F}_d^n$ as well as those from its adjacent layers $\mathbf{F}_{d-1}^n$ and $\mathbf{F}_{d+1}^n$, which means that $\mathbf{H}_d^n$ contains more rich and accurate semantics because it integrates different-resolution information. In addition, $\mathbf{H}_d^n$ has much less amount of data than the simple combination of $\mathbf{F}_{d-1}^n$,$\mathbf{F}_d^n$ and $\mathbf{F}_{d+1}^n$. As a result, the redundancy among $\mathbf{F}_{d-1}^n$,$\mathbf{F}_d^n$ and $\mathbf{F}_{d+1}^n$ are reduced from $\mathbf{H}_d^n$ by using the proposed ADFC module.

Finally, it should be noted that the idea of ADFC is similar to those of the feature pyramid network (FPN) [55] and the hierarchical feature integration mechanism (HIFI) [56]. All of the three modules investigate the integration of multi-level features to improve the saliency detection performance. But in HIFI, the features from all convolutional layers of different levels are integrated. While, in ADFC and FPN, only the features from the last convolutional layers of different levels are integrated, considering that the features from the deepest layer of each level are the strongest [55]. As a result, HIFI has much more to-be-learned parameters than ADFC and FPN, and thus has higher computational complexity. In addition, the integration of large numbers of features by HIFI will also introduce much more redundant information and degrade the subsequent saliency detection performance. In FPN, only the features from two adjacent levels (i.e., the current level and its deeper level) are integrated. Differently, in ADFC, features from three adjacent levels (i.e., the current level, its deeper level and its shallower level) are integrated. Accordingly, more spatial details will be captured by ADFC than by FPN, which will improve the subsequent salient detection results. This will be verified in the later experimental part.

B. Fusion of multi-modal image features using MGF modules

Given the features $\{\mathbf{H}_d^n | d = 1, 2, 3; n = 1, 2\}$ of RGB and thermal infrared images generated from ADFC modules, most of existing models capture the cross-modal features between the two modalities by first simply concatenating $\mathbf{H}_d^1$ and $\mathbf{H}_d^2$ at the same level d and then performing the Conv+ReLU operators on the concatenated features, as shown in Fig. 4. This may well capture the wholly cross-modal correlations among the concatenated features. However, some salient features from individual single-modal image may also be drowned in the concatenated features because of the large numbers of features, which will diminish the discriminability of the subsequent fused features [57]. In addition, under the premise of multi-level features, direct concatenation may increase the network parameters, which is not desirable for the training of multi-modal methods.

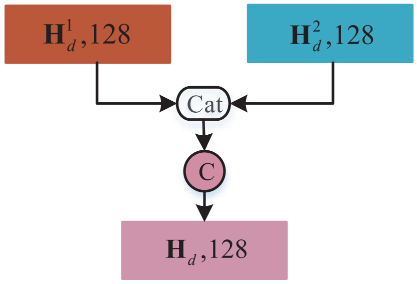


Fig. 4. Traditional feature fusion module.

Such problems can be solved by using group convolution, which may date back to the AlexNet [58] or even earlier and is supported by Caffe [59], PyTorch [60], and so on. The basic idea behind group convolution is split-transform-merge, similar to the Inception models [61]–[63]. In group convolution, the input features are first divided into a few small groups along the channel. Then, a set of regular 3×3 or 5×5 convolutions are performed on these small groups. All of the outputs from these small groups are concatenated as the final output. But different from Inception models, where each path may be carefully customized, group convolution shares the same topology among all the paths. More specifically, Xie *et al.* [34] proposed to use stacked group convolutions in the process of transformation to reduce complexity and model sizes. Although group convolution may greatly reduce parameters, it can only capture the partly cross-modal correlations among the features within the same group, which may weaken the cross-modal correlations among all of the feature maps.

In this work, we propose a multi-branch group fusion (MGF) module to fuse the features $\{\mathbf{H}_d^1 | d = 1, 2, 3\}$ and $\{\mathbf{H}_d^2 | d = 1, 2, 3\}$. MGF is expected to effectively capture the cross-modal features between RGB and thermal infrared images but at a considerably lower computational complexity.

As shown in Fig. 5, the proposed MGF consists of two branches for feature fusion at each level. One branch (named multi-group fusion branch) has *M* (e.g., *M*=8 in this paper) paths via group convolutions to reduce the network parameters while the other branch (named single-group fusion branch) has just one path to capture the wholly cross-modal features as in the traditional fusion module in Fig. 4. The two branches produce the same number of feature maps (e.g., 64 in this paper), so the number of the finally fused feature maps is doubled (e.g., 128 in this paper).

The details of the proposed MGF module are described as follows. In the multi-group fusion branch, the input single-modal features $\mathbf{H}_d^1, \mathbf{H}_d^2$ are first divided into *M* groups along the channel, resulting in two sets of features $\{\mathbf{H}_{d,m}^1 | m = 1, 2, ..., M\}$ and $\{\mathbf{H}_{d,m}^2 | m = 1, 2, ..., M\}$ with the same number of channels (i.e., $8/M$), respectively. Then the features $\mathbf{H}_{d,m}^1$ and $\mathbf{H}_{d,m}^2$ with the same group index *m* are concatenated and fused by two stacked convolution layers , i.e., a 1×1 convolutional layer with $64/M$ channels followed by a 3×3 convolutional layer with $64/M$ channels. Both layers

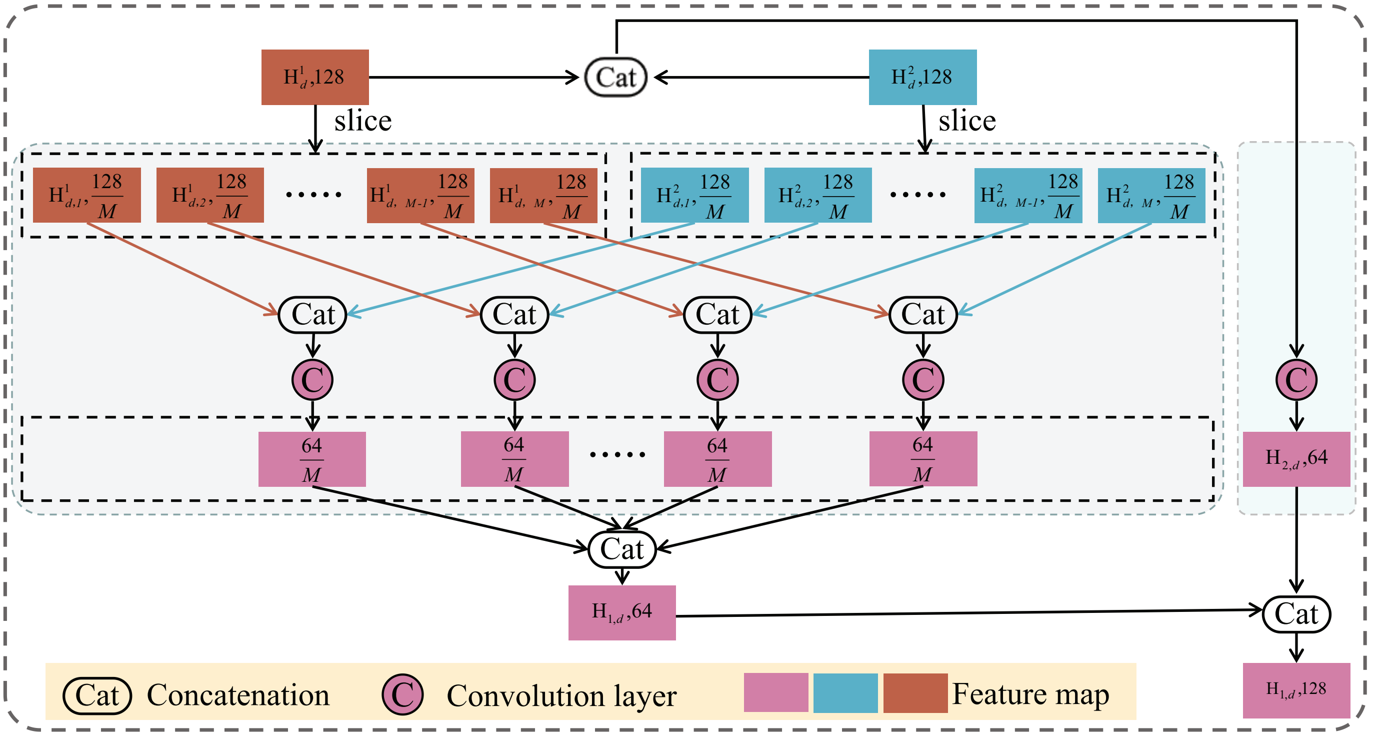


Fig. 5. Illustration of the proposed MGF module. The left part is the multi-group fusion branch, where the concatenated input features are first divided into M groups along the channel and then fused at each group by using several convolutional layers. The right part is the single-group fusion branch, where two stacked regular convolutional layers are directly performed on the concatenated input features to obtain the fused features. The finally fused features are obtained by concatenating the outputs from the two branches.

adopt a ReLU activation function. Finally, the outputs from the M groups are concatenated to obtain the fused features $\mathbf{H}_{1,d}$ via the multi-group fusion branch. Mathematically, the multi-group fusion branch is expressed as

$$\begin{aligned}\mathbf{H}_{1,d} = \mathrm{Cat}\big(&\mathrm{Z}_1\big(\mathrm{Cat}\big(\mathbf{H}^1_{d,1}, \mathbf{H}^2_{d,1}\big); \varphi_{d,1}\big), ..., \\ &\mathrm{Z}_\mathrm{M}\big(\mathrm{Cat}\big(\mathbf{H}^1_{d,M}, \mathbf{H}^2_{d,M}\big); \varphi_{d,M}\big)\big),\end{aligned} \tag{2}$$

where $\mathrm{Z}_\mathrm{m}\left(*, \varphi_{d,m}\right)$ denotes the stacked convolutions with ReLU activation function mentioned above, and $\varphi_{d,m}$ denotes the network parameters in the m-th path.

