Multimodal Salient Object Detection via Adversarial Learning with Collaborative Generator

Zhengzheng Tu[a], Wenfang Yang[a], Kunpeng Wang[a], Amir Hussain[c], Bin Luo[a], Chenglong Li[b,*]

[a]*Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, School of Computer Science and Technology, Anhui University, Hefei, 230601, China*
[b]*Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, Institute of Artificial Intelligence, Hefei Comprehensive National Science Center, School of Artificial Intelligence, Anhui University, Hefei, 230601, China*
[c]*School of Computing, Merchiston Campus, Edinburgh Napier University, Edinburgh, EH10 5DT, Scotland, U.K.*

Abstract

Multimodal salient object detection(MSOD), which utilizes multimodal information (e.g., RGB image and thermal infrared or depth image) to detect common salient objects, has received much attention recently. Different modalities reflect different appearance properties of salient objects, some of which could contribute to improving the precision and/or recall of MSOD. To greatly improve both Precision and Recall by fully exploring multimodal data, in this work, we propose an effective adversarial learning framework based on a novel collaborative generator for accurate multimodal salient object detection. In particular, the collaborative generator consists of three generators(generator1, generator2 and generator3), which aim at decreasing the false positive and false negative of the generated saliency maps and improving F-measure of the final saliency maps respectively. Generator1 and generator2 contain two encoder-decoder networks for multimodal inputs, and we propose a new co-attention model to perform adaptive interactions between different modalities. Furthermore, we apply generator3 to integrate feature maps from generator1 and generator2 in a complementary way. Through adversarially learning the collaborative generator and discriminator, both Precision and Recall of the predicted maps are boosted with the complementary benefits of multimodal data. Extensive experiments on three RGBT datasets and six RGBD datasets show that our method performs quite well against state-of-the-art MSOD methods.

Keywords:
Multimodal salient object detection, Collaborative generator, Adversarial learning.

1. Introduction

Image salient objects detection (SOD) aims to find out the most conspicuous regions in visible images, which plays an important role in many computer visual tasks, such as person re-identification (Zhao et al., 2013), object tracking (Mahadevan and Vasconcelos, 2009), image caption (Fang et al., 2015) and content-aware image editing (Zhu et al., 2014), etc. With the rapid development of deep learning in many fields, many deep learning-based SOD methods (Noori et al., 2020; Wei et al., 2020) have appeared in the past decade. However, most of SOD methods on visible images can not do very well in complex scenes, such as similar foreground and background, clutter background, bad weather and low light. Therefore, thermal infrared and depth modality are introduced to boost the performance of SOD in recent years, so that researches for multimodal salient object detection appear.

Multimodal salient object detection (MSOD) utilizes the information from multiple modalities to detect common salient objects through modality fusion, mainly including RGB-Depth (RGBD) SOD and RGB-Thermal (RGBT) SOD. Thermal infrared sensors convert invisible temperature distribution of object surface into thermal image, which

*Corresponding author
Email addresses: zhengzhengahu@163.com (Zhengzheng Tu), yangwf1102@163.com (Wenfang Yang), kp.wang@foxmail.com (Kunpeng Wang), hussain.doctor@gmail.com (Amir Hussain), luobin@ahu.edu.cn (Bin Luo), lcl1314@foxmail.com (Chenglong Li)

is insensitive to lighting condition and has a strong ability to penetrate haze and smog. Integrating visible image(that is RGB image) and thermal infrared image pairs has been proven that it is effective for several computer vision tasks (Li et al., 2016, 2019; Ye et al., 2018; Hao et al., 2019). Depth information can catch the distance between object and camera, and provide better position and boundary information of the salient object as supplementary for RGB modality. So RGBD SOD usually takes depth information as an auxiliary modality to handle some problems, such as clutter background, similar foreground and background.

For MSOD task, false positive (i.e., FP) and false negative (i.e., FN) are two common errors. FP is caused by some noises in the background, meaning that some regions in the background are mistakenly regarded as the salient foreground object. FN is caused by some pixels in the object that are not detected. For example, as shown in Fig. 1, there are FP and FN errors in the results of RGBT SOD method MIDD (Tu et al., 2021) and RGBD SOD method ICNet (Li et al., 2020). FP and FN directly affect two primary indexes that are Precision and Recall respectively. The larger FP is, the smaller Precision is. The larger FN is, the smaller Recall is. In addition, F-measure is taken to consider Precision and Recall comprehensively. Although existing methods achieve good performances in tasks of RGBT SOD (Li et al., 2018; Tu et al., 2020; Tu et al., 2021) or RGBD SOD (Li et al., 2020; Zhang et al., 2020a), they ignore considering this task from the perspective of exploring Precision and Recall and optimizing them simultaneously, resulting in bottlenecks. For traditional machine learning-based MSOD methods, improving Recall requires a low threshold for salient confidence maps, while improving Precision requires a high threshold. Previous methods (Li et al., 2018; Tu et al., 2020) use the adaptive threshold calculated from the original image, but they fail to achieve the desired balance due to complex input images. In deep learning-based MSOD methods, improving Recall requires the model to pay more attentions to global features and improving Precision requires the model to focus on local features. However, the existing deep learning-based methods (Li et al., 2020; Tu et al., 2021; Huo et al., 2022a) usually minimize the overall detection error and ignoring the difference between FN and FP, which makes the model to focus on improving either Recall or Precision, rather than suppressing them together and achieving an optimal balance. In other computer vision task, the work (Wang et al., 2019) suppresses miss detection and false alarming separately for segmenting infrared small objects accurately. However, it cannot guarantee that the two sub-tasks that are reducing miss detection and suppressing false alarming are optimal in the whole, which requires an additional model to balance the two sub-tasks.

Considering Precision and Recall are two competing evaluation metrics and boosting them at the same time may make the model tend to optimize one of Precision and Recall. Therefore, if we can boost Precision and Recall separately and achieve a balance between them, FP and FN will be correspondingly reduced. Inspired by the idea of decoupling in (Wang et al., 2019), we propose a novel adversarial learning framework with collaborative generator for accurate multimodal salient object detection. In this paper, we propose a method to directly improve the salient prediction implemented by a collaborative generator, which contains three generators (generator1, generator2 and generator3) to optimize Recall, Precision, and F-measure, respectively. Each of the three generators has its own role, and the corresponding loss function and discriminator constrain the results they generate. In this way, erroneous prediction regions can be reduced and a balance can be achieved, i.e., reducing the predictions about false positive (i.e., FP) and false negative (i.e., FN). To make the three predicted saliency maps generated by the corresponding three generators as similar as possible to the ground truth, we take advantage of a discriminator to classify these four maps. Through the adversarial learning, three generators optimize the saliency map from the perspectives of Recall, Precision and F-measure respectively, and help each other coverage towards the ground truth. As shown in Fig. 1, compared to some latest methods, our method has both fewer FN and FP predictions. The contributions of our work can be summarized as followings:

- We analyze the effects of different modalities on improving Precision and/or Recall, and propose a novel collaborative generator with three sub-generators to improve Precision, Recall and F-measure of saliency maps respectively through adversarial learning.
- We design a co-attention module to fuse features from multiple modalities. It not only enhances salient regions effectively, but also restrains redundant noises from both the channel and the spatial levels.
- We make extensive experiments and compare our method with many RGBD-based and RGBT-based methods. Our experimental results indicate that our model is effective and achieves state-of-the-art performance.

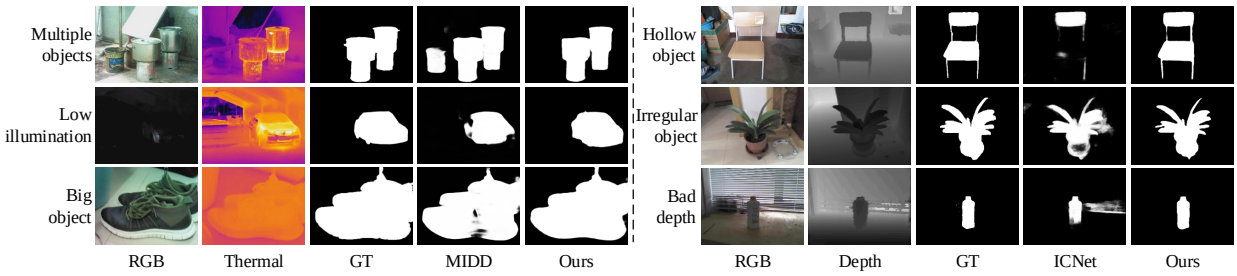

Figure 1: Comparison of MSOD methods on some challenges. **Left:** Examples of our method compared with RGBT SOD method MIDD (Tu et al., 2021). **Right:** Examples of our method compared with RGBD SOD method ICNet (Li et al., 2020).

The rest of the paper is organized as follows. Section 2 gives a brief overview of some relevant work of Multi-modal Salient Object Detection. Section 3 describes the proposed method. Section 4 shows our experimental results on several public RGBT SOD and RGBD SOD datasets. Section 5 summarizes this paper. Section 6 gives the acknowledgement.

2. Related Work

2.1. RGBT SOD Methods

With complementary information from thermal images, RGBT SOD is insensitive to illumination and has a strong ability to penetrate smog and gaze. In the early stage, some works use handcrafted features for RGBT saliency detection. Li et al. (2018) propose a novel method using multi-task manifold ranking with cross-modality consistency and construct the first RGBT image dataset. Tu et al. (2019) propose a multi-scale and multi-modal algorithm to predict the optimal ranking seed based on the manifold ranking method. Then, Tu et al. (2020) take super-pixels as graph nodes and design a novel collaborative graph learning method for RGBT SOD. Then deep learning-based RGBT SOD methods have attracted attentions and achieved great performances. Tu et al. (2022b) propose an effective baseline method for RGBT SOD, which aggregates multi-level multi-modal features with attention mechanism, and shows a great improvement against previous methods. Zhang et al. (2020b) design a network to fuse multi-modal information at various stages with several modules embedded. Zhang et al. (2020c) explore feature fusion for mining intrinsic RGBT saliency patterns and propose a novel deep feature fusion network, which consists of multi-scale, multi-modality and multi-level feature fusion modules. Tu et al. (2021) propose a siamese decoder network to integrate the cross-modal feature representation, also use global context to guide the decoding process. Zhou et al. (2022a) design an effective and consistent cross-modal fusion network to predict saliency maps. Huo et al. (2022a) propose an efficient context-guided cross modality fusion module to suppress noise and explore the complementarity of two modalities.

2.2. RGBD SOD Methods

RGBD SOD introduces depth information as an auxiliary modality to detect the common salient objects. Traditional algorithms (Zhu et al., 2017; Liang et al., 2018; Wang and Wang, 2017) extract hand-crafted features to compute saliency confident scores. But the robustness and the generalization performance of these methods are very weak. Recently, deep learning-based methods have achieved better performances. Some methods (Liu et al., 2019; Cong et al., 2019) predict saliency maps by early fusion. Some methods apply the late fusion strategy (Han et al., 2018; Wang and Gong, 2019) to fuse features of different modalities. Many existing RGBD SOD methods apply the middle fusion strategy (Chen and Li, 2018; Fu et al., 2020; Li et al., 2020; Liu et al., 2021c) to make a full fusion of different modalities. Chen and Li (2018) propose a complementarity-aware fusion network to extract and fuse the features of RGB modality and depth modality. Fu et al. (2020) employ the siamese structure with an encoder-decoder network and fuse the high-level RGB and depth features interactively, then detect salient objects adaptively. Li et al. (2020) apply an interactive and adaptive way to fuse high-level RGB and depth features. Liu et al. (2021c) leverage multi-scale features, and then design a fusion strategy based on the attention mechanism for RGBD SOD. In this

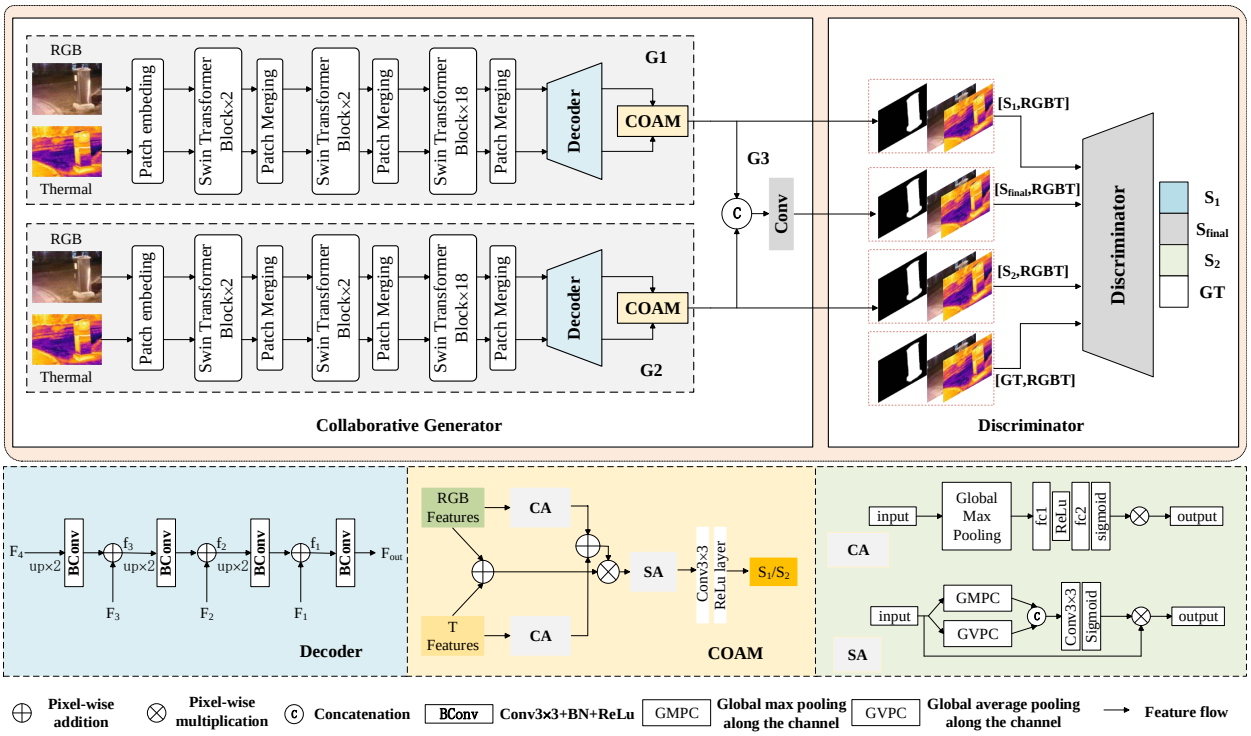

Figure 2: The overall architecture of our network, our network consists of three generators, $\mathbf{G}_1$ and $\mathbf{G}_2$ predict saliency maps $\mathbf{S}_1$ and $\mathbf{S}_2$, and $\mathbf{G}_3$ combines output features of two generators to predict the final saliency $\mathbf{S}_{final}$. The right is the discriminator network.

paper, we propose a novel collaborative generator network for multimodal SOD, and fully consider the influence of different modalities on Precision and Recall of predicted maps. We design a typical weight shared encoder-decoder network to extract features, and apply a novel co-attention module to fuse multimodal features adaptively.

2.3. Conditional GAN Methods

Because the SOD task is an image-to-image process, our method mainly utilizes a conditional generative adversarial network (CGAN) (Isola et al., 2016). In recent years, generative adversarial network (GAN) (Goodfellow et al., 2014) has made great success in many computer vision tasks, such as semantic segmentation (Dou et al., 2018; Pan et al., 2021), object detection (Chen et al., 2018), and lesion detection (Ben-Cohen et al., 2019), etc. GAN model produces a good output through the adversarial learning of two modules in the framework: generator (G) and discriminator (D). In SOD tasks (Pan et al., 2017; Jiang et al., 2020; Liu et al., 2020), D is designed to discriminate the similarity between predicted map and the ground truth. G is used to produce a binary segmentation map close to the ground truth and attempt to fool D. They work together to obtain more accurate segmentation maps. Pan et al. (2017) first use CGAN to predict saliency maps. Except for the adversarial loss, it uses BCE loss between saliency map and the ground truth. Jiang et al. (2020) learn an optimal view-invariant and consistent pixel-level representation for RGB and depth images via a novel adversarial learning framework, which also takes advantage of attention mechanism and edge detection. Liu et al. (2020) use the depth-wise separable residual convolution to deal with deep semantic information and combine processed feature with side-output features of CGAN-based encoder network. In this paper, our model is based on CGAN which is divided into three sub-networks, and these sub-networks work together competitively and cooperatively to complete their tasks. As a result, the last predicted saliency maps have a great performance.

3. Our Method

In this section, we will elaborate our multimodal salient object detection method via adversarial learning with collaborative generator (ALCG). Firstly, we will introduce the overall architecture of our network. Next, we will describe the collaborative generator and the discriminator separately. Then, the details of the co-attention module will be depicted. At last, we will introduce the loss functions we adopted. It should be explained here that we adopt RGBT image pair as the input example when elaborating the proposed network.

3.1. Overall Architecture

As illustrated in Fig. 2, the overall architecture of ALCG mainly includes two parts: a collaborative generator with three sub-generator networks and a discriminator network. Different from the classical conditional GAN, we design three generators (i.e., generator1 ($\mathbf{G}_1$), generator2 ($\mathbf{G}_2$), and generator3 ($\mathbf{G}_3$)) to complete different tasks respectively. $\mathbf{G}_1$ and $\mathbf{G}_2$ convert the paired RGBT image pair $\mathbf{I}_{(RGB,T)}$ into binary segmentation maps, which are designed to improve Recall and Precision respectively. This process can be represented as $\mathbf{G}_1(\mathbf{I}_{(RGB,T)}) \rightarrow \mathbf{S}_1$ and $\mathbf{G}_2(\mathbf{I}_{(RGB,T)}) \rightarrow \mathbf{S}_2$, where $\mathbf{S}_1$ and $\mathbf{S}_2$ mean the predicted segmentation results, $\mathbf{G}_1$ and $\mathbf{G}_2$ are two siamese generators. Besides, we design a co-attention module embedded into $\mathbf{G}_1$ and $\mathbf{G}_2$ to fuse the multi-modality features and constraint the redundant noises at the channel and the space levels. In order to improve the F-measure of the saliency maps, we design $\mathbf{G}_3$, which predicts the final saliency map $\mathbf{S}_{final}$ with the output features of $\mathbf{G}_1$ and $\mathbf{G}_2$. It can be expressed as $\mathbf{G}_3[\mathbf{S}_1, \mathbf{S}_2] \rightarrow \mathbf{S}_{final}$, where [,] indicates channel-wise concatenation. To form the complete adversarial learning, we apply a discriminator to distinguish three segmentation results (i.e., $\mathbf{S}_1$, $\mathbf{S}_2$, $\mathbf{S}_{final}$) and the ground truth (i.e., GT). The details of ALCG will be introduced in the following sections.

3.2. Collaborative Generator

Precision and Recall are two primary indexes to evaluate the performance of salient object detection methods. The calculation formulas are as follows:

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

As shown in Eq. 1 and Eq. 2, Precision consists of TP and FP, and Recall consists of TP and FN, where FP and FN determine Precision and Recall, respectively. Reducing FP and FN can improve Precision and Recall separately. Departing from the traditional way that relying on a single objective to jointly reduce both FP and FN, we decompose this difficult task into two sub-tasks handled by two generators trained adversarially and separately, each sub-task focuses on reducing either FP or FN. Therefore, we design $\mathbf{G}_1$ and $\mathbf{G}_2$, each of which focuses on increasing Recall and Precision respectively, for reducing FN and FP. For Eq. 3, F-measure is taken to consider Precision and Recall comprehensively, and it cannot be determined by one of them independently, but the F-measure can be improved by balancing Precision and Recall. $\mathbf{G}_3$ integrates the feature maps of $\mathbf{G}_1$ and $\mathbf{G}_2$ in a complementary way, optimizing the saliency map from the perspective of F-measure.

$$Fm = \frac{(1 + \beta^2) \times Precision \times Recall}{\beta^2 \times Precision + Recall} \tag{3}$$

where β^2 is set to 0.3 as suggested by (Achanta et al., 2009) to balance Precision and Recall.

As shown in Fig. 2, the inputs of $\mathbf{G}_1$ and $\mathbf{G}_2$ are an RGB image and its corresponding thermal image respectively. We use a network with shared weights to encode and decode features, the co-attention module (i.e., COAM) with channel attention and spatial attention fuse these cross-modality features from the decoder. Then, we use a sigmoid function to predict the saliency maps (i.e., $\mathbf{S}_1$ and $\mathbf{S}_2$) of $\mathbf{G}_1$ and $\mathbf{G}_2$. Note that, our two siamese generators are simple and have the same structure. Furthermore, we concatenate the output feature maps from the COAM of $\mathbf{G}_1$ and $\mathbf{G}_2$ to form the input feature of $\mathbf{G}_3$, then use two convolutional layers to predict the final saliency map (i.e., $\mathbf{S}_{final}$).

Inspired by the U-net (Ronneberger et al., 2015), our siamese generators adopt a typical encoder-decoder structure to segment the common salient objects of RGB modality and thermal modality, as shown in Fig. 2. The encoder

extracts useful features, and the decoder upsamples them to predict saliency maps. $\mathbