Multi-Interactive Dual-Decoder for RGB-Thermal Salient Object Detection

Zhengzheng Tu, Zhun Li, Chenglong Li, Yang Lang, and Jin Tang

Abstract—RGB-thermal salient object detection (SOD) aims to segment the common prominent regions of visible image and corresponding thermal infrared image that we call it RGBT SOD. Existing methods don't fully explore and exploit the potentials of complementarity of different modalities and multi-type cues of image contents, which play a vital role in achieving accurate results. In this paper, we propose a multi-interactive dual-decoder to mine and model the multi-type interactions for accurate RGBT SOD. In specific, we first encode two modalities into multi-level multi-modal feature representations. Then, we design a novel dual-decoder to conduct the interactions of multi-level features, two modalities and global contexts. With these interactions, our method works well in diversely challenging scenarios even in the presence of invalid modality. Finally, we carry out extensive experiments on public RGBT and RGBD SOD datasets, and the results show that the proposed method achieves the outstanding performance against state-of-the-art algorithms. The source code has been released at: https://github.com/lz118/Multi-interactive-Dual-decoder.

Index Terms—Salient object detection, information fusion, dual-decoder, multiple interactions.

I. INTRODUCTION

RGB-THERMAL salient object detection aims at estimating the common conspicuous objects or regions in an aligned visible and thermal infrared image pair. Thermal infrared sensor captures the infrared ray and generates an image that can present the temperature of objects. Thermal infrared images can provide many informative cues, making many hard tasks be solved well, which has been proven in lots of computer vision tasks, such as RGBT tracking [1]–[3] and multi-spectral person re-identification [4], [5]. In this paper,

we focus on exploring the saliency of visual perception and temperature field, which is named RGBT SOD for short. It is meaningful in some applications such as automatic driving in foggy weather and nighttime, and abnormal object detection in some important places such as power stations.

There is a similar task called RGBD SOD that introduces depth images to provide complementary information for coping with the challenges in SOD in visible images. RGBT SOD differs from it as follows. RGBD SOD aims to introduce depth information as an auxiliary modality to handle some difficulties such as similar foreground and cluttered background. As the depth information can catch the distance between objects and camera, it can easily distinguish salient object from similar objects and cluttered background. In most cases, the saliency in the depth image is also in accord with visual saliency. Therefore, existing RGBD SOD methods mainly focus on exploring depth information as supplementary. While RGBT SOD treats RGB and thermal modalities equally, and takes advantages of different modalities to discover common conspicuous objects in both modalities.

Existing works have explored the task of RGBT SOD to some extent. Wang *et al.* [6] propose a multi-task manifold ranking algorithm, which introduces a weight for each modality to describe the reliability and then adaptively fuses two modalities with these weights. Based on [6], Tang *et al.* [7] further take both of collaboration and heterogeneity into account to fuse different modalities more effectively. Tu *et al.* [8] propose a collaborative graph learning algorithm to integrate multi-level deep features. However, handcraft features cannot represent the semantic relevance among pixels well, and the superpixel-based methods need to segment superpixels accurately. Thus the above-mentioned graph-based RGBT SOD methods cann't have robust performances on various challenges. In recent years, deep learning has shown its superiority for feature representation. There are also some deep learning methods for RGBT SOD. For example, Tu *et al.* [9] use the attention mechanism to refine multi-level features of two modalities and adopt another encoding branch to fuse these features. More recently, Zhang *et al.* [10] take advantages of feature fusion in CNN to handle the challenge of RGBT SOD. They design three modules respectively for combining adjacent-depth features, capturing the cross-modal features and predicting saliency maps by integrating the multi-level fused features.

Although above-mentioned methods have achieved remarkable performance, there are two key problems not addressed

well. The first one is to achieve effective complementation between modalities and prevent the interference from noises. The second one is to suppress saliency bias, which means we should focus on common conspicuous regions and avoid obtaining the results dominated by one modality.

So, with overall consideration of above problems, we design a dual-decoder to decode two modalities severally and enable decoded features interaction to adjust the saliency, thus achieving common saliency of two modalities. Specifically, we design a multi-interaction block (MIB) to model the interactions among dual modalities, multi-level features and global contexts. On the one hand, we take modality interaction between two branches to learn from each other for mutual complementation. On the other hand, there are also interactions between global contexts and multi-level features in dual-decoder. We embed three MIBs into each stream of dual-decoder to decode features in a progressive way. The above design has the following major properties. 1) It propagates the useful complementary features of dual modalities to each other, and the good performance can be achieved even though one modality is invalid or with many noises. 2) It captures the hierarchical encoded features to restore more spatial details. The up-sampled decoded features can be complemented by these spatial details, thus we can achieve clearer structure of salient objects. 3) It receives global context, which can help highlighting complete salient regions. The global context contains strong semantic relevance of pixels within and between modalities, and is thus helpful for locating salient objects and hold back background noises.

For suppressing the modality bias, we force the two decoder branches of dual-decoder to capture more consistent features from different encoded features. We set two prediction heads on top of dual-decoder to compute saliency maps of two modalities, and employ label supervision to drive feature learning in each modality. With the supervision driven, two decoder streams trend to discover common salient regions from two modalities. It should be noticed that the two decoder streams have their independent parameters since they receive the encoded features from different modalities. The output features from dual-decoder are fused to predict the final saliency map and the common conspicuous regions can thus be highlighted.

To make the dual-decoder more robust to the problem that one modality is invalid or with many noises, we develop a data augmentation strategy. In specific, we randomly replace the input of a modality with zero or a noise map sampled in standard normal distribution. By this way, the proposed network can be trained to adapt these situations. Some challenging cases are shown in Fig. 1, which demonstrates our multi-interactive dual-decoder network has the ability to capture the complementary information and suppress the noises of defective inputs.

Note that RGBD SOD usually treats depth data as the supplement of RGB data and thus employs depth information to assist RGB SOD. While RGBT SOD treats RGB and thermal infrared data equally and thus aims at estimating the common conspicuous objects or regions in both modalities. To this end, we use the same network structure for each
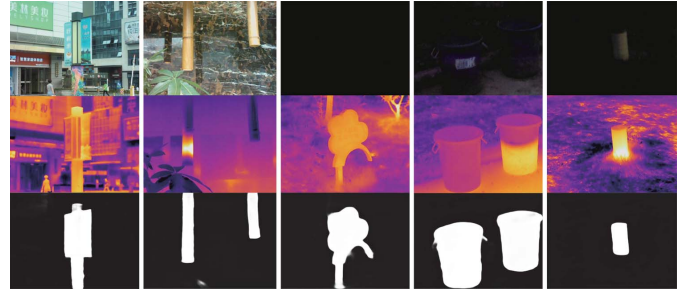

Fig. 1. Our results on some challenges. The first and the second rows are visible images and thermal images respectively. The third row shows our results, demonstrating our network performs well with defective inputs.

modality to model all modalities equally. Such design is specific for RGBT SOD, but it can also be applied to RGBD SOD if we treat RGB and depth information equally, which is not mainstream in the community of RGBD SOD.

Extensive experiments on several public RGBT and RGBD SOD datasets show that the proposed method achieves the state-of-the-art performance against the existing methods. The contributions of this paper are summarized as follows:

- We propose a novel dual-decoder architecture for effective RGBT SOD. Existing methods usually perform multimodal fusion in the encoding phase, but such a fusion strategy is more difficult to be optimized than fusion in the decoding phase. In back propagation, the path of gradient computation in the decoder is shorter than in the encoder, and the optimization procedure of decoder is thus less to be influenced by vanishing gradient or exploding gradient. Therefore, the decoder is easier to be optimized than the encoder.

- We propose a unified model to seamlessly integrate the multi-type interactions for robust RGBT SOD. In particular, the multi-interactive model can extract multi-type cues including correlation between modalities, spatial details and global context. Different from using different cues separately to refine decoded features in existing works, we seamlessly integrate the multi-interactive block in the dual-decoder as the basic module of the decoder. In addition, our network also has the interaction between two decoders, which is added into every decoding step to gradually improve the decoded features.

- A simple yet effective data augmentation strategy is designed to train the proposed network, and the performance is thus further boosted in the presence of invalid or unreliable modalities. We use random zero or noise map to simulate the challenging cases of RGBT image pairs that significantly enhance the diversity of training samples. From experimental results we can see that although this operation is simple, the effect is prominent.

II. RELATED WORK

A. Salient Object Detection

In recent years, deep learning-based methods have achieved great progress in salient object detection. Wang *et al.* [11] utilize the neural network to perform the fusion of local spatial features and global semantic features, then combine

local estimation with global search to predict saliency map. Liu *et al.* [12] design a deep hierarchical network to predict a coarse saliency map and then refine it hierarchically and progressively. Next, many researches come out based on fully convolutional network with its successful applications in semantic segmentation. Wang *et al.* [13] propose a recurrent FCN to constantly refine the saliency. Hou *et al.* [14] add several short connections from the deeper side output to the shallower side output for enhancing more accuracy. Luo *et al.* [15] aggregate hierarchical features and contrast features to generate local saliency map, and then use the global semantic features to polish it. Liu *et al.* [16] propose a novel pixel-level context-aware network that learns to pay selective attention to the context information of each pixel. Deng *et al.* [17] propose a module to learn the residual between the intermediate saliency prediction and the ground truth by alternatively using the low-level features and the high-level features. Wu *et al.* [18] design a cascaded partial decoder for fast and accurate salient object detection. Zhao *et al.* [19] introduce edge detection to SOD for more accurate boundaries. Wei *et al.* [20] design a reasonable module to aggregate hierarchical features effectively and propose a weighted binary cross entropy loss function to emphasize the region that is difficult to predict. Numerous studies have made SOD algorithms more robust. However, with existing algorithms, it is difficult to deal with some challenges such as bad imaging conditions, which might cause defects or semantic ambiguity in visible images.

B. RGBT Salient Object Detection

With the availability of thermal sensors, many works [21], [22] introduce the thermal infrared images in SOD by utilizing their complementary information. As the original work, Wang *et al.* [6] construct the first RGBT SOD dataset and propose a multi-task manifold ranking algorithm. Tu *et al.* [23] use a multi-modal multi-scale manifold ranking to achieve the fusion of different features and introduce an intermediate variable to infer the optimal ranking seeds. Furthermore, Tu *et al.* [8] propose a collaborative graph learning method for RGBT SOD, which takes the superpixel as the graph node and uses hierarchical deep features to learn the graph affinity and node saliency. These RGBT SOD methods use traditional graph based technologies, which have the limited capability for feature representation. As the deep learning based methods show excellent performance for SOD, Tu *et al.* [9] build a large scale dataset containing 5000 image pairs. They also propose an effective baseline, which uses the attention mechanism to refine multi-level features of two modalities and adopts another encoding branch to fuse these features. Then the multi-level fused features are inputted into the decoder in a progressive way. Zhang *et al.* [10] also handle the challenge of RGBT SOD with feature fusion. Different from [9], they divide the fusion process into three parts and elaborately design three modules respectively for combining adjacent-depth features, capturing the cross-modal features and predicting saliency maps by integrating the multi-level fused features. Both of the methods focus on fusing encoded features of two modalities, and then use fused features to decode. However, without appropriate constraint, this kind of method

cannot handle the modality bias well. In this paper, we propose a more suitable network with multi-interaction dual-decoder to utilize the multi-type cues in a reasonable way and take the modalities bias into account simultaneously.

C. RGBD Salient Object Detection

RGBD SOD has been studied a lot in recent years. Qu *et al.* [24] design the handcraft features of multi-modality inputs and then use them as the input of network to predict saliency maps. Song *et al.* [25] also use the multi-modality inputs to design kinds of features and exploit low-level feature contrasts, mid-level feature weighted factors and high-level location priors to calculate saliency measures. Liu *et al.* [26] directly concatenate the modalities as a four-channel input, and then feed it to an encoder to obtain hierarchical feature representations. Finally, they use a depth recurrent network to render salient object outline from deep to shallow hierarchically and progressively. The multimodal fusion strategy of these methods is called early fusion, which simply combines the modalities as single input of network. In contrast to early fusion, the middle fusion strategy has more extensive researches. Many existing RGBD SOD methods adopt this kind of strategy to make full fusion of modalities. Liu *et al.* [27] perform the fusion during the decoding phase by directly adding the features of two modalities with the feature from the previous decoding step. Instead of directly adding together, Chen *et al.* [28] design a complementarity-aware fusion module to fuse the features of two modalities. Chen *et al.* [29] further propose a three-stream attention-aware network, in which a fusion stream is introduced that accompanies RGB-specific stream and depth-specific stream to obtain the new fused features for each layer, then the fused features are used for decoding. Piao *et al.* [30] extract multi-level complementary RGB and depth features and then fuse them by using residual connections. Then they model the relationship between depth information and object scales, and use a new recursive attention module to generate more accurate saliency maps. More recently, Fu *et al.* [31] use a siamese network to extract features of two modalities and propose a densely-cooperative fusion strategy. Zhang *et al.* [32] take background information into account. They introduce a Bilateral Attention Module that contains foreground-first attention and background-first attention, capturing more meaningful foreground and background cues. The middle fusion based methods can effectively fuse encoded features.

Different from these RGBD SOD methods which focus on features fusion [29]–[31] or using background information [32], we take the fusion during decoding instead of encoding phase, and we design a dual-decoder to decode two modalities severally and enable decoded features interaction to adjust the saliency, thus obtaining common saliency of two modalities.

III. Multi-Interactive Dual-Decoder Network

A. Overall Architecture

As shown in Fig. 2, we use two independent backbones to extract features from RGB images and thermal infrared
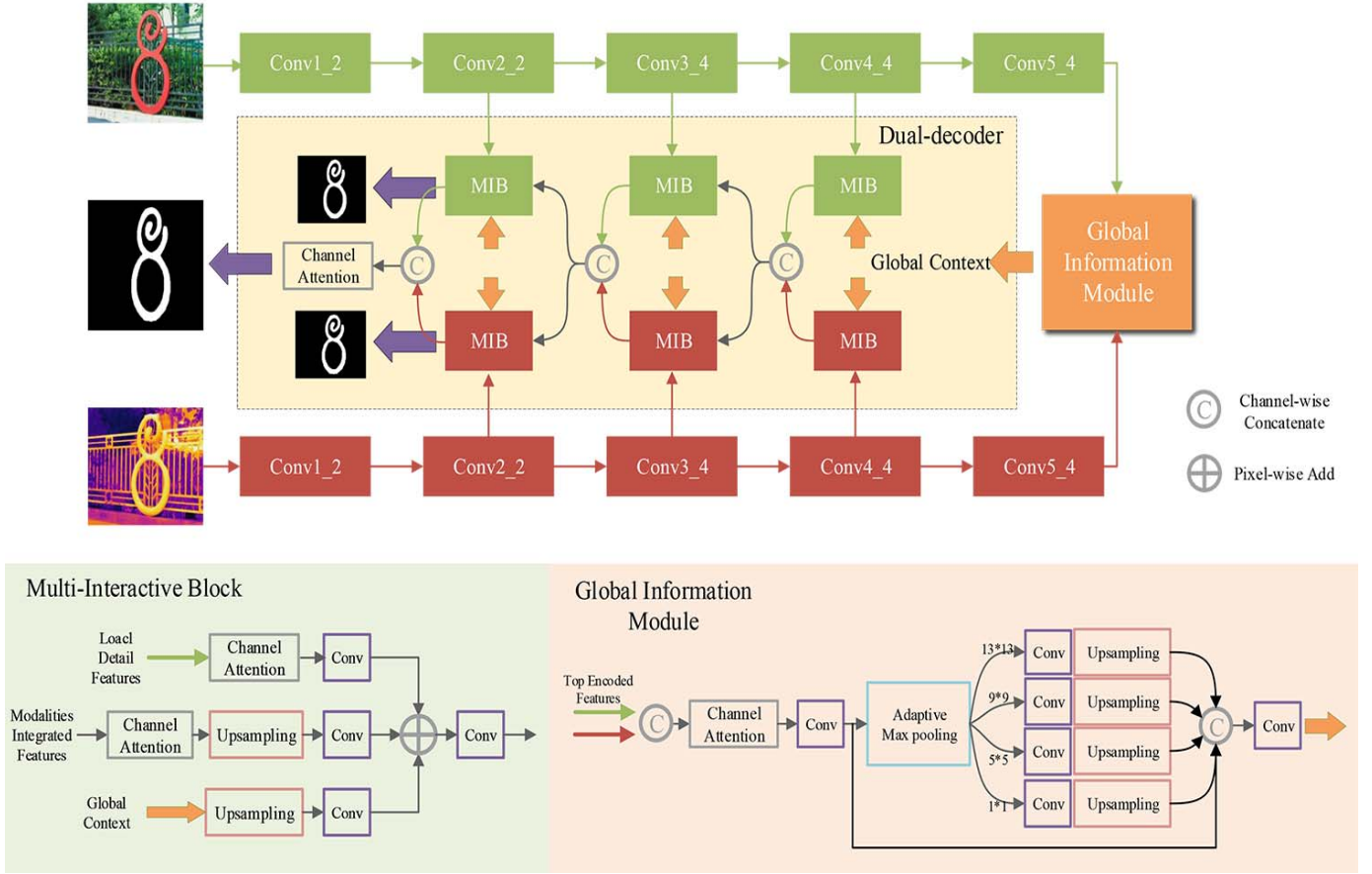
Fig. 2. Framework of our proposed method. The blocks related with RGB image are painted in green and the blocks related with the thermal image are painted in red. The global information is painted in orange. The 'Score' block is a convolutional layer with $1 * 1$ kernel size and outputs predicted saliency map with single channel.

images respectively. Then the global information module can combine the highest-level features of two modalities to obtain the global features with various receptive fields, which are then used as global contexts for accurate location of salient regions. In the decoding phase, we adopt a dual-decoder based on two modalities to achieve saliency computing progressively. To be specific, we design the multi-interaction block (MIB) and then embed it into the dual-decoder in a cascade way to achieve sufficient fusion by multi-type cues. Intuitively, we make two outputs of the dual-decoder be consistent by using the same saliency supervision. Therefore the joint computation is achieved in the progressive interactive decoding phase, while the bias and noise of two modalities are suppressed. Finally, we fuse the final features of dual-decoder to predict the final saliency map. Details of each part are presented in the following subsections.

B. Encoder Network

For universality and simplicity, we use VGG16 [33] as our encoder to extract hierarchical features with different resolutions from the images of two modalities. Herein, we remove the last pooling layer and two fully connected layers in VGG16. As we know, features from deeper layers encode high-level semantic information while features from shallower

layers contain more spatial details. The features from the shallowest layer capture lots of details which are not conducive to saliency prediction, and extracting these features has a high computational complexity. To improve the efficiency and performance, we abandon these features of each modality. For convenience, the remaining features extracted from RGB image are denoted as $R_2 \sim R_5$ and those from thermal infrared image are denoted as $T_2 \sim T_5$.

C. Global Information Module

Global context is vital for locating the regions in RGBT SOD task. Region based context is helpful for maintaining the completeness of salient region and suppressing background noise. Therefore we should extract the global features with multiple receptive fields for performance boosting. Inspired by the pyramid pooling module (PPM) [34] which is widely used in capturing multiple region contexts, we simply change PPM and then embed it into our network. The details can be seen in the bottom right of Fig. 2. We collect the top encoded features of two modalities (R_5 and T_5) and then concatenate them in a channel-wise way. A channel attention mechanism is used for selective recombination of the two features. For simplification, we use the channel attention in CBAM [35]. And we replace the multiple perceptrons in CBAM by $1 * 1$

convolution layer.

$$Y = \sigma(f_1(AvgPooling_1(X)) + f_1(MaxPooling_1(X))) * X. \tag{1}$$

where $f_1(*), AvgPooling_1(*)$ and $MaxPooling_1(*)$ are respectively represent $1 * 1$ convolution layer, global average pooling and global max pooling. X is the input feature and Y is the output feature. The $\sigma(*)$ is sigmoid function that maps the value to the range of 0 to 1. For simplification, this channel attention operation is marked as $CA(*)$ in subsequent formulas.

The combination of convolution layer, batch normalization [36] and Relu [37] is marked as $Conv(*)$. We adopt a convolutional block to decrease the channel number to 256, and the output is marked as F:

$$F = Conv(CA([R_5, T_5])). \tag{2}$$

where $[*]$ is the channel-wise concatenation. Then four operations of adaptive global max pooling with different sizes are used to obtain four feature maps with different receptive fields. We use four convolutional blocks to reconstruct four feature maps respectively. After that, we up-sample all the four feature maps to the size of F and then concatenate them with F. Finally, we apply a convolutional layer to the concatenated features and generate the reconstructed features G which contain the information from the global receptive field.

$$\tilde{F}_i = UP(Conv(MaxPooling_n(F))). \tag{3}$$
$$G = Conv([\tilde{F}_i]), i = 1, \dots, 4. \tag{4}$$

where $UP(*)$ is the up-sampling operation and $MaxPooling_n(*)$ is an adaptive max pooling with $n * n$ output size. $\tilde{F}_i$ represents the i-th branch's output where $i = 1, 2, 3, 4$ with $n = 1, 5, 9, 13$ respectively. And the output G is the global contexts shown in Fig. 2.

D. Dual-Decoder Network

The architecture of our Dual-decoder can be seen in Fig. 2. In general, the hierarchical encoded features are always useful for the up-sampling of decoded features. And the global context is vital for region location. Therefore, we adopt a multi-interactive manner to make full use of multi-type cues. Specially, we design a multi-interaction block (MIB) and embed it into the decoder in a cascade way, which can achieve the interactions of dual modalities, hierarchical features and global contexts. The details of MIB are shown as the bottom left of Fig. 2. In this subsection, we will focus on the three kinds of interactions to present MIB in dual-decoder.

1) Interaction With Hierarchical Features: In each decoder stream, we use the encoded features to restore the spatial details step by step. We use the channel attention to emphasize the more useful features, and then decrease the number of channels to 128.

$$\tilde{A}_i = Conv(CA(A_i)), i = 2, 3, 4. \tag{5}$$

where A_i represents the corresponding RGB encoded features R_i or the thermal encoded features T_i. The corresponding

(a) Image Pairs (b) Spatial Details (c) Modal Fusion Context (d) Global Context (e) Outputs
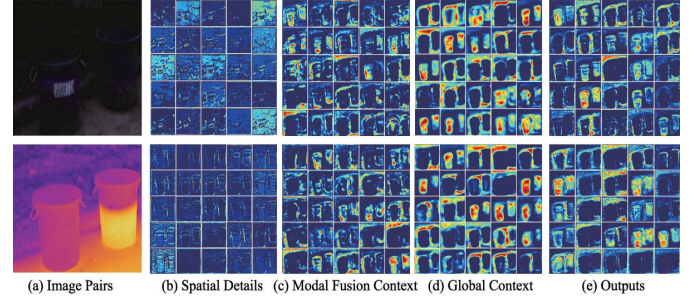
Fig. 3. Feature visualization in the second group MIBs in dual-decoder. We present the feature maps of the first 25 channel for each feature. (a) Input image pairs. (b) Spatial details ($\tilde{A}_i$) computed by encoded features (A_i). (c) Fused features of modalities ($\tilde{M}_i$). (d) Global contexts ($\tilde{G}_i$) adjusted to MIB. (e) Outputs of MIBs in two branches.

encoded features in current decode stream are used for refining the spatial details of the previous output which has been up-sampled. Because of adopting the encoded features in specific modality, the outputs of the two decode streams can maintain strong characteristics of corresponding modalities. Thus we can effectively prevent useful details from fading away after the fusion step. For example, in Fig. 3, we visualize the spatial details used in the second group MIBs in dual-decoder. From the maps in the column (b), we can see there are abundant details in encoded features of shallower layers. And in column (e), the modality characteristic with abundant details can be maintained.

2) Interaction With Dual Modalities: For the interactions of dual modalities, we concatenate the previous MIB outputs from two decoders marked as M_i. These output features contain previous fused information and the specific information of two modalities. In the top-down path, they are regarded as high-level semantic features with strong relevance of pixels. Then we use the channel attention to adaptively select useful information and achieve the feature reconstruction from the concatenated features. We up-sample the reconstructed features to the size as same as A_i and adopt a convolutional block to reduce the number of channels of reconstructed features to 128. Thus, we get $\tilde{M}_i$.

$$\tilde{M}_i = Conv(UP(CA(M_i))), i = 2, 3, 4. \tag{6}$$

Here, we make the fusion of modalities by a channel attention because we want to prevent the multi-modal information from excessive disturbance by each other. And we think the decoder can implicitly achieve the complementarity between modalities as the dual-decoder is driven by the same saliency supervision. In Fig. 3, the third column shows the fused features of the modalities in the second group MIBs. We can see that the deficient features in RGB modality have been complemented by the thermal modality. And the features will be polished to be better in subsequent steps in the decoder.

3) Interaction With Global Contexts: In the top-down path, with the fusion of spatial details, the semantic information of high levels will be gradually diluted [38], which leads to inaccurate positioning for salient regions, as positioning for salient regions depends on high-level semantic information. Besides, although the high-level encoded features model the

semantic relevance of pixels, the receptive field of VGG16 is too small so that the network cannot obtain sufficient semantic information, which might lead to missing some parts in the big object or multiple objects. To handle these problems, we design the global information module as Section. III-C introduced. G is the global context computed by global information module. We integrate it into each MIB to maintain the location of salient region. In detail, we up-sample G to the size of A_i and use a convolutional block to decrease the number of channels to 128. Therefore the global context can be adjusted to interact with other information.

$$\tilde{G}_i = Conv(UP(G)), i = 2, 3, 4. \qquad (7)$$

Obviously, in Fig. 3 we can find that the global context effectively highlights the salient region. With the global context interacted, the main part of salient region is emphasized and the noise of background is suppressed.

4) Integration of Multi-Type Cues: We sum the above three kinds of features directly and obtain the reconstructed fused features Z through a convolutional block as the output of a *MIB*:

$$Z_i = Conv(\tilde{M}_i + \tilde{G}_i + \tilde{A}_i), i = 2, 3, 4. \qquad (8)$$

On the one hand, the pixel-wise addition does not need any parameter and the calculation is simple and efficient. On the other hand, the operation of pixel-wise addition means that we treat these features equally, which avoid the network learning a bias on one kind of features.

Other MIBs have the same operation except for different inputs. It should be noted that the interactions with dual modalities only perform on the outputs of previous MIBs in two paths, and we still use the original features of the dual modalities to obtain the low-level spatial details. Therefore, the final features from both of decoders can effectively maintain the characteristics of the corresponding modality to some extent and also contain valid information. Through the three steps of cascaded decoder, the interactions of dual modalities, hierarchical features and global contexts can be performed well, and the resolution of the last MIB's output will be restored to 1/4 of the input image.

Finally, we fuse the final features from MIBs in dual-decoder by the concatenation and a simple channel-wise attention to predict the final saliency map. And then we up-sample the saliency map four times to the same size of the input image, marked as S_f. Besides, we also use these outputted features to predict two branches of saliency maps respectively with the same supervision operation. Therefore, although the two branches of dual-decoder receive different type of cues, they trend to infer the same salient regions, benefited from the interactions in dual-decoder. By this way, we can explore the complementarity of dual modalities and eliminate their bias.

We visualize the whole procedure of dual-decoder in Fig. 4. We can see that the outputs of MIBs are gradually improved with more accurate details. And the features in two branch are trending to be consistent, which verifies the ability of our network for complementarity of modalities and eliminating the modalities bias.

(a) Image Pairs   (b) MIB₂   (c) MIB₃   (d)MIB₄   (e) Predicted Maps
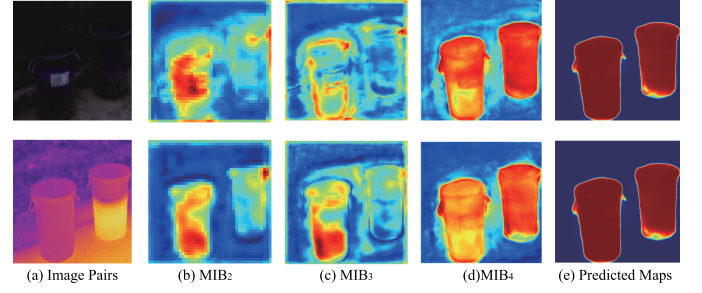
Fig. 4. Visualization of all MIBs' outputs and the final predicted saliency maps of dual-decoder. The presented feature maps are averaged in a channel-wise way. Herein, MIB_i is the group of MIBs corresponding to the i-th level encoded features.

E. Loss Function

Given the predicted saliency map $S = \{S_i | i = 1, \ldots, T\}$ and the corresponding ground truth $Y = \{Y_i | i = 1, \ldots, T\}$, where T is number of total pixels, the binary cross entropy (BCE) loss commonly used in SOD task is formulated as follows:

$$\mathcal{L}(S, Y) = -\sum_{i=1}^{T} (Y_i * log(S_i) + (1 - Y_i) * log(1 - S_i)). \quad (9)$$

We predict two saliency maps from the two branches of dual-decoder denoted as S_1 and S_2, and then compute the BCE loss with the ground truth Y:

$$\mathcal{L}_d = \mathcal{L}(S_1, Y) + \mathcal{L}(S_2, Y). \qquad (10)$$

To make the global information module be learned better, we predict a saliency map S_g from the global context G. For the same size, we down-sample Y with the factor of 16 to obtain Y_g. Then, a BCE loss is used:

$$\mathcal{L}_g = -\sum_{i=1}^{T_g} (Y_{gi} * log(S_{gi}) + (1 - Y_{gi}) * log(1 - S_{gi})).(11)$$

where the T_g is the number of total pixels of S_g. For the final predicted map S_f, the loss function is:

$$\mathcal{L}_f = \mathcal{L}(S_f, Y). \qquad (12)$$

Furthermore, we use the smoothness loss [39] as a constraint to achieve region consistency and obtain clearer edges. We compute first-order derivatives of the saliency map in the smoothness term as [40] does.

$$\mathcal{L}_s = \sum_{i=1}^{T} \sum_{d \in \vec{x}, \vec{y}} \Psi \left(\left| \partial_d S_{f_i} \right| e^{-\alpha |\partial_d Y_i|} \right). \qquad (13)$$

$$\Psi(s) = \sqrt{s^2 + 1e^{-6}}. \qquad (14)$$

where ∂_d represents the partial derivatives on $\vec{x}$ and $\vec{y}$ directions. And we set $\alpha = 10$ as [40] does. Therefore, our total loss is:

$$\mathcal{L} = \mathcal{L}_d + \mathcal{L}_g + \mathcal{L}_f + \beta \mathcal{L}_s. \qquad (15)$$

We empirically set $\beta = 0.5$ to balance the effect of smoothness loss. Our network can be trained well with the cooperation of this four kinds of constraints.

In general, considering modality weight in loss function is a suitable way to deal with the difference of modalities. However, in our method, we design the dual-decoder to decode from two modalities with equal importance, which aims at capturing more consistent saliency features by gradually interacting. When coming a defective input, we don't want to weak the learning ability by decreasing the weight of loss function. We prefer to train the dual-decoder to adapt to these cases and it can also capture more consistent features even with a defective input. Therefore, we consider the effectiveness of the modality during data enhancement, which aims at creating more defective samples to train our dual-decoder.

F. Noisy Data Augmentation

We observe that some modalities in RGBT image pairs are not always useful. For example, the RGB modality is indiscernible when the image is captured in low illumination while the thermal modality is invalid when the foreground and background have close temperature. Even in the complex scene, both of the two modalities have large noises. Therefore, we design a simple yet effective data augmentation strategy to train the proposed network. We randomly set one modality to a zero map or noisy maps sampled from the standard normal distribution. The zero map is used to simulate the case that there is one invalid modality. Therefore the network cannot extract any useful information from it. The noisy maps are used to simulate the noise modality and the network can thus obtain lots of useless information. Through training on this kind of data, our network can learn to conquer these difficulties and to care more about the interaction between modalities. In our implementation, we set the probability at 10% to randomly use zero map or noisy maps. Both of two modalities have the probability of 50% to be replaced and the probability of two kinds of maps are also 50%.

IV. EXPERIMENT

In this section, we give the details of our experiments. At first, we introduce the used datasets. Then we show the experimental setup including implementation details and evaluation metrics. Finally, we conduct and analyze the comparison experiments and ablation experiments to demonstrate the effectiveness of our method.

A. Dataset

There are three RGBT SOD dataset publicly available, including VT821 [6], VT1000 [8] and VT5000 [9]. VT821 contains 821 registered RGBT image pairs. To strengthen the challenge of the dataset, some noises are added to some images. As RGBT image pairs in VT821 are registered manually, there are vacant regions in the thermal infrared images. VT1000 contains 1000 RGBT image pairs, with relatively simple scenes and well aligned images. VT5000 collects 5000 aligned RGBT image pairs, and has more complex scenes and various objects. There are many challenges in above datasets including big salient object (BSO), small salient object (SSO), multiple salient object (MSO), low illumination (LI),

center bias (CB), cross image boundary (CIB), similar appearance (SA), thermal crossover (TC), image clutter (IC), out of focus (OF) and bad weather (BW), which basically covers all the problems in RGBT SOD. In this work, 2500 various image pairs in VT5000 are chosen as our training set, and the rest image pairs together with VT821 and VT1000 are taken as testing sets.

B. Experimental Setup

1) Implementation Details: Our network is implemented based on Pytorch and trained with a single Titan Xp GPU. We use the stochastic gradient descent (SGD) to optimize parameters with the weight decay of 5e-4 and the momentum of 0.9. We train 100 epochs with batch size of 4. The initial learning rate is 1e-3, and it becomes 1e-4 after 20 epochs and 1e-5 after 50 epochs. For the RGBT inputs, we resize all the images to the size of $352 * 352$. In addition, we compute the mean and standard deviation of all the images in training set and use them to normalize the inputs.

2) Evaluation Metrics: For evaluating different methods, we use F-measure, S-measure, E-measure [41] and mean absolute error [42], which are widely used in SOD.

The formula of F-measure is expressed as follows:

$$F_m = \frac{(1 + \beta^2) \cdot Precision \cdot Recall}{\beta^2 \cdot Precision + Recall}. \quad (16)$$

where $\beta^2 = 0.3$ emphasizes the importance of precision, suggested by [43]. The $Precison$ is the ratio of the correctly predicted foreground pixels to the totally predicted foreground pixels. The $Recall$ is the ratio of the correctly predicted foreground pixels to the ground truth. The mean value of the predicted saliency map is doubled as the threshold to binarize the saliency map, and then F-measure can be computed and marked as F_m. In addition, we follow [44] to compute the weighted F-measure as another metric marked as wF. For PR curve, we equally divide the range of saliency score to x parts as the thresholds to binarize the predicted saliency map and then we calculate corresponding $Precison$ and $Recall$. In our evaluation code, we set x to 20. Thus 20 F-measure values can be calculated and the PR curve can be plotted.

The formula of mean absolute error (MAE) is expressed as follows:

$$MAE = \frac{1}{T} \sum_{i=1}^{T} |S_i - Y_i|. \quad (17)$$

where S_i is a predicted saliency map and Y_i is the ground truth. T is the total number of pixels in a map. Thus MAE can evaluate the difference between the predicted map and its ground truth. We average all the MAE values of test samples as another evaluation metric.

S-measure(S_m) is used to evaluate the similarity of spatial structure, which combines the region-aware structural similarity S_r and the object-aware structural similarity S_o:

$$S_m = \alpha * S_o + (1 - \alpha) * S_r. \quad (18)$$

where we set $\alpha = 0.5$, more details are introduced in [45].

TABLE I

PERFORMANCE COMPARISON WITH 12 METHODS ON THREE TESTING DATASETS. THE BEST SCORES ARE HIGHLIGHTED IN **RED**, THE SECOND BEST SCORES ARE HIGHLIGHTED IN **GREEN**, AND THE THIRD BEST SCORES ARE HIGHLIGHTED IN **BLUE**

Methods	VT821					VT1000					VT5000				
	Em	Sm	Fm	MAE	wF	Em	Sm	Fm	MAE	wF	Em	Sm	Fm	MAE	wF
MTMR	0.815	0.725	0.662	0.108	0.462	0.836	0.706	0.715	0.119	0.485	0.795	0.680	0.595	0.114	0.397
M3S-NIR	0.859	0.723	0.734	0.140	0.407	0.827	0.726	0.717	0.145	0.463	0.780	0.652	0.575	0.168	0.327
SGDL	0.847	0.765	0.730	0.085	0.583	0.856	0.787	0.764	0.090	0.652	0.824	0.750	0.672	0.089	0.559
ADF	0.842	0.81	0.716	0.077	0.627	0.921	0.910	0.847	0.034	0.804	0.891	0.864	0.778	0.048	0.722
DMRA	0.691	0.666	0.577	0.216	0.546	0.801	0.784	0.716	0.124	0.699	0.696	0.672	0.562	0.195	0.532
S2MA	0.834	0.829	0.723	0.081	0.702	0.914	0.921	0.852	0.029	0.850	0.869	0.855	0.751	0.055	0.734
PFA	0.756	0.761	0.592	0.096	0.526	0.809	0.813	0.688	0.078	0.635	0.737	0.748	0.563	0.099	0.498
R3Net	0.803	0.782	0.681	0.081	0.656	0.903	0.886	0.835	0.037	0.831	0.856	0.812	0.729	0.059	0.703
BASNet	0.856	0.823	0.735	0.067	0.716	0.923	0.909	0.847	0.030	0.861	0.878	0.839	0.764	0.054	0.742
PoolNet	0.811	0.788	0.652	0.082	0.573	0.852	0.849	0.751	0.063	0.690	0.809	0.788	0.643	0.080	0.570
CPD	0.843	0.818	0.718	0.079	0.686	0.923	0.907	0.863	0.031	0.844	0.894	0.855	0.787	0.046	0.748
EGNet	0.856	0.830	0.726	0.063	0.662	0.922	0.910	0.848	0.033	0.817	0.888	0.853	0.775	0.050	0.712
MIDD	0.895	0.871	0.804	0.045	0.760	0.933	0.915	0.882	0.027	0.856	0.897	0.868	0.801	0.043	0.763

E-measure(E_m) is an enhanced alignment measure, proposed recently in [41], which can jointly capture statistic information in image level and matching information in pixel level. By using all above metrics, we can make a comprehensive evaluation of our method.

C. Comparison With State-of-the-Art Methods

We compare our method with 12 existing methods as follows. Three traditional RGBT SOD methods are SDGL [8], MTMR [6] and M3S-NIR [23]. A deep learning based RGBT SOD method is ADF [9]. Two deep learning based RGBD SOD methods are DMRA [30] and S2MA [46]. And six deep learning based single-modality SOD methods include R3Net [17], PFA [47], CPD [18], EGNet [19], PoolNet [38] and BASNet [48]. To keep the architectures of those single-modality SOD methods well, we apply early fusion strategy to them for fairness. In addition, it should be mentioned that our method doesn't use any post-processing such as fully connected conditional random fields (CRF) [49] used in R3Net [17].

1) Quantitative Evaluation: Tab.I shows the results of our method and other eleven methods on our testing sets. First, it can be seen that our method outperforms well against the existing four RGBT SOD methods with the above metrics. Especially in VT821, our method shows tremendous superiority. The performance of three traditional methods are inferior to the deep learning based methods, which is mainly caused by the weakness of feature representations and the limitation of super-pixel representations.

Second, DMRA [30] is a state-of-the-art RGBD SOD method which has been proven to be effective on RGBD dataset. It uses the middle fusion strategy to fuse two modalities. We train it on VT5000 training set, but it shows the poorer performance on RGBT data. The S2MA [46] is the latest RGBD SOD method. From the Tab. I, we can find that it can also perform well on RGBT SOD datasets. We have made an analysis that the RGBD SOD task is different from RGBT as the former focuses on modalities complementary and the depth maps are used as auxiliary information, while the later focuses on joint inferring to eliminate the bias between modalities and the thermal maps are with equivalent importance. Therefore RGBD methods can be used on RGBT task, but it is not impeccable. Although DMRA symmetrically fuses features of two modalities, it also adopts a depth-specific module to adjust the features. However, the thermal maps cannot be used for that.

Third, we study six single-modality SOD methods, which have the most advanced performance. Because most of visible images have useful information, these methods also work well on simple cases. The early fusion strategy regards two modalities as an integrated information. The relationship between modalities can be explored in encoding phase, and the modal complementarity can thus also be achieved. Then benefiting from the various advanced segmentation tricks, these methods also work well on RGBT SOD. A further study for the performance of our method will be taken in the ablation experiments.

2) Qualitative Evaluation: The qualitative comparison can be seen in Fig. 5. We can see our method outperforms all comparison methods in the challenging image pairs. Our network has considered the specialties of two modalities and reasonably integrated features with multi-type cues. Although one of modalities is useless, the other modality is rarely affected by the destructive information. When both of two modalities are informative, our network can take advantage of the complementarity of two modalities to gain more reliable relevance in pixels and make a consistent prediction by suppressing the bias. Even both of modalities are unreliable such as the 6*th* column in Fig. 5, our method can combine the useful information in two modalities to make a better prediction. Those advanced SOD methods work well in informative visible images, but they can't deal with the images with deficiency and semantic ambiguity. As we can see in Fig. 5,

TABLE II

PERFORMANCE COMPARISON WITH 12 METHODS ON 11 CHALLENGES AND 2 QUALITY ATTRIBUTIONS. THE BEST SCORES ARE HIGHLIGHTED IN RED, THE SECOND BEST SCORES ARE HIGHLIGHTED IN GREEN, AND THE THIRD BEST SCORES ARE HIGHLIGHTED IN BLUE

|	BSO	CB	CIB	IC	LI	MSO	OF	SSO	SA	TC	BW	Bad RGB	Bad T
DMRA	0.795	0.675	0.736	0.662	0.716	0.663	0.690	0.355	0.675	0.652	0.648	0.687	0.640
S2MA	0.873	0.817	0.860	0.794	0.859	0.805	0.844	0.735	0.811	0.810	0.773	0.825	0.810
PFA	0.803	0.746	0.743	0.735	0.747	0.73	0.754	0.689	0.725	0.763	0.666	0.730	0.755
R3Net	0.829	0.790	0.821	0.74	0.787	0.775	0.757	0.657	0.727	0.725	0.747	0.735	0.715
BASNet	0.859	0.808	0.824	0.774	0.832	0.799	0.816	0.716	0.763	0.792	0.768	0.787	0.787
PoolNet	0.801	0.725	0.741	0.720	0.753	0.708	0.761	0.658	0.728	0.718	0.749	0.732	0.716
CPD	0.873	0.845	0.863	0.811	0.840	0.829	0.823	0.766	0.828	0.811	0.793	0.809	0.801
EGNet	0.874	0.838	0.856	0.817	0.848	0.818	0.819	0.700	0.794	0.790	0.772	0.792	0.775
MTMR	0.666	0.574	0.579	0.562	0.688	0.621	0.707	0.697	0.652	0.567	0.605	0.668	0.562
M3S-NIR	0.645	0.563	0.571	0.554	0.700	0.597	0.709	0.600	0.609	0.560	0.636	0.557	0.640
SGDL	0.755	0.703	0.693	0.680	0.740	0.709	0.74	0.755	0.656	0.674	0.641	0.672	0.662
ADF	0.881	0.853	0.862	0.834	0.868	0.839	0.839	0.805	0.837	0.841	0.804	0.820	0.832
MIDD	0.891	0.869	0.869	0.840	0.878	0.843	0.861	0.813	0.855	0.850	0.812	0.844	0.838
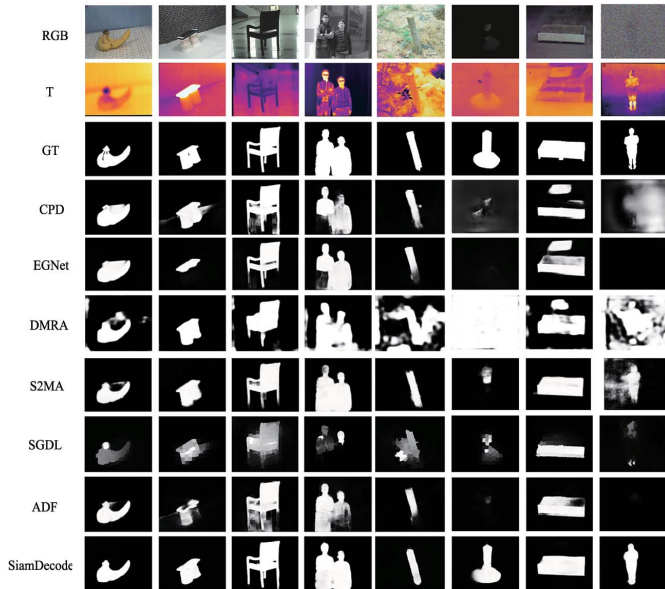
Fig. 5. Qualitative comparison of the proposed method with other methods. We select 8 RGBT image pairs with diverse challenges to compare the quality of the predicted maps.

Fig. 6. The statistics of mean F-measure score of all the methods on 13 challenges.

single-modality SOD methods with early fusion strategy are impressionable to noise. Therefore they always detect the extra wrong pixels or miss the correct pixels. The traditional RGBT methods can accurately detect the salient region but they always have lower confidence. The deep learning based RGBT and RGBD methods have better saliency maps and our method is more robust in challenging cases.

3) Challenge-Based Quantitative Evaluation: We further make a study on all challenges labeled by VT5000. The quantitative comparisons are presented in Tab. II. The first eleven columns are the challenge attributes and the last two columns are quality attributes which present the deficient modalities. As we can see in Tab. II, our method has the best performance on all challenges. We compute the max F-measure scores of all methods on thirteen challenges and use
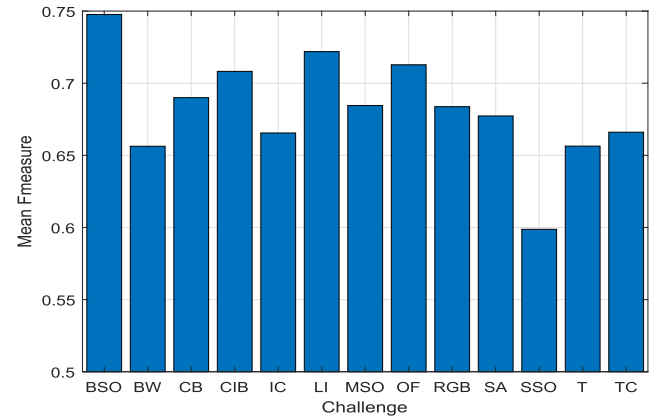
a histogram to present them in Fig. 6. The low illumination (LI), small salient object (SSO) and bad weather (BW) seem to be the most three difficult challenges. Because these cases are more likely to lead a deficient modality or ambiguous context. Once the discrepancy between modalities, the joint inference becomes difficult.

D. Ablation Study

In this section, we mainly study the effects of different configurations on network performance, as shown in Tab. III. These experiments are performed on RGBT dataset. We first explore the supervisions on the two branches of dual-decoder, which is used for guiding the modal complementarity and eliminating the modality bias. Then we study the modality interaction and global context interaction of our network. Since the hierarchical features interaction is necessary for decoder, we do not study for it. Without the modality interaction, our network is degenerated into an original late fusion based framework as shown in the (*a*) of Fig. 7. Furthermore, we study the single decoder stream with the middle fusion strategy on our network as shown in the (*b*) of Fig. 7.

TABLE III
ABLATION STUDIES WITH DIFFERENT SETUPS. w/o MEANS DISABLING THE CORRESPONDING COMPONENT

	VT821				VT1000				VT5000			
	Sm	Em	Fm	MAE	Sm	Em	Fm	MAE	Sm	Em	Fm	MAE
w/o $\mathcal{L}_d$	0.863	0.881	0.790	0.048	0.901	0.912	0.868	0.030	0.860	0.883	0.783	0.047
w/o Noise Data	0.852	0.878	0.776	0.052	0.907	0.928	0.871	0.029	0.856	0.891	0.789	0.046
w/o Global Interaction	0.857	0.873	0.771	0.057	0.892	0.906	0.843	0.036	0.852	0.874	0.764	0.051
w/o Modality Interaction	0.864	0.882	0.781	0.050	0.898	0.911	0.846	0.031	0.861	0.890	0.792	0.045
w/o Channel Attention	0.868	0.888	0.794	0.049	0.915	0.922	0.860	0.029	0.865	0.891	0.793	0.045
Single Decoder	0.862	0.884	0.779	0.051	0.907	0.921	0.868	0.031	0.852	0.875	0.774	0.049
MIDD	**0.871**	**0.895**	**0.804**	**0.045**	0.915	**0.933**	**0.882**	0.027	0.868	0.897	0.801	**0.043**
MIDD(R)	**0.871**	0.882	0.790	0.047	**0.921**	0.927	0.875	**0.025**	**0.874**	0.896	0.800	0.044
MIDD(R+)	0.870	0.882	0.796	0.049	**0.921**	0.929	0.881	0.027	**0.874**	**0.900**	**0.810**	0.044
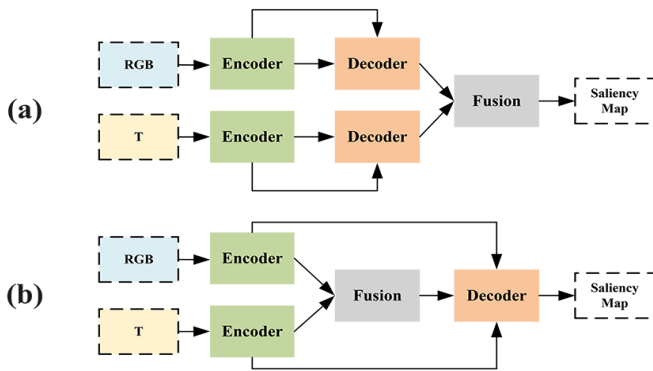
Fig. 7. (a) Framework of 'w/o Modality Interaction'. (b) Framework of 'Single Decoder'.

(a) Image Pairs (b) w/o L$_d$ (c) SiamDecoder

Fig. 8. Without the supervision on two branches of Dual-decoder. The feature maps of two decoder streams are less likely to be consistent.

Then we disable the used data augmentation strategy and verify the effectiveness of it. Finally, we replace VGG16 by ResNet50 [50] as the encoder block of our network. Since the features extracted by ResNet50 have lower resolutions, we further add a MIB block to fusion the bottom encoded features. We will make detailed analysis of these factors in following parts.

1) Supervision Effectiveness in Dual-Decoder: In Tab. III, 'w/o $\mathcal{L}_d$' represents that we don't use the output features of the two decoder branches to predict the saliency maps, and the supervision only exists in the final predicted saliency map. As we can see in Tab. III, the performance in four metrics have declined without $\mathcal{L}_d$. We visualize the feature maps of final decoder step in Fig. 8. When the supervisions are absent, the ability of modality complementarity is weaken. Therefore Dual-decoder cannot cooperate well though the multiple interactions exist.

2) Effectiveness of MIB: In our network, the two decoder streams consist of three MIBs. Each MIB performs the interactions and fusions of multi-type cues. Therefore, to evaluate the ability of MIB, we should explore the usefulness of global information interaction and modality interaction. The global contexts play a vital role in our network because we rely on it to suppress the long range background noises and highlight foreground regions. In addition, it also provide a coarse modality relevance, which is helpful for seeking common salient regions. Without the global context interaction, the average
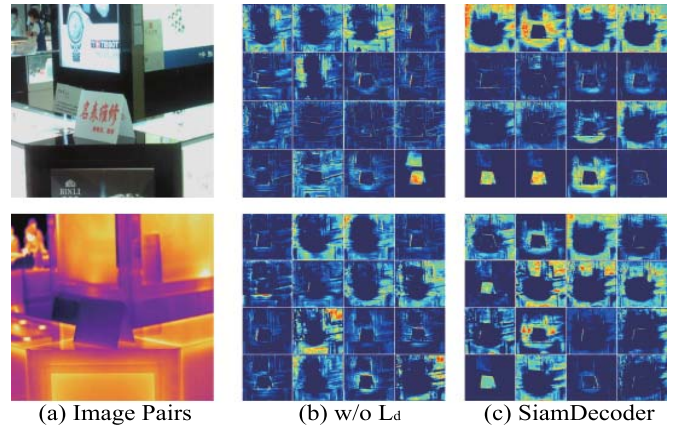
indexes of four metrics(Sm, Em, Fm and MAE) on three test datasets respectively decline 2.0%, 2.7%, 4.6% and 18.1%.

The modality interaction are conducted in all the decoder steps. Without these interactions, the framework is degraded into a common late fusion based methods shown in Fig. 7. We compare the saliency maps of our network with and without this interaction in Fig. 9. As we can see, the saliency maps have larger confidence both in foreground regions and background regions when the modality interaction exists. If we disable modality interaction in dual-decoder and only fusion the output features of dual-decoder to predict the saliency map, the network cannot withstand the noise coming from one deficient modality and has limited ability to eliminate the modality bias, thus leading the uncertain results.

3) Effectiveness of Channel Attention: For the final fusion of features in dual-decoder, we use a channel attention mechanism referred to [35]. It has been proven that this mechanism has the effectivity of selecting more informative channels in [35]. Therefore, we adopt the channel attention to fuse two top decoded features to relieve the distraction of information in channel, which is more reasonable than immediately concatenating or summing features. We further make an ablation study for this channel attention. We abandon the channel attention and directly predict saliency map by using the concatenated feature. The evaluation results are presented at the

TABLE IV
ABLATION STUDIES ON PARAMETERS OF DUAL-DECODER. "SHARED DD" SHARES THE PARAMETERS OF DUAL-DECODER WHILE THE
"INDEPENDENT DD" HAS INDEPENDENT PARAMETERS

	VT821					VT1000					VT5000-Test				
	Em	Sm	Fm	MAE	wF	Em	Sm	Fm	MAE	wF	Em	Sm	Fm	MAE	wF
Independent DD	**0.895**	**0.871**	**0.804**	**0.045**	**0.760**	**0.933**	**0.915**	**0.882**	**0.027**	**0.856**	**0.897**	**0.868**	**0.801**	**0.043**	**0.763**
Shared DD	0.887	0.869	0.792	0.048	0.752	0.931	0.913	0.876	0.030	0.840	0.893	0.865	0.795	0.045	0.748

(a) RGB   (b) Thermal   (c) w/o Modality Interaction   (d) SiamDecoder
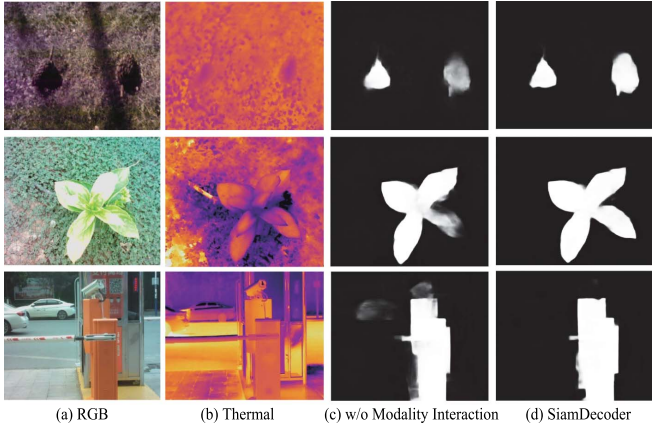
Fig. 9.   Comparison on saliency maps of our network with/without the modality interaction.

Fig. 10.   Max F-measure scores and MAE on VT821 in the whole training procedure to verify the effectiveness of the used data augmentation strategy.

5*th* rows in III. Compared with the 7*th* rows, the performance on three datasets averagely declines $0.4\%, 0.9\%, 1.3\%$ and 6.5% in five metrics. The results suggest that the channel attention is more helpful for decreasing the miss detection and false alarm, which is benefited from the informative channel selection.

4) Effectiveness of Dual-Decoder Network: We further achieve a single decoder based version of our network shown in (*b*) of Fig. 7. This framework maintains the global context interaction and then directly concatenate the hierarchical encoded features of two modalities to interact in a single decoder stream. From the comparisons on different metrics, we can find the performance of this method is obviously inferior to our baseline, which verifies the effectiveness of our dual-decoder network.

Furthermore, we make an ablation study on the parameters of Dual-decoder as shown in Tab. IV. Compared with the shared dual-decoder, the setup of independent parameters has gained the maximum $2.0\%, 0.4\%, 1.5\%, 10.0\%$ and 2.0% respectively in the used four evaluation metrics. Using independent parameters achieves better performance since the dual-decoder is designed to focus on specific saliency of two modalities. With independent parameters, our dual-decoder has the ability to learn to capture the saliency of each modality. And then two streams of decoder gradually interact with each other to obtain common saliency.

5) Effectiveness of Noisy Data Augmentation: Finally, we compare the performance of our network trained with or without noisy data augmentation. As we can see in Tab. III, training with noisy data can improve the performance on four
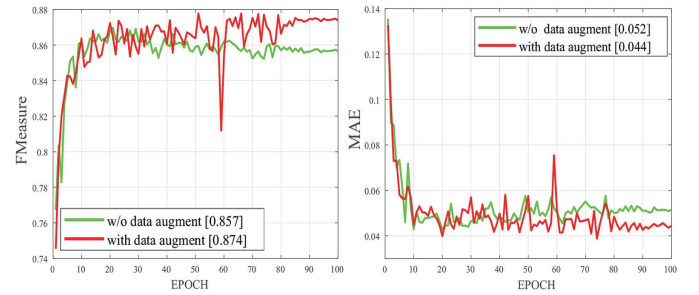
evaluation metrics by average $1.5\%, 1.0\%, 2.1\%$ and 9.0% on three test datasets. We present the results of max F-measure and MAE in the whole training procedure in Fig. 10. The network converges after 90 epochs, and we can find the stable gains when our network trained with noisy data augmentation.

6) Effectiveness of Different Backbones: We replace VGG16 by ResNet50 as the backbone of our network. This setup is marked as "MIDD(R)" in Tab. III. We all know that ResNet50 has stronger ability in feature representation than VGG16, but the performance of two backbones have little difference as we can find in Tab. III. This is mainly caused by the lower resolution of features extracted by ResNet50. The resolution of the output is half of VGG16, which has less spatial details and coarser structure. Since the features extracted by ResNet50 have lower resolutions, we further add a MIB block to fusion the bottom encoded features, which is marked as "MIDD(R+)". The output has higher resolution and more details through the extra MIB, but more background details are also integrated into it. Thus, the "MIDD(R+)" can improve the performance in some aspects but it is not optimal setup. These three setups of our network can show their superiority in different scenes and all of them outperform the state-of-the-art methods.

E. Experiment on RGBD SOD Dataset

1) Dataset: To fully verify the effectiveness of the our method on RGBD SOD, we perform the experiments on seven RGBD SOD benchmark datasets. The NJU2K [51] contains 1,985 RGBD image pairs with labeled saliency maps, which are collected from the Internet and 3D movies or taken by a Fuji W3 stereo camera. NLPR [52] collects 1,000 RGBD image pairs captured by Microsoft Kinect. DES [53] is also named RGBD135 since it contains 135 RGBD image pairs about indoor scenes collected by Microsoft Kinect. LFSD [54] has 100 image pairs with depth information and labelled

TABLE V

PERFORMANCE COMPARISON WITH TEN METHODS ON 5 RGBD SOD DATASETS. THE BEST SCORES ARE HIGHLIGHTED IN **Red**, THE SECOND BEST SCORES ARE HIGHLIGHTED IN **Green**, AND THE THIRD BEST SCORES ARE HIGHLIGHTED IN **Blue**

	DES				LFSD				SIP				SSD				STERE			
	Sm	Em	Fm	MAE	Sm	Em	Fm	MAE	Sm	Em	Fm	MAE	Sm	Em	Fm	MAE	Sm	Em	Fm	MAE
DF	0.752	0.877	0.753	0.093	0.791	0.844	0.806	0.138	0.653	0.794	0.673	0.185	0.747	0.812	0.724	0.142	0.757	0.838	0.742	0.141
PCF	0.842	0.912	0.782	0.049	0.794	0.842	0.792	0.112	0.842	0.899	0.825	0.071	0.841	0.886	0.791	0.062	0.875	0.897	0.826	0.064
CTMF	0.863	0.911	0.778	0.055	0.796	0.851	0.782	0.119	0.716	0.824	0.684	0.139	0.776	0.838	0.710	0.099	0.848	0.864	0.771	0.086
MMCI	0.848	0.904	0.762	0.065	0.787	0.840	0.779	0.132	0.833	0.886	0.795	0.086	0.813	0.860	0.748	0.082	0.873	0.901	0.829	0.068
AFNet	0.770	0.874	0.730	0.068	0.738	0.810	0.742	0.133	0.720	0.815	0.705	0.118	0.714	0.803	0.694	0.118	0.825	0.886	0.807	0.075
TANet	0.858	0.919	0.795	0.046	0.801	0.845	0.794	0.111	0.835	0.893	0.809	0.075	0.839	0.879	0.767	0.063	0.871	0.906	0.835	0.060
DMRA	0.867	0.920	0.806	0.033	0.767	0.824	0.792	0.111	0.800	0.858	0.815	0.088	0.857	0.892	0.821	0.058	0.834	0.899	0.844	0.066
D3Net	0.898	0.951	0.870	0.031	0.832	0.833	0.801	0.099	0.86	0.902	0.835	0.063	0.857	0.897	0.814	0.058	0.899	0.920	0.859	0.046
S2MA	0.910	0.888	0.776	0.033	0.837	0.863	0.820	0.094	0.872	0.911	0.854	0.057	0.708	0.781	0.633	0.138	0.890	0.907	0.855	0.051
cmSalGAN	0.913	0.948	0.869	0.028	0.830	0.870	0.831	0.097	0.865	0.902	0.849	0.064	0.791	0.851	0.717	0.086	0.896	0.914	0.863	0.050
MIDD	0.916	0.972	0.913	0.022	0.846	0.875	0.839	0.080	0.862	0.908	0.856	0.057	0.814	0.890	0.785	0.066	0.894	0.922	0.870	0.041
MIDD(R)	0.903	0.962	0.906	0.025	0.858	0.884	0.852	0.074	0.869	0.913	0.866	0.054	0.817	0.904	0.802	0.062	0.895	0.921	0.872	0.043
MIDD(R+)	0.915	0.971	0.917	0.022	0.85	0.882	0.844	0.080	0.883	0.926	0.879	0.047	0.837	0.901	0.805	0.056	0.901	0.927	0.881	0.040

ground truths. SSD [55] is also a small dataset that contains 80 image pairs picked up from three stereo movies. SIP [56] collect 1,000 image pairs in outdoor scenarios and it has many challenging situations. STERE [57] also contains 1000 various image pairs which include real-world scenes and virtual scenes. We randomly sample 1485 image pairs and 700 image pairs from the NJU2K [51] and NLPR [52] datasets as the training set, which is common used in RGBD SOD methods. For fairness, we test on DES, LFSD, SIP, SSD and STERE.

2) Experiment Setup: We randomly sample 1485 image pairs and 700 image pairs from the NJU2K [51] and NLPR [52] datasets as the training set, which is common used in RGBD SOD methods. For fairness, we test on DES, LFSD, SIP, SSD and STERE.

We compare 10 applicable RGBD SOD methods which include DF [24], PCF [28], CTMF [58], MMCI [59], AFNet [60], TANet [29], DMRA [30], D3Net [56], S2MA [46] and cmSalGAN [61]. These methods relatively equally deal with RGB and depth, and thus it is fair to compare with our methods. We use the available codes with provided models to predict saliency maps or directly use the provided saliency maps to make evaluations on S-measure(Sm), E-measure(Em), F-measure(Fm) and MAE. For our method, we follow the experimental setup of RGBT SOD except that the input size is adjust to 256 and the noisy data augmentation strategy is not used since the RGBD image pairs have less possibility of catching invalid informations. In addition, we compute the mean and standard deviation of all the images in training set and use them to normalize the inputs.

3) Quantitative Evaluation: As we can see in Tab. V, our method also performs well on RGBD datasets. Compared with ten RGBD SOD methods, our method can even surpass the most of them. On SSD, the performance doesn't follow the law of other datasets, which is mainly caused by the less samples and lower otherness in samples of SSD. Thus the

performance shows lower stability. On the other datasets, our method can basically outperform other compared methods. And our method based on ResNet50 [50] with four MIBs can further improve the performance on larger datasets or higher resolution samples like SIP and STERE.

V. FUTURE WORKS

The extended tasks of salient object detection like RGB-T and RGB-D SOD have been explored a lot in recent years. However, there are still many potential problems have not been solved well. For example, in this paper, although we consider the saliency bias of two modalities by implicitly learning, we think it is more meaningful to establish the correlation of two modalities explicitly. Furthermore, the captured image pairs are not naturally aligned. Considering constructing a large scale dataset with aligned images costs lots of labour, therefore, it is necessary to propose the alignment-free multi-modal SOD algorithms. With the success of co-saliency works such as [62], [63], we deem that alignment-free multi-modal SOD is theoretically feasible. Moreover, the uncertainty of visual saliency is also a noteworthy problem, which has been first studied by [64]. Besides above mentioned saliency related problems, a kind of reverse study is introduced recently, which is called camouflaged object detection. Camouflaged object detection aims to identify objects embedded in their surroundings, which has been studied in [65], [66]. Addressing this task requires the visual perception knowledge, and this new issue deserves to be further explored.

VI. CONCLUSION

In this paper, we propose a multi-interactive dual-decoder network for RGBT SOD task. Considering the different relevance among the dual modalities, the encoded hierarchical features and the global contexts, we design a dual-decoder

network with cascaded multi-interactive modules to achieve sufficient fusion of different source data. The proposed network can prevent information of two modalities from excessive influence by each other during the interactions. And with the same supervision, two decoder branches trend to be consistent, and the bias of dual modalities is thus suppressed implicitly. Experimental results show the superiority of our method on both RGBT SOD and RGBD SOD. Furthermore, we discuss more potential problems which are worth to study and we will further explore to solve these problems in the future.

REFERENCES

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

[2] C. Li, X. Liang, Y. Lu, N. Zhao, and J. Tang, "RGB-T object tracking: Benchmark and baseline," *Pattern Recognit.*, vol. 96, Dec. 2019, Art. no. 106977.

[3] A. Lu, C. Li, Y. Yan, J. Tang, and B. Luo, "RGBT tracking via multi-adapter network with hierarchical divergence loss," *IEEE Trans. Image Process.*, 2021, doi: 10.1109/TIP.2021.3087341.

[4] M. Ye, Z. Wang, X. Lan, and P. C. Yuen, "Visible thermal person re-identification via dual-constrained top-ranking," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, p. 2.

[5] Y. Hao, N. Wang, J. Li, and X. Gao, "HSME: Hypersphere manifold embedding for visible thermal person re-identification," in *Proc. AAAI Conf. Artif. Intell.*, Jul. 2019, pp. 8385–8392.

[6] G. Wang, C. Li, Y. Ma, A. Zheng, J. Tang, and B. Luo, "RGB-T saliency detection benchmark: Dataset, baselines, analysis and a novel approach" in *Proc. Chin. Conf. Image Graph. Technol.*, 2018, pp. 359–369.

[7] J. Tang, D. Fan, X. Wang, Z. Tu, and C. Li, "RGBT salient object detection: Benchmark and a novel cooperative ranking approach," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 12, pp. 4421–4433, Dec. 2020.

[8] Z. Tu, T. Xia, C. Li, X. Wang, Y. Ma, and J. Tang, "RGB-T image saliency detection via collaborative graph learning," *IEEE Trans. Multimedia*, vol. 22, no. 1, pp. 160–173, Jan. 2020.

[9] Z. Tu, Y. Ma, Z. Li, C. Li, J. Xu, and Y. Liu, "RGBT salient object detection: A large-scale dataset and benchmark," 2020, *arXiv:2007.03262*. [Online]. Available: https://arxiv.org/abs/2007.03262

[10] Q. Zhang, N. Huang, L. Yao, D. Zhang, C. Shan, and J. Han, "RGB-T salient object detection via fusing multi-level CNN features," *IEEE Trans. Image Process.*, vol. 29, pp. 3321–3335, Dec. 2020.

[11] L. Wang, H. Lu, X. Ruan, and M.-H. Yang, "Deep networks for saliency detection via local estimation and global search," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3183–3192.

[12] N. Liu and J. Han, "DHSNet: Deep hierarchical saliency network for salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 678–686.

[13] J. Wei and B. Zhong, "Saliency detection using fully convolutional network," in *Proc. Chin. Autom. Congr. (CAC)*, Nov. 2018, pp. 3902–3907.

[14] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, and P. Torr, "Deeply supervised salient object detection with short connections," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3203–3212.

[15] Z. Luo, A. Mishra, A. Achkar, J. Eichel, S. Li, and P.-M. Jodoin, "Non-local deep features for salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6609–6617.

[16] N. Liu, J. Han, and M.-H. Yang, "PiCANet: Learning pixel-wise contextual attention for saliency detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3089–3098.

[17] Z. Deng *et al.*, "R³net: Recurrent residual refinement network for saliency detection," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 684–690.

[18] Z. Wu, L. Su, and Q. Huang, "Cascaded partial decoder for fast and accurate salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3907–3916.

[19] J. Zhao, J.-J. Liu, D.-P. Fan, Y. Cao, J. Yang, and M.-M. Cheng, "EGNet: Edge guidance network for salient object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8779–8788.

[20] J. Wei, S. Wang, and Q. Huang, "F³Net: Fusion, feedback and focus for salient object detection," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 12321–12328.

[21] C. Li, C. Zhu, Y. Huang, J. Tang, and L. Wang, "Cross-modal ranking with soft consistency and noisy labels for robust RGB-T tracking," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 808–823.

[22] C. L. Li, A. Lu, A. H. Zheng, Z. Tu, and J. Tang, "Multi-adapter RGBT tracking," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 2262–2270.

[23] Z. Tu, T. Xia, C. Li, Y. Lu, and J. Tang, "M3S-NIR: Multi-modal multi-scale noise-insensitive ranking for RGB-T saliency detection," in *Proc. IEEE Conf. Multimedia Inf. Process. Retr. (MIPR)*, Mar. 2019, pp. 141–146.

[24] L. Qu, S. He, J. Zhang, J. Tian, Y. Tang, and Q. Yang, "RGBD salient object detection via deep fusion," *IEEE Trans. Image Process.*, vol. 26, no. 5, pp. 2274–2285, May 2017.

[25] H. Song, Z. Liu, H. Du, G. Sun, O. L. Meur, and T. Ren, "Depth-aware salient object detection and segmentation via multiscale discriminative saliency fusion and bootstrap learning," *IEEE Trans. Image Process.*, vol. 26, no. 9, pp. 4204–4216, Sep. 2017.

[26] 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, Oct. 2019.

[27] D. Liu, Y. Hu, K. Zhang, and Z. Chen, "Two-stream refinement network for RGB-D saliency detection," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 3925–3929.

[28] H. Chen and Y. Li, "Progressively complementarity-aware fusion network for RGB-D salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3051–3060.

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

[30] Y. Piao, W. Ji, J. Li, M. Zhang, and H. Lu, "Depth-induced multi-scale recurrent attention network for saliency detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7254–7263.

[31] K. Fu, D.-P. Fan, G.-P. Ji, and Q. Zhao, "JL-DCF: Joint learning and densely-cooperative fusion framework for RGB-D salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3052–3062.

[32] Z. Zhang, Z. Lin, J. Xu, W.-D. Jin, S.-P. Lu, and D.-P. Fan, "Bilateral attention network for RGB-D salient object detection," *IEEE Trans. Image Process.*, vol. 30, pp. 1949–1961, Jan. 2021.

[33] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: https://arxiv.org/abs/1409.1556

[34] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2881–2890.

[35] S. Woo, J. Park, J.-Y. Lee, and I. So Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 3–19.

[36] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015, *arXiv:1502.03167*. [Online]. Available: https://arxiv.org/abs/1502.03167

[37] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. Int. Conf. Mach. Learn.*, 2010, pp. 1–8.

[38] 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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3917–3926.

[39] C. Godard, O. M. Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 270–279.

[40] Y. Wang, Y. Yang, Z. Yang, L. Zhao, P. Wang, and W. Xu, "Occlusion aware unsupervised learning of optical flow," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4884–4893.

[41] D.-P. Fan, C. Gong, Y. Cao, B. Ren, M.-M. Cheng, and A. Borji, "Enhanced-alignment measure for binary foreground map evaluation," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 1–7.

[42] F. Perazzi, P. Krahenbuhl, Y. Pritch, and A. Hornung, "Saliency filters: Contrast based filtering for salient region detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 733–740.

[43] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 1597–1604.

[44] R. Margolin, L. Zelnik-Manor, and A. Tal, "How to evaluate foreground maps," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 248–255.

[45] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, and A. Borji, "Structure-measure: A new way to evaluate foreground maps," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4548–4557.

[46] N. Liu, N. Zhang, and J. Han, "Learning selective self-mutual attention for RGB-D saliency detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13756–13765.

[47] T. Zhao and X. Wu, "Pyramid feature attention network for saliency detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3085–3094.

[48] X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan, and M. Jagersand, "BASNet: Boundary-aware salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7479–7489.

[49] Y. Zhang and T. Chen, "Efficient inference for fully-connected CRFs with stationarity," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 582–589.

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

[51] R. Ju, L. Ge, W. Geng, T. Ren, and G. Wu, "Depth saliency based on anisotropic center-surround difference," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2014, pp. 1115–1119.

[52] H. Peng, B. Li, W. Xiong, W. Hu, and R. Ji, "RGBD salient object detection: A benchmark and algorithms," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 92–109.

[53] Y. Cheng, H. Fu, X. Wei, J. Xiao, and X. Cao, "Depth enhanced saliency detection method," in *Proc. Int. Conf. Internet Multimedia Comput. Service (ICIMCS)*, 2014, pp. 23–27.

[54] N. Li, J. Ye, Y. Ji, H. Ling, and J. Yu, "Saliency detection on light field," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 2806–2813.

[55] G. Li and C. Zhu, "A three-pathway psychobiological framework of salient object detection using stereoscopic technology," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2017, pp. 3008–3014.

[56] D.-P. Fan, Z. Lin, Z. Zhang, M. Zhu, and M.-M. Cheng, "Rethinking RGB-D salient object detection: Models, data sets, and large-scale benchmarks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 5, pp. 2075–2089, May 2021.

[57] Y. Niu, Y. Geng, X. Li, and F. Liu, "Leveraging stereopsis for saliency analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 454–461.

[58] J. Han, H. Chen, N. Liu, C. Yan, and X. Li, "CNNs-based RGB-D saliency detection via cross-view transfer and multiview fusion," *IEEE Trans. Cybern.*, vol. 48, no. 11, pp. 3171–3183, Nov. 2018.

[59] H. Chen, Y. Li, and D. Su, "Multi-modal fusion network with multi-scale multi-path and cross-modal interactions for RGB-D salient object detection," *Pattern Recognit.*, vol. 86, pp. 376–385, Feb. 2019.

[60] N. Wang and X. Gong, "Adaptive fusion for RGB-D salient object detection," *IEEE Access*, vol. 7, pp. 55277–55284, 2019.

[61] B. Jiang, Z. Zhou, X. Wang, J. Tang, and B. Luo, "CmSalGAN: RGB-D salient object detection with cross-view generative adversarial networks," *IEEE Trans. Multimedia*, vol. 23, pp. 1343–1353, May 2020.

[62] D.-P. Fan, Z. Lin, G.-P. Ji, D. Zhang, H. Fu, and M.-M. Cheng, "Taking a deeper look at co-salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2919–2929.

[63] W.-D. Jin, J. Xu, M.-M. Cheng, Y. Zhang, and W. Guo, "ICNet: Intra-saliency correlation network for co-saliency detection," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 1–11.

[64] J. Zhang *et al.*, "Uncertainty inspired RGB-D saliency detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Apr. 15, 2021, doi: 10.1109/TPAMI.2021.3073564.

[65] D.-P. Fan, G.-P. Ji, G. Sun, M.-M. Cheng, J. Shen, and L. Shao, "Camouflaged object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2777–2787.

[66] J. Yan, T.-N. Le, K.-D. Nguyen, M.-T. Tran, T.-T. Do, and T. V. Nguyen, "MirrorNet: Bio-inspired camouflaged object segmentation," 2020, *arXiv:2007.12881*. [Online]. Available: https://arxiv.org/abs/2007.12881
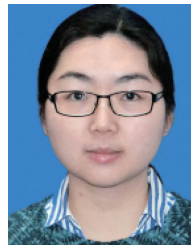
Zhengzheng Tu received the M.S. and Ph.D. degrees from the School of Computer Science and Technology, Anhui University, Hefei, China, in 2007 and 2015, respectively. She is currently an Associate Professor with the School of Computer Science and Technology, Anhui University. Her current research interests include computer vision and deep learning.

Zhun Li received the B.S. degree from Anhui University, Hefei, China, in 2019, where he is currently pursuing the M.S. degree with the School of Computer Science and Technology. His current research interests include computer vision and deep learning.
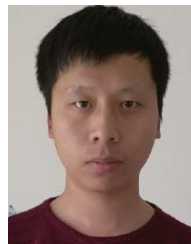
Chenglong Li received the M.S. and Ph.D. degrees from the School of Computer Science and Technology, Anhui University, Hefei, China, in 2013 and 2016, respectively. From 2014 to 2015, he worked as a Visiting Student with the School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China. He was a Postdoctoral Research Fellow with Center for Research on Intelligent Perception and Computing (CRIPAC), National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences (CASIA), China. He is currently an Associate Professor with the School of Computer Science and Technology, Anhui University. His research interests include computer vision and deep learning. He was a recipient of the ACM Hefei Doctoral Dissertation Award in 2016.

Yang Lang is currently pursuing the B.S. degree with Anhui University, Hefei, China. His current research interest includes RGBT salient object detection based on deep learning.

Jin Tang received the B.Eng. degree in automation and the Ph.D. degree in computer science from Anhui University, Hefei, China, in 1999 and 2007, respectively. He is currently a Professor with the School of Computer Science and Technology, Anhui University. His current research interests include computer vision, pattern recognition, machine learning, and deep learning.