The single-group fusion branch in MGF module can be seen as a special case of the multi-group fusion branch with $M = 1$. Therefore, the single-group fusion branch can be simply expressed by

$$\mathbf{H}_{2,d} = \mathrm{Z}^*\big(\mathrm{Cat}\big(\mathbf{H}^1_d, \mathbf{H}^2_d\big); \varphi^*_d\big), \tag{3}$$

where $\mathbf{H}_{2,d}$ is the d-th level of fused features from the single-group fusion branch, and $\mathrm{Z}^*\left(*, \varphi^*_d\right)$ consists of two stacked convolution layers (a 1×1 convolutional layer with 64 channels followed by a 3×3 convolutional layer with 64 channels). Similarly, the two convolutional layers also have a ReLU activation function. φ^*_d denotes the network parameters for Z^*.

The final fused features $\mathbf{H}_d$ for the d-th level are obtained by simply concatenating $\mathbf{H}_{1,d}$ and, $\mathbf{H}_{2,d}$ i.e.,

$$\mathbf{H}_d = \mathrm{Cat}\big(\mathbf{H}_{1,d}, \mathbf{H}_{2,d}\big). \tag{4}$$

As discussed above, MGF module can capture the wholly cross-modal correlations among the features of RGB-T images via the single-group fusion branch. As well, it can extract more salient features from each single-modal input image via the multi-group fusion branch. As a result, the proposed MGF module can potentially better capture the cross-modal features of RGB-T images than those exiting fusion methods [36]–[38]. By using multiple MGF modules, different levels of fused features containing semantic information as well as visual details can be extracted for RGB-T salient object detection. More importantly, due to the employed group convolution, MGF module requires much fewer network parameters[2] than the traditional fusion method shown in Fig. 4, which first concatenates $\mathbf{H}^1_d$ and, $\mathbf{H}^2_d$ and then performs a 1×1 convolutional layer with 128 channels and a 3×3 convolutional layer with 128 channels.

C. Saliency map prediction using JABMP module

With multiple MGF modules, three levels of fused features $\{\mathbf{H}_d | d = 1, 2, 3\}$ are obtained, which will be used to predict the final saliency map. A straightforward method is to perform the side output on each level $\mathbf{H}_d$ independently, and then derive the final saliency map by adding a new convolutional layer

[2] Assume that 128-channels of fused features are generated by using two sets of 128 channels of single-modal features. The number of parameters in traditional fusion method is $a_1 = (128 + 128) \times 128 \times 1 \times 1 + 128 \times 128 \times 3 \times 3 = 180224$, and the number of parameters in MGF is $a_2 = 8 \times (128 + 128)/8 \times 64/8 \times 1 \times 1 + 64/8 \times 3 \times 3 + 64/8 \times 1 \times 1 = 54080$. For each level fusion, the number of parameters is reduced by about 130000.

to fuse these side outputs of different levels. Although this method can detect salient objects with features at different levels, the inner correlations among different levels of features are missing. As a result, the prediction may not be optimal, and some post-processing may be further needed as in [64]. To facilitate the interaction among multiple predictions, a series of connections from deeper side output layers to shallower ones were suggested in [24]. This method only considered the information transmitted from deeper layers to shallower ones, but ignored the information flow from shallower layers to deeper ones. Thus, the deep side outputs still lack the low-level details contained in the shallow layers. For that, Zhang *et al.* [25] proposed a bi-directional message passing module for salient object detection by concatenating feature maps from both high levels and low levels. However, their module just used simple concatenation operations to integrate multi-level features without considering their importance. As the multi-level features are not always equally useful for every input image, this aggregation method would lead to information redundancy. Considering that, a gated bi-directional message passing module was presented in [35], where a gate function was employed to transmit the useful features and suppress the superfluous features.

Inspired by the work in [35], we propose a joint attention guided bi-directional message passing (JABMP) module for saliency map prediction by effectively integrating the multi-level features from MGF modules. The module can capture the cross-level complementarity among the fused features in two directions. With the proposed module, high-level semantic information in deeper layers will be passed to shallower layers and low-level spatial details contained in shallower layers will also be passed to deeper layers. As is well known, each convolutional layer has multiple channels of feature maps. But not all these channels are effective for saliency prediction. Similarly, in each feature map, features from different spatial positions may play different roles in salient object detection. Therefore, a joint channel-spatial attention (JCSA) block, instead of a gate function as in [35], is introduced to control the message passing in the proposed JABMP module.

Fig. 6 illustrates the architecture of the proposed JABMP module, which consists of two directional connections. One is the bottom-up information stream, where the features from the current level and the weighted features from the previous level are integrated to produce the current level of attentive features via the JCSA block. The other is the top-down information stream, where we hierarchically propagate the predictions from higher-level to lower-level to obtain more accurate side outputs. Next,we discuss each step in detail.

1) Bottom-up information stream with joint channel-spatial attention: Given the multi-level fused features $\{\mathbf{H}_d | d = 1, 2, 3\}$ from MGF modules, the attentive features at different levels are sequentially generated by using the introduced JCSA block. Mathematically, the process of the message passing from shallower layer to deeper layer is described by

$$\widetilde{\mathbf{H}}_d = \begin{cases} \mathrm{A}\left(\left(\mathbf{H}_d + \phi\left(\mathrm{C}\left(\widetilde{\mathbf{H}}_{d-1};\widetilde{\theta}_d, 2\right)\right)\right), s, \tau\right), d = 2, 3 \\ \mathrm{A}\left(\mathbf{H}_d, s, \tau\right), d = 1 \end{cases}, \tag{5}$$

where $\mathrm{C}\left(*, \widetilde{\theta}\_d, 2\right)$ denotes a 3×3 convolutional layer to ensure the adjacent-level features have the same number of channels (i.e., 128) and the same spatial resolutions. $\phi\left(\cdot\right)$ is a ReLU activation. $\mathrm{A}\left(*, s, \tau\right)$ is the joint attention function to weight the features. $s = \left[s_1, s_2, ..., s_Q\right]^T \in R^Q$ is a set of channel-wise weights, and $\tau \in R^{W \times H}$ denotes the importance of each local spatial position in the feature maps. Q, W, and H represent the number of channels, width and length of the input features for the JCSA block, respectively. Detailed implementation of JCSA will be described as follows.

As shown in the bottom of Fig. 6, the proposed JCSA block consists of a “Squeeze-and-Excitation” (SE) block [65] and a “Spatial Attention” (SA) block [66]. The SE block reflects the global channel-wise importance of each feature map by introducing some channel-dependent weights, and the SA block indicates the local spatial importance of features by introducing some position-dependent weights.

Suppose that the input features $\mathbf{H} = \left[\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_Q\right] \in R^{W \times H \times Q}$ for JCSA contain Q channels of feature maps, and $\mathbf{h}_q \in R^{W \times H}$ is the q-th feature map . $\widetilde{\mathbf{H}} = \left[\widetilde{\mathbf{h}}_1, \widetilde{\mathbf{h}}_2, ..., \widetilde{\mathbf{h}}_Q\right] \in R^{W \times H \times Q}$ is the output of JCSA, i.e., $\widetilde{\mathbf{H}} = \mathrm{A}\left(\mathbf{H}, s, \tau\right)$.

Similar to that in [65] , a global average pooling is first performed on $\mathbf{H}$ to generate a set of channel-wise statistical features $v = \left[v_1, v_2, ..., v_Q\right]^T \in R^Q$ in the SE block. Then two fully connected (FC) layers and a simple sigmoid activation function are performed on v, and a set of channel-wise weights s are obtained. The output $\mathbf{H}' = \left[\mathbf{h}'_1, \mathbf{h}'_2, ..., \mathbf{h}'_Q\right] \in R^{W \times H \times Q}$ of the SE block can be obtained by the following channel-wise multiplication

$$\mathbf{h}'_q = s_q \times \mathbf{h}_q, \tag{6}$$

where s_q is the q-th element of s, and $\mathbf{h}'_q$ is the q-th feature map in $\mathbf{H}'$.

The output $\mathbf{H}'$ of the SE block is further fed into the subsequent SA block. More specifically, in the SA block, a 1×1 convolutional operation and a simple sigmoid activation function are performed on $\mathbf{H}'$ to obtain the spatial weight map τ. Then the output of the SA block, i.e., the final output $\widetilde{\mathbf{H}} = \left[\widetilde{\mathbf{h}}_1, \widetilde{\mathbf{h}}_2, ..., \widetilde{\mathbf{h}}_Q\right]$ of JCSA block, is obtained by the element-wise product between each feature map in $\mathbf{H}'$ and the spatial weights (or importance) , i.e.,

$$\widetilde{\mathbf{h}}_q = \tau \circ \mathbf{h}'_q, \tag{7}$$

where $\circ$ represents element-wise product, and $\widetilde{\mathbf{h}}_q$ is the q-th feature map in $\widetilde{\mathbf{H}}$.

As shown in Eq. (6) and Eq. (7), some important features with higher channel attention (or weights) as well as spatial attention (or weights) will be selected and transmitted to the next level, and some superfluous features will be suppressed in the bottom-up information stream by using the proposed

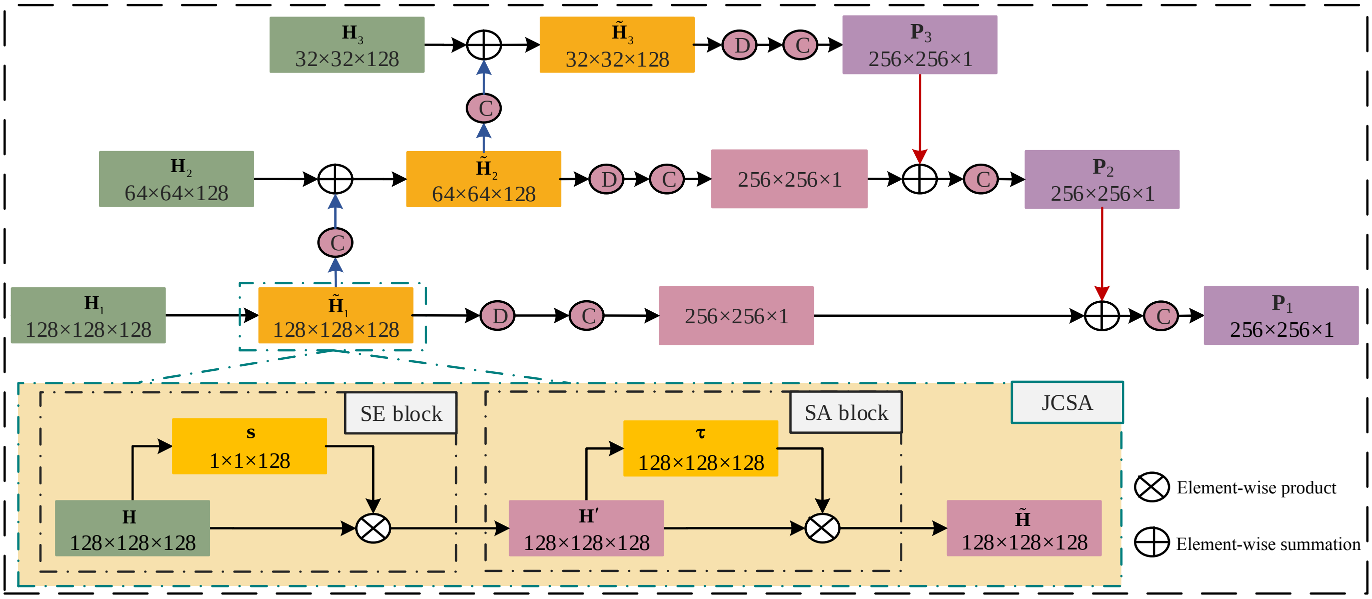


Fig. 6. Illustration of the proposed JABMP module. This module consists of two directional connections. One is the bottom-up information stream and the other is the top-down information stream. In the bottom-up stream, the features from the current level and the weighted features from the previous level are integrated to produce the current level of attentive features via a joint channel-spatial attention (JCSA) block. In the top-down stream, the predictions from higher-level to lower-level are hierarchically propagated to obtain more accurate saliency detection results.

JCSA block. This will boost the discriminability of the fused features when predicting the saliency maps.

2) Top-down information stream: Having obtained the attentive features $\left\{\widetilde{\mathbf{H}}_d | d=1,2,3\right\}$, the multiple side outputs $\{\mathbf{P}_d | d=1,2,3\}$ for different levels can be sequentially obtained in a deep-to-shallow manner, i.e.,

$$\mathbf{P}_d = \begin{cases} \mathrm{C}\Big(\big(\phi\big(\mathrm{C}\big(\mathrm{D}\big(\widetilde{\mathbf{H}}_d;\gamma_d,(1/2)^d\big);\theta_d^1,1\big)\big)+\mathbf{P}_{d+1}\big);\theta_d^2,1\Big), \\ \hfill d=1,2 \\ \mathrm{C}\left(\mathrm{D}\left(\widetilde{\mathbf{H}}_d;\gamma_d,(1/2)^d\right);\theta_d^2,1\right), d=3 \end{cases} \tag{8}$$

where $\mathrm{D}\left(*;\gamma_d,(1/2)^d\right)$ is a $2^d \times 2^d$ deconvolutional layer to ensure the features to be fused have the same spatial resolutions, $\mathrm{C}\left(*;\theta_d^1,1\right)$ and $\mathrm{C}\left(*;\theta_d^2,1\right)$ denote two 1×1 convolutional layers, which are used to fuse features and obtain side outputs, respectively. It should be noted that all of the side outputs $\{\mathbf{P}_d | d=1,2,3\}$ have the same spatial resolutions as those of the input images because of the employed deconvolution layers.

3) Saliency map prediction: Let $\{\mathbf{S}_d | d=1,2,3\}$ denote the side output maps, and they can be computed by $\mathbf{S}_d = \sigma\left(\mathbf{P}_d\right)$. $\sigma\left(\cdot\right)$ is a sigmoid activation function.

These side outputs are further fused to obtain the fusion output $\mathbf{P}_0$ by using a 1×1 convolutional layer $\mathrm{C}\left(*;\theta^0,1\right)$ i.e.,

$$\mathbf{P}_0 = \mathrm{C}\left(\mathrm{Cat}\left(\mathbf{P}_1,\mathbf{P}_2,\mathbf{P}_3\right);\theta^0,1\right). \tag{9}$$

Thus, the fusion output map $\mathbf{S}$ can be computed by $\mathbf{S}_0 = \sigma\left(\mathbf{P}_0\right)$, and we take $\mathbf{S}_0$ as the final saliency map of our model. Then, the proposed model is trained end to end using the cross-entropy loss L between the ground truth and the predicted results $\{\mathbf{S}_t | t=0,1,2,3\}$, which is defined as [67]

$$\begin{aligned} L = &-\beta \sum_{i=0}^{3}\sum_{i,j}\mathbf{G}\left(i,j\right)\log\left(\mathbf{S}_t\left(i,j\right)\right) \\ &-\left(1-\beta\right)\sum_{i=0}^{3}\sum_{i,j}\left(1-\mathbf{G}\left(i,j\right)\right)\log\left(1-\mathbf{S}_t\left(i,j\right)\right), \end{aligned} \tag{10}$$

where $\mathbf{G}\left(i,j\right) \in \{0,1\}$ is the label of the pixel (i,j) in the ground truth, and $\mathbf{S}_t\left(i,j\right)$ is the probability of pixel (i,j) belonging to the foreground in the predicted saliency map $\mathbf{S}_t$. To increase the detection accuracy for salient objects of various sizes, a class-balancing weight β is used to balance the foreground and background, and is set to the ratio of the number of background pixels to that of all the pixels in the ground truth.

IV. Experiments

In this section, we first describe the experimental setup and the employed evaluation metrics. Afterwards, we compare the proposed RGB-T salient object detection model with the state-of-the-art (SOTA) methods on some publicly available datasets. Finally, we perform several sets of ablation experiments to show the validity of each component in our proposed saliency detection model.

A. Experimental setup

1) Datasets: We train and evaluate our approach on three public datasets, including RGB-thermal dataset [27], Grayscale-thermal dataset [26] and MSRA-B dataset [68].

RGB-thermal dataset [27] contains 821 aligned RGB-T image pairs under different conditions to ensure the diversity and richness of data scenarios, such as multiple salient objects, low illumination, and similar appearances.

Grayscale-thermal dataset [26] includes 25 aligned grayscale-thermal video pairs with high diversity. However, this dataset is collected for object tracking rather than for salient object detection. For object tracking, only moving objects need to be detected, while for salient object detection, stationary objects may also be salient. Moreover, moving objects sometimes are too small or occluded so that they may not be salient objects. Considering these, we selected 843 pairs of images from the dataset in our experiments, which can be divided into two sets, i.e., a pedestrian set with 537 frame pairs from 10 aligned video pairs and a car set with 306 frame pairs from another 10 aligned video pairs.

MSRA-B dataset [68] contains 5000 RGB images (2500 images as training set and 2500 images as testing set) and is widely used for single-modal image salient object detection.

2) Training: We start with the backbone VGG-16 nets in our proposed model, whose convolutional layers are initialized with the weights that are pre-trained on the ImageNet dataset [31]. Then we adopt a 3-step training strategy to ensure that our proposed network is converged quickly. First, we train the RGB branch by using the cross entropy loss between the predicted saliency map and the ground truth. For that, we remove the MGF modules from the whole network, i.e., the outputs from the ADFC modules in the RGB branch model are directly fed into the JCSA block in the proposed JABMP module for saliency prediction. Secondly, we train the thermal infrared branch in a similar way as that in the training of RGB branch. Finally, the whole multi-modal salient object detection model is trained, where the network parameters for each single-modal feature extraction branch, including the VGG-16 net and the ADFC modules, are initialized by using their corresponding pre-trained ones in the first two steps.

Due to the lack of large RGB-thermal image datasets, we have to use different training data during the network training, which is similar to that in [36] [37]. Concretely, we randomly select 410 RGB-thermal image pairs from the RGB-thermal dataset and 200 RGB-thermal image pairs from the selected Grayscale-thermal dataset (i.e., the car and pedestrian sets) as the training set. The rest of RGB-thermal image pairs in RGB-thermal dataset and Grayscale-thermal dataset are used as the testing set. Then, 830 samples are randomly selected from the training set of the MSRA-B as an auxiliary set to train the RGB/thermal branch of the proposed model.

Subsequently, in the first training step, the RGB images in the training set and those in the auxiliary set are employed to fine-tune the RGB branch model. In the second training step, the thermal images in the training set and the red color channels of RGB images in the auxiliary set are used to fine-tune the thermal infrared branch model. In the third training step, the RGB-thermal image pairs in training set are used to fine-tune the whole multi-modal saliency detection network.

3) Implementation: The proposed network model is implemented on the MATLAB R2014b platform with the Caffe toolbox [59] and a NVIDIA 1080Ti GPU (with 11G memory). The stochastic gradient descent (SGD) method is adopted to train the proposed network with a momentum 0.9 and a weight decay 0.0001. The base learning rate is set to 10^{-8}, and then turned into a tenth of the previous set when the training loss reaches a flat. During training and testing, all the input images are rescaled to the spatial resolution of 256×256.

B. Evaluation metrics

We adopt six widely used metrics [11]–[25], including the precision-recall (PR) curves, F-measure curves, average F-measure (F_{ave}), mean absolute error (*MAE*), S-measure (S_α) [69] and weighted F_β^ω-measure (F_β^ω) [70], to objectively evaluate different saliency detection models.

Given a predicted saliency map $\mathbf{S}$ of size $W \times H$, a binary mask $\mathbf{B}$ is first obtained by using a threshold. Then precision and recall can be, respectively, computed by $Precsion = |\mathbf{B} \cap \mathbf{G}|/|\mathbf{B}|$ and $Recall = |\mathbf{B} \cap \mathbf{G}|/|\mathbf{G}|$, where $\mathbf{G}$ is the ground-truth and $|\cdot|$ denotes the non-zero entries in a mask. F-measure is a weighted harmonic mean of precision and recall, and is defined by

$$F_\beta = \frac{\left(1+\beta^2\right) \times Precision \times Recall}{\beta^2 \times Precision + Recall}, \tag{11}$$

where β^2 is set to 0.3 as suggested in [16] and [25]. With different thresholds, the PR and F-Measure curves are thus obtained. F_{ave} is the mean of all the F_β values obtained by different thresholds.

Weighted F_β^ω-measure (F_β^ω) is an intuitive generalization of the F-measure, which is computed by

$$F_\beta^\omega = (1+\beta^2) * \frac{Precision^\omega \cdot Recall^\omega}{\beta^2 \cdot Precision^\omega + Recall^\omega}, \tag{12}$$

where $Precision^\omega$ and $Recall^\omega$ are weighted precision and recall, respectively. Here, β^2 is also set to 0.3 as default. More details are seen in [70].

MAE is computed by

$$MAE = \frac{1}{W \times H} \sum_{i=1}^{W} \sum_{j=1}^{H} |\mathbf{S}\,(i,j) - \mathbf{G}\,(i,j)\,|. \tag{13}$$

S-measure (S_α) [69] is employed for the important structure information evaluation, which combines a region-aware (S_r) and an object-aware (S_o) structural similarity as their final structure metric:

$$S_\alpha = \alpha * S_o + (1-\alpha) * S_r, \tag{14}$$

where $\alpha \in [0,1]$ is the balance parameter and is set to 0.5 as default.

C. Comparison with the state-of-the-art methods

To validate the proposed RGB-T salient detection model, we compare our model with 10 SOTA methods, which are further divided into three types, i.e., (1)RGB salient object detection methods: PoolNet [39], R3Net [40], and CPDNet [41]; (2) RGB-D salient object detection methods: AFNet [45], TSAA [46], PDNet [47], and SSRC [48]; and (3) RGB-T salient object detection methods: MFSR [28], GCL [49], and MRCM [27].

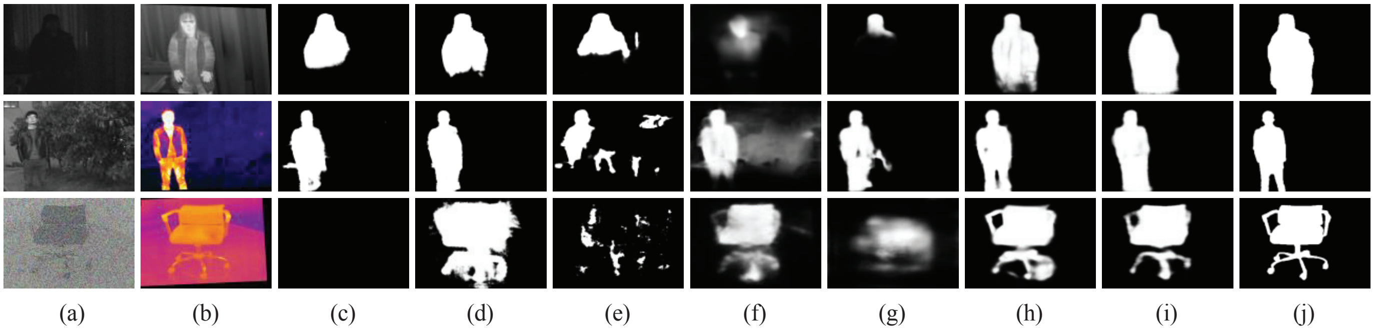


Fig. 7. Illustrations of the saliency detection results by some RGB methods and their modified multi-modal versions. (a) RGB images; (b) Thermal infrared images; (c) and (d) Saliency maps for RGB and RGB-T images obtained by R3Net and R3Net+, respectively; (e) and (f) Saliency maps for RGB and RGB-T images obtained by PoolNet and PoolNet+, respectively; (g) and (h) Saliency maps for RGB and RGB-T images obtained by CPDNet and CPDNet+, respectively; (i) Saliency maps for RGB-T images obtained by our proposed method; (j) Ground truth.

For fair comparisons, we modify these RGB and RGB-D salient object detection methods for RGB-T saliency detection. For those RGB methods, their original networks are seen as single-modal branches for RGB or thermal infrared image feature extraction. The outputs before the final saliency predictions in these networks for RGB and thermal infrared images are first concatenated and then fed into the saliency prediction layers to obtain the final multi-modal saliency maps. Those RGB-D methods are also re-trained for RGB-T saliency detection, where the input channels of depth images are replaced by the thermal images. These multi-modal versions (PoolNet+, R3Net+, and CPDNet+, for short, respectively) modified from the RGB models and those re-trained RGB-D models are fine-tuned in the same way as described in Subsection IV-A.

1) Visual Evaluation: Fig. 7 illustrates some saliency detection results by those RGB methods and their multi-modal versions. As shown in Fig. 7, compared with those single-modal salient detection methods for RGB images, their multi-modal counterparts generally perform better, especially when the input RGB images have poor visual quality (e.g, low contrast or much more noise). This further indicates that the use of the complementary information between the input RGB-T image pairs can improve the saliency detection performance.

Fig. 8 illustrates some visual comparisons of different methods in various cases, including large objects, small objects, simple background, complex background, poor illumination, and low contrast. As shown in the first four rows of Fig. 8 , under the condition of simple background and sufficient illumination, most of these methods work well for small salient objects, but poorly for large objects. Differently, our method can detect objects of various sizes effectively. Comparing the results in the fifth and sixth rows, we can also see that our method can effectively capture saliency information from RGB-T images when the infrared image or the RGB image has low contrast. As shown in the last four rows of Fig. 8, for those images with poor illumination and complex background, some SOTA methods cannot achieve desirable results. For example, some salient objects are not uniformly detected and even mistakenly detected. Parts of the backgrounds are not well suppressed during the saliency detection. In contrast, our proposed method still works well for these images. This may be attributed to the good collaborations among the different modules in our proposed network, i.e., ADFC for multi-level feature extraction of each single-modal image, MGF for cross-modal feature fusion of RGB-T images, and JABMP for final saliency prediction.

2) Quantitative Evaluation: PR and F-measure curves of different methods are shown in Fig. 9[3]. F_{ave}, F_{β}^{ω}, S_{α} and MAE values of different methods are listed in Table I. In Table I, the type of ‘RGB’ means that these methods are specifically designed for RGB salient object detection, which have been fine-tuned by using the RGB images in our RGB-T datasets. The evaluation values for these methods are obtained by performing these RGB salient object detection methods just on the RGB images in our RGB-T datasets. The type of ‘RGB→RGB-T’ means that these methods are modified from the RGB salient detection methods, as discussed previously. The type of ‘RGB-D→RGB-T’ means that these methods are re-trained versions of those RGB-D salient detection methods. Finally, the type of ‘RGB-T’ means that these methods are designed for RGB-T salient object detection. As shown in Fig. 9, the proposed method scores the best on both PR and F-measure curves among these methods. Similar conclusions can also be drawn from Table I. It can also be found from Table I that their multi-modal versions of those RGB salient object detection methods significantly outperform their original versions that are designed for RGB images. In addition, the processing time of different methods is also provided in Table I, which indicates that the computational complexity of the proposed method is acceptable. Especially, it has the highest computational efficiency among the four RGB-T salient detection methods.

D. Ablation analysis

1) Validity of multi-level feature extraction using ADFC: As shown in Fig. 3, multiple levels of features can be extracted from each single-modal input image by using several ADFC modules. In order to verify the validity of multiple levels of extracted features for saliency detection, we compare our proposal with another six schemes with different numbers of

[3] PR and F-measure curves of the RGB salient detection methods are not provided in Fig. 9 for better displaying.

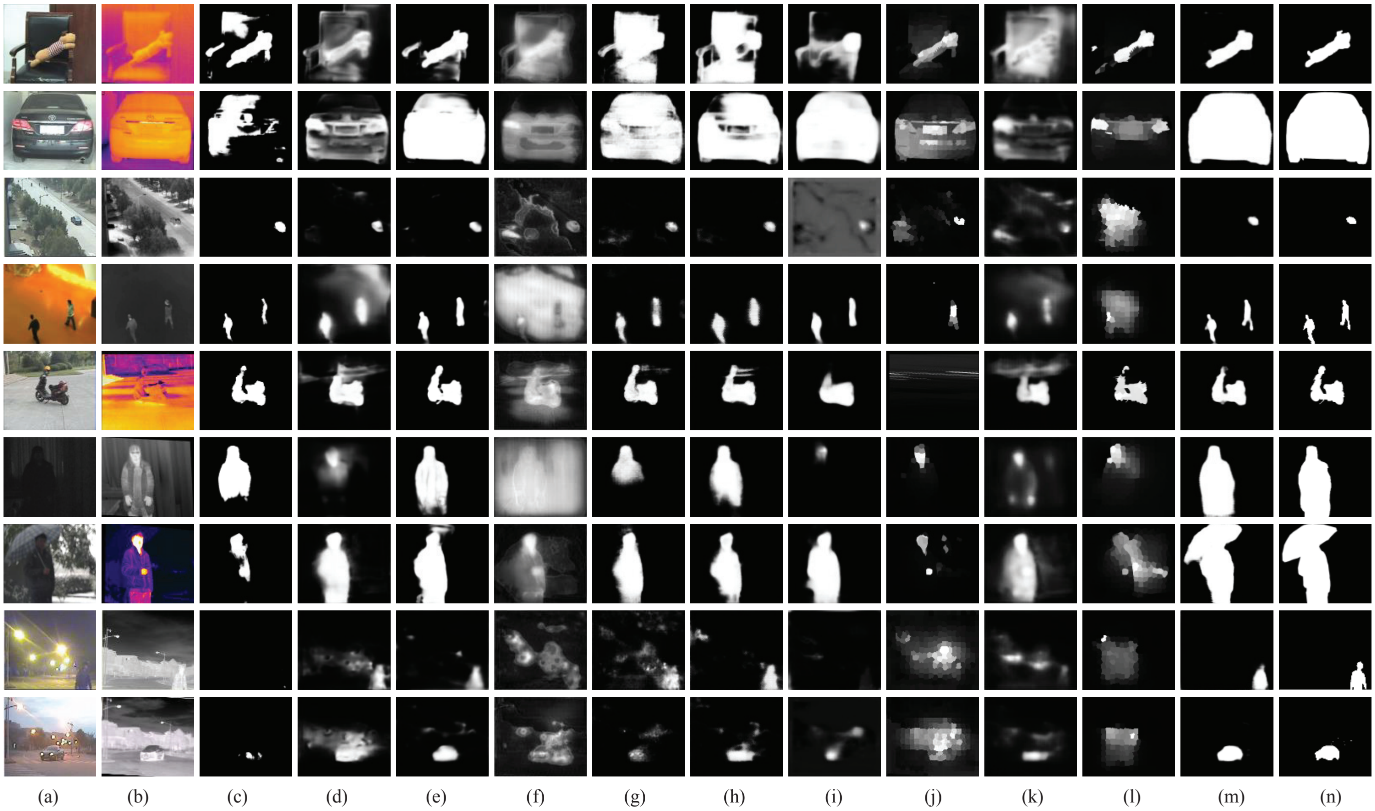
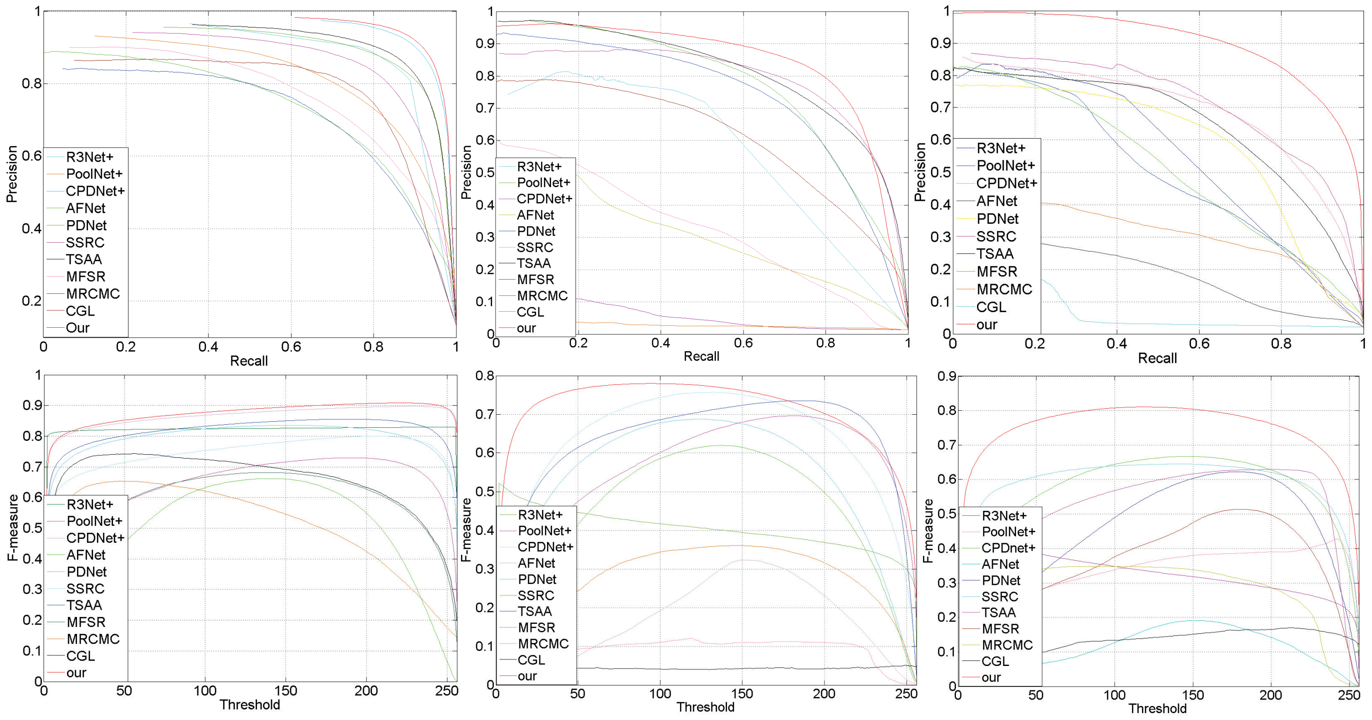


Fig. 8. Visual comparisons of different methods. From left to right: (a) RGB images; (b) Thermal infrared images; (c) R3Net+; (d) PoolNet+; (e) CPDNet+; (f) AFNet; (g) PDNet; (h) TSAA; (i) SSRC; (j) MRCMC; (k) MFSR; (l) CGL; (m) Ours; (n) Ground truth.



(a) RGB-thermal

(b) Car

(c) Pedestrian

Fig. 9. Quantitative comparisons between the proposed algorithm and the other state-of-the-art methods. The first and second rows are the PR curves and F-measure curves of different methods on different data sets, respectively.

TABLE I
PERFORMANCE OF DIFFERENT SALIENCY DETECTION METHODS ON THE THREE DATASETS. THE BEST RESULTS ARE SHOWN IN BOLD.

Methods	Type	RGB-thermal				Car				Pedestrian				Runtime(s)
		F_{ave}	F_β^ω	MAE	S_α	F_{ave}	F_β^ω	MAE	S_α	F_{ave}	F_β^ω	MAE	S_α	
R3Net	RGB	0.803	0.755	0.046	0.831	0.602	0.528	0.022	0.661	0.274	0.249	0.047	0.572	0.16
R3Net+	RGB→RGB-T	0.852	0.785	0.049	0.838	0.532	0.358	0.014	0.591	0.492	0.307	0.015	0.574	0.33
PoolNet	RGB	0.674	0.645	0.070	0.763	0.119	0.025	0.013	0.498	0.035	0.031	0.019	0.489	0.09
PoolNet+	RGB→RGB-T	0.716	0.654	0.051	0.863	0.183	0.307	0.029	0.699	0.209	0.217	0.099	0.601	0.19
CPDNet	RGB	0.788	0.768	0.041	0.861	0.199	0.301	0.055	0.668	0.165	0.154	0.079	0.578	0.08
CPDNet+	RGB→RGB-T	0.860	0.838	0.028	0.889	0.319	0.481	0.016	0.811	0.341	0.463	0.018	0.750	0.16
AFNet	RGB-D→RGB-T	0.700	0.682	0.062	0.841	0.143	0.044	0.162	0.495	0.070	0.044	0.338	0.412	0.18
PDNet	RGB-D→RGB-T	0.803	0.750	0.048	0.869	0.561	0.399	0.016	0.751	0.451	0.286	0.041	0.658	0.11
TSAA	RGB-D→RGB-T	0.817	0.778	0.040	0.882	0.642	0.532	0.013	0.793	0.529	0.453	0.021	0.730	0.13
SSRC	RGB-D→RGB-T	0.750	0.694	0.055	0.833	0.309	0.068	0.189	0.507	0.635	0.573	0.014	0.773	0.24
MRCMC	RGB-T	0.661	0.428	0.109	0.688	0.078	0.048	0.072	0.501	0.319	0.065	0.074	0.613	1.99
MFSR	RGB-T	0.701	0.673	0.073	0.823	0.177	0.155	0.046	0.613	0.201	0.138	0.106	0.569	0.18
CGL	RGB-T	0.771	0.585	0.086	0.765	0.103	0.088	0.033	0.455	0.109	0.062	0.117	0.495	2.33
Ours	RGB-T	**0.873**	**0.858**	**0.025**	**0.911**	**0.708**	**0.620**	**0.007**	**0.795**	**0.745**	**0.667**	**0.010**	**0.838**	0.12

ADFC modules in our proposed network. Seven schemes as well as their corresponding saliency detection results on the RGB-thermal dataset are listed in Table II. In the first three schemes, only a pair of ADFC modules are employed. One is used to extract a specific level of features from RGB images (e.g., $\mathbf{H}_1^1$), and the other is used to extract the corresponding level of features from thermal infrared images (e.g., $\mathbf{H}_1^2$). In the fourth, fifth and sixth schemes, two pairs of ADFC modules are employed to extract two specific levels of features from RGB images (e.g., $\mathbf{H}_1^1$ and $\mathbf{H}_2^1$) and thermal infrared images (e.g., $\mathbf{H}_1^2$ and $\mathbf{H}_2^2$), respectively. In the last scheme, i.e., the scheme employed in our proposed network, three pairs of ADFC modules are employed to extract the RGB image features (e.g., $\mathbf{H}_1^1$, $\mathbf{H}_2^1$, and $\mathbf{H}_3^1$) and thermal infrared image features (e.g., $\mathbf{H}_1^2$,$\mathbf{H}_2^2$, and $\mathbf{H}_3^2$), respectively.

From the experimental results in Table II, it can be concluded that higher saliency detection performance can be obtained by those schemes extracting multiple levels of features than those just extracting one specific level of features. Especially, the proposed scheme that extracts three levels of features achieves the best performance among the schemes mentioned here.

TABLE II
PERFORMANCE OF DIFFERENT SCHEMES IN THE PROPOSED FEATURE EXTRACTION MODULE FOR SINGLE-MODAL INPUT IMAGE ON RGB-THERMAL DATASET.

Schemes	Extracted features	F_{ave}	MAE	S_α
1	$\{\mathbf{H}_1^1, \mathbf{H}_1^2\}$	0.810	0.041	0.877
2	$\{\mathbf{H}_2^1, \mathbf{H}_2^2\}$	0.823	0.038	0.883
3	$\{\mathbf{H}_3^1, \mathbf{H}_3^2\}$	0.820	0.039	0.887
4	$\{\mathbf{H}_1^1, \mathbf{H}_1^2\}$,$\{\mathbf{H}_2^1, \mathbf{H}_2^2\}$	0.851	0.029	0.891
5	$\{\mathbf{H}_2^1, \mathbf{H}_2^2\}$,$\{\mathbf{H}_3^1, \mathbf{H}_3^2\}$	0.858	0.027	0.906
6	$\{\mathbf{H}_1^1, \mathbf{H}_1^2\}$,$\{\mathbf{H}_3^1, \mathbf{H}_3^2\}$	0.856	0.030	0.896
7	$\{\mathbf{H}_1^1, \mathbf{H}_1^2\}$,$\{\mathbf{H}_2^1, \mathbf{H}_2^2\}$,$\{\mathbf{H}_3^1, \mathbf{H}_3^2\}$	**0.873**	**0.025**	**0.911**

TABLE III
PERFORMANCE OF USING DIFFERENT FEATURE EXTRACTION MODULES ON RGB-THERMAL DATASET.

Schemes	Module	F_{ave}	MAE	S_α	Runtime (s)
1	baseline	0.739	0.052	0.863	0.19
2	+FPN	0.752	0.037	0.872	0.14
3	+HIFI	0.747	0.071	0.821	0.23
4	+RFC	0.817	0.038	0.873	0.52
5	+ADFC	0.873	0.025	0.911	0.12

Furthermore, we compare the proposed ADFC module with several SOTA feature extraction modules for RGB images, including resolution based feature combination (RFC) structure in Amulet [25], Hierarchical feature integration(HIFI) module in [56] and FPN in [55]. For that, we first design a baseline method, where the ADFC modules are removed from our proposed method and the features from the backbone networks are directly fed into the subsequent feature fusion and saliency inference modules. Then several versions of our proposed methods are compared by replacing ADFC with the other feature extraction modules mentioned above, while the rest modules in our proposed method are kept unchanged. The experimental results in Table III demonstrate that the saliency detection performance can be improved via performing multi-level feature extraction modules on the backbone networks. In addition, the proposed ADFC module achieves higher saliency detection performance than the other feature extraction modules. It also has the highest computational efficiency among the four feature extraction modules.

2) Validity of multi-modal image feature fusion using MGF : In order to verify the validity of the proposed MGF module, we also provide six fusion schemes for multi-modal image features. Specifically, in the first two schemes (ADD and MGF1, for short, respectively), only the single-group fusion branch

TABLE IV
PERFORMANCE OF DIFFERENT SCHEMES IN THE PROPOSED FEATURE FUSION MODULE FOR MULTI-MODAL IMAGES ON RGB-THERMAL DATASET.

Metrics	ADD	MGF1	MGF2 (*M*=8)	MGF_4 (*M*=4)	MGF_8 (*M*=8)	MGF_16 (*M*=16)	MGF_32 (*M*=32)
F_{ave}	0.749	0.843	0.849	0.853	**0.873**	0.854	0.860
MAE	0.039	0.033	0.030	0.030	**0.025**	0.031	0.028
S_α	0.867	0.898	0.899	0.899	**0.911**	0.900	0.907

TABLE V
PERFORMANCE OF DIFFERENT VERSIONS OF THE PROPOSED BI-DIRECTIONAL MESSAGE PASSING MODULE ON RGB-THERMAL DATASET.

Metrics	No_Bi	Bi_NA	Bi_CA	Bi_SA	Bi_JCSA
F_{ave}	0.818	0.845	0.855	0.850	**0.873**
MAE	0.037	0.032	0.030	0.032	**0.025**
S_α	0.868	0.895	0.899	0.897	**0.911**

is employed. Simple element-wise addition and concatenation are first performed on the features from the input images, respectively, in the two schemes. Then the regular convolution is applied to obtain the fused features. In the third scheme (MGF2, for short), only the multi-group fusion branch with *M*=8 is employed. In the rest of schemes, the two branches are jointly employed but with different numbers of groups (MGF_4 with *M=4*, MGF_8 with *M=8*, MGF_16 with *M=16* and MGF_32 with *M=32*, for short, respectively). The output channels of these models are all set to 128 for fair comparison.

The performance of different schemes is given in Table IV. By comparing ADD and MGF1, it can be easily found that the simple concatenation fusion strategy significantly outperforms the element-wise addition fusion strategy for multi-modal RGB-T images salient object detection. This may be attributed to the fact that simple element-wise addition may easily weaken the discriminability of the fused features because of the polarity inverse between the RGB and thermal image intensities. The experiment also demonstrates that multi-group fusion schemes can obtain better results than the traditional single-group fusion schemes. The performance can be further improved by combining the two schemes. Especially, when the number of groups M is set to 8 in the proposed MGF module, the best saliency detection performance can be obtained. As discussed in [57], some salient features from individual single-modal image may be easily drowned in the concatenated features because of the large number of features. It is hard for single-group fusion scheme (i.e., the regular convolution) to boost those salient features from all of the input features. When the group numbers are too large (e.g., $M > 8$), the correlations among the feature maps will be weaken, since fewer channels of input features will be fed into each convolutional filter. This will also diminish the discriminability of the subsequent fused features [71]. Therefore, in our proposed MGF module, M is set 8.

3) Validity of the proposed JCSA block for saliency prediction: In this part, five versions (No_Bi, Bi_NA, Bi_CA, Bi_SA, Bi_JCSA, for short, respectively) of the proposed bi-directional message passing module are compared to test the validity of the JCSA block. In No_Bi, the proposed JABMP is removed from our proposed network. Instead, as in [55], simple up-sampling and concatenation strategies are employed to integrate the multi-level fused features for the final saliency prediction in a coarser-to-finer way. In Bi_NA, no attention guidance is employed in the bi-directional message passing module. In Bi_CA, only the channel-wise attention mechanism (i.e., SE block in [65]) is employed to control message passing in the bi-directional message passing model. Similarly, only the spatial attention mechanism (i.e., CA block in [66]) is employed in Bi_SA, and the joint channel-spatial attention (JCSA) mechanism is employed in Bi_JCSA (i.e., the proposed attention guided bi-directional message passing module, shown in Fig. 6). The experimental results in Table V indicate that the attention mechanism, especially the joint channel-spatial attention mechanism, can greatly improve the saliency detection results.

V. CONCLUSION

In this paper, we have presented a novel end-to-end deep neural network model for RGB-T salient object detection, where the multi-modal saliency detection is formulated as a CNN feature fusion problem. The proposed model consists of three components, i.e., multi-level feature extraction of single-modal images using multiple adjacent-depth feature combination (ADFC) modules, cross-modal feature fusion of RGB-T image pairs using multi-branch group fusion (MGF) modules, and saliency prediction using a joint attention guided bi-directional message passing (JABMP) module. A joint channel-spatial attention (JCSA) mechanism is further employed in the proposed message passing module to focus on those important features with high channel attention as well as spatial attention but suppress those superfluous features. Experimental results demonstrate that the proposed RGB-T salient object detection method performs better than the state-of-the-art methods, especially for those challenging scenes with poor illumination, complex background or low contrast. One possible future work is to ap- ply our saliency detector to industrial applications, such as image classification [72], object tracking [73], [74] and instance-level object retrieval [75], [76].

REFERENCES

[1] H. Wu, G. Li, and X. Luo, “Weighted attentional blocks for probabilistic object tracking,” *The Visual Computer*, vol. 30, no. 2, pp. 229–243, 2014.

[2] F. Wang, Y. Zhen, B. Zhong, and R. Ji, "Robust infrared target tracking based on particle filter with embedded saliency detection," *Information Sciences*, vol. 301, pp. 215–226, 2015.

[3] D. Gao, S. Han, and N. Vasconcelos, "Discriminant saliency, the detection of suspicious coincidences, and applications to visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 6, pp. 989–1005, 2009.

[4] A. Abdulmunem, Y. Lai, and X. Sun, "Saliency guided local and global descriptors for effective action recognition," *Computational Visual Media*, vol. 2, no. 1, pp. 97–106, 2016.

[5] C. Guo and L. Zhang, "A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression," *IEEE Transactions on Image Processing*, vol. 19, no. 1, pp. 185–198, 2010.

[6] L. Shen, Z. Liu, and Z. Zhang, "A novel H. 264 rate control algorithm with consideration of visual attention," *Multimedia Tools and Applications*, vol. 63, no. 3, pp. 709–727, 2013.

[7] B. Cheng, L. Jin, and G. Li, "General fusion method for infrared and visual images via latent low-rank representation and local non-subsampled shearlet transform," *Infrared Physics & Technology*, vol. 92, pp. 68–77, 2018.

[8] D. P. Bavirisetti and R. Dhuli, "Two-scale image fusion of visible and infrared images using saliency detection," *Infrared Physics & Technology*, vol. 76, pp. 52–64, 2016.

[9] J. Han, D. Zhang, G. Cheng, N. Liu, and D. Xu, "Advanced deep learning techniques for salient and category-specific object detection: A survey," *IEEE Signal Processing Magazine*, vol. 35, no. 1, pp. 84–100, 2018.

[10] R. Quan, J. Han, D. Zhang, F. Nie, X. Qian, and X. Li, "Unsupervised salient object detection via inferring from imperfect saliency models," *IEEE Transactions on Multimedia*, vol. 20, no. 5, pp. 1101–1112, 2018.

[11] M. Cheng, N. J. Mitra, X. Huang, P. H. Torr, and S.-M. Hu, "Global contrast based salient region detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 3, pp. 569–582, 2015.

[12] J. Kim, D. Han, Y. Tai, and J. Kim, "Salient region detection via high-dimensional color transform and local spatial support," *IEEE Transactions on Image Processing*, vol. 25, no. 1, pp. 9–23, 2016.

[13] F. Huang, J. Qi, H. Lu, L. Zhang, and X. Ruan, "Salient object detection via multiple instance learning," *IEEE Transactions on Image Processing*, vol. 26, no. 4, pp. 1911–1922, 2017.

[14] H. Lu, X. Li, L. Zhang, X. Ruan, and M. H. Yang, "Dense and sparse reconstruction error based saliency descriptor," *IEEE Transactions on Image Processing*, vol. 25, no. 4, pp. 1592–1603, 2016.

[15] C. Yang, L. Zhang, H. Lu, X. Ruan, and M. H. Yang, "Saliency detection via graph-based manifold ranking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3166–3173.

[16] H. Peng, B. Li, H. Ling, W. Hu, W. Xiong, and S. J. Maybank, "Salient object detection via structured matrix decomposition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 818–832, 2017.

[17] T. Zhao, L. Li, X. Ding, Y. Huang, and D. Zeng, "Saliency detection with spaces of background-based distribution," *IEEE Signal Processing Letters*, vol. 23, no. 5, pp. 683–687, 2016.

[18] A. Borji, M. M. Cheng, H. Jiang, and J. Li, "Salient object detection: A benchmark," *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 5706–5722, 2015.

[19] D. Zhu, Y. Luo, L. Dai, X. Shao, Q. Zhou, L. Itti, and J. Lu, "Salient object detection via a local and global method based on deep residual network," *Journal of Visual Communication and Image Representation*, vol. 54, pp. 1–9, 2018.

[20] X. Li, L. Zhao, L. Wei, M. H. Yang, F. Wu, Y. Zhuang, H. Ling, and J. Wang, "Deepsaliency: Multi-task deep neural network model for salient object detection," *IEEE Transactions on Image Processing*, vol. 25, no. 8, pp. 3919–3930, 2016.

[21] P. Zhang, D. Wang, H. Lu, H. Wang, and B. Yin, "Learning uncertain convolutional features for accurate saliency detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 212–221.

[22] G. Li and Y. Yu, "Visual saliency detection based on multiscale deep CNN features," *IEEE Transactions on Image Processing*, vol. 25, no. 11, pp. 5012–5024, 2016.

[23] G. Lee, Y. W. Tai, and J. Kim, "Eld-net: An efficient deep learning architecture for accurate saliency detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 7, pp. 1599–1610, 2018.

[24] Q. Hou, M. Cheng, X. Hu, A. Borji, Z. Tu, and P. Torr, "Deeply supervised salient object detection with short connections." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 4, p. 815, 2019.

[25] P. Zhang, D. Wang, H. Lu, H. Wang, and X. Ruan, "Amulet: Aggregating multi-level convolutional features for salient object detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 202–211.

[26] C. Li, X. Wang, L. Zhang, J. Tang, H. Wu, and L. Lin, "Weighted low-rank decomposition for robust grayscale-thermal foreground detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 4, pp. 725–738, 2017.

[27] C. Li, G. Wang, Y. Ma, A. Zheng, B. Luo, and J. Tang, "A unified RGB-T saliency detection benchmark: dataset, baselines, analysis and a novel approach," *arXiv preprint arXiv:1701.02829*, 2017.

[28] Y. Ma, D. Sun, Q. Meng, Z. Ding, and C. Li, "Learning multiscale deep features and SVM regressors for adaptive RGB-T saliency detection," in *Proceedings of the International Symposium on Computational Intelligence and Design*, 2017, pp. 389–392.

[29] S. Gundimada, V. K. Asari, and N. Gudur, "Face recognition in multi-sensor images based on a novel modular feature selection technique," *Information Fusion*, vol. 11, no. 2, pp. 124–132, 2010.

[30] V. N. Gangapure, S. Nanda, and A. S. Chowdhury, "Superpixel-based causal multisensor video fusion," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 6, pp. 1263–1272, 2018.

[31] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[32] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[34] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1492–1500.

[35] L. Zhang, J. Dai, H. Lu, Y. He, and G. Wang, "A bi-directional message passing model for salient object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1741–1750.

[36] D. Konig, M. Adam, C. Jarvers, G. Layher, H. Neumann, and M. Teutsch, "Fully convolutional region proposal networks for multi-spectral person detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 49–56.

[37] J. Wagner, V. Fischer, M. Herman, and S. Behnke, "Multispectral pedestrian detection using deep fusion convolutional neural networks," in *Proceedings of 24th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, 2016, pp. 509–514.

[38] J. Liu, S. Zhang, S. Wang, and D. N. Metaxas, "Multispectral deep neural networks for pedestrian detection," in *Proceedings of the 27th British Machine Vision Conference*, 2016, pp. 1–13.

[39] J.-J. Liu, Q. Hou, M.-M. Cheng, J. Feng, and J. Jiang, "A simple pooling-based design for real-time salient object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3917–3926.

[40] Z. Deng, X. Hu, L. Zhu, X. Xu, J. Qin, G. Han, and P.-A. Heng, "R3net: Recurrent residual refinement network for saliency detection," in *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, 2018, pp. 684–690.

[41] Z. Wu, L. Su, and Q. Huang, "Cascaded partial decoder for fast and accurate salient object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3907–3916.

[42] H. Chen and Y. Li, "Progressively complementarity-aware fusion network for RGB-D salient object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3051–3060.

[43] D. P. Fan, Z. Lin, J.-X. Zhao, Y. Liu, Z. Zhang, Q. Hou, M. Zhu, and M.-M. Cheng, "Rethinking RGB-D salient object detection: Models, datasets, and large-scale benchmarks," *arXiv preprint arXiv:1907.06781*, 2019.

[44] J. X. Zhao, Y. Cao, D. P. Fan, M. M. Cheng, X. Y. Li, and L. Zhang, "Contrast prior and fluid pyramid integration for RGB-D salient object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3927–3936.

[45] N. Wang and X. Gong, "Adaptive fusion for RGB-D salient object detection," *IEEE Access*, vol. 7, pp. 55 277–55 284, 2019.

[46] H. Chen and Y. Li, "Three-stream attention-aware network for RGB-D salient object detection," *IEEE Transactions on Image Processing*, vol. 28, no. 6, pp. 2825–2835, 2019.

[47] C. Zhu, X. Cai, K. Huang, T. H. Li, and G. Li, "PDNet: Prior-model guided depth-enhanced network for salient object detection," in *Proceedings of the IEEE International Conference on Multimedia and Expo*, 2019, pp. 199–204.

[48] Z. Liu, S. Shi, Q. Duan, W. Zhang, and P. Zhao, "Salient object detection for RGB-D image by single stream recurrent convolution neural network," *Neurocomputing*, vol. 363, pp. 46–57, 2019.

[49] Z. Tu, T. Xia, C. Li, X. Wang, Y. Ma, and J. Tang, "RGB-T image saliency detection via collaborative graph learning," *arXiv preprint arXiv:1905.06741*, 2019.

[50] Y. Liu, X. Chen, J. Cheng, H. Peng, and Z. Wang, "Infrared and visible image fusion with convolutional neural networks," *International Journal of Wavelets, Multiresolution and Information Processing*, vol. 16, no. 3, pp. 1–20, 2018.

[51] C. Li, X. Wu, N. Zhao, X. Cao, and J. Tang, "Fusing two-stream convolutional neural networks for RGB-T object tracking," *Neurocomputing*, vol. 281, pp. 78–85, 2018.

[52] C. Li, H. Cheng, S. Hu, X. Liu, J. Tang, and L. Lin, "Learning collaborative sparse representation for grayscale-thermal tracking," *IEEE Transactions on Image Processing*, vol. 25, no. 12, pp. 5743–5756, 2016.

[53] C. Li, X. Liang, Y. Lu, N. Zhao, and J. Tang, "RGB-T object tracking: Benchmark and baseline," *arXiv preprint arXiv:1805.08982*, 2018.

[54] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th International Conference on Machine Learning*, 2010, pp. 807–814.

[55] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2117–2125.

[56] K. Zhao, W. Shen, S. Gao, D. Li, and M. M. Cheng, "Hi-fi: Hierarchical feature integration for skeleton detection," in *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, 2018, pp. 1191–1197.

[57] Y. Liu, J. Han, Q. Zhang, and C. Shan, "Deep salient object detection with contextual information guidance," *IEEE Transactions on Image Processing*, vol. 29, pp. 360–374, 2019.

[58] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2012, pp. 1097–1105.

[59] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proceedings of the 22nd ACM International Conference on Multimedia*, 2014, pp. 675–678.

[60] PyTorch, "Pytorch: Tensors and dynamic neural networks in python with strong gpu acceleration," Website, http://pytorch.org/ .

[61] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1–9.

[62] S. Ioffe and C. Szegedy, "Batch normalization: accelerating deep network training by reducing internal covariate shift," in *Proceedings of the 32nd International Conference on International Conference on Machine Learning*, 2015, pp. 448–456.

[63] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2818–2826.

[64] G. Li and Y. Yu, "Deep contrast learning for salient object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 478–487.

[65] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7132–7141.

[66] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, and T.-S. Chua, "Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5659–5667.

[67] S. Xie and Z. Tu, "Holistically-nested edge detection," *International Journal of Computer Vision*, vol. 125, no. 1, pp. 3–18, 2017.

[68] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H. Y. Shum, "Learning to detect a salient object," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 2, pp. 353–367, 2011.

[69] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, and A. Borji, "Structure-measure: A new way to evaluate foreground maps," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 4548–4557.

[70] R. Margolin, L. Zelnik-Manor, and A. Tal, "How to evaluate foreground maps?" in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 248–255.

[71] X. Zhang, X. Zhou, M. Lin, and J. Sun, "Shufflenet: An extremely efficient convolutional neural network for mobile devices," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6848–6856.

[72] G. Ding, Y. Guo, K. Chen, C. Chu, J. Han, and Q. Dai, "Decode: Deep confidence network for robust image classification," *IEEE Transactions on Image Processing*, vol. 28, no. 8, pp. 3752–3765, 2019.

[73] B. Zhang, W. Yang, Z. Wang, L. Zhuo, J. Han, and X. Zhen, "The structure transfer machine theory and applications," *IEEE Transactions on Image Processing, 10.1109/TIP.2019.2954178*.

[74] J. H. Han, E. Pauwels, P. de Zeeuw, and P. de With, "Employing a rgb-d sensor for real-time tracking of humans across multiple re-entries in a smart environment," *IEEE Transactions on Consumer Electronics*, vol. 58, no. 2, pp. 255–263, 2012.

[75] G. Wu, J. Han, Y. Guo, L. Liu, G. Ding, Q. Ni, and L. Shao, "Unsupervised deep video hashing via balanced code for large-scale video retrieval," *IEEE Transactions on Image Processing*, vol. 28, no. 4, pp. 1993 – 2007, 2019.

[76] G. Wu, J. Han, Z. Lin, G. Ding, B. Zhang, and Q. Ni, "Joint image-text hashing for fast large-scale cross-media retrieval using self-supervised deep learning," *IEEE Transactions on Industrial Electronics*, vol. 66, no. 12, pp. 9868 – 9877, 2019.

Qiang Zhang is a professor with the Automatic Control Department at Xidian University, China. His research interests include image processing and pattern recognition.

Nianchang Huang is currently pursuing the Ph.D. degree in School of Mechano-Electronic Engineering at Xidian University, China. His research interests include deep learning and multimodal image processing in computer vision.

Lin Yao is currently working towards the M. S. degree in Control Theory and Control Engineering at Xidian University, China. Her current research interests include deep learning and multimodal image based salient object detection.

Dingwen Zhang is currently an associate professor at Xidian University. His research interests include computer vision and multimedia processing, especially on saliency detection, co-saliency detection, and weakly supervised learning.

Caifeng Shan is currently a Senior Scientist and Project Leader with Philips Research, Eindhoven, The Netherlands. His research interests include medical image processing, and computer vision.

Jungong Han is currently an associate professor with WMG Data Science at University of Warwick, UK. His research interests span the fields of video analysis, computer vision and applied machine learning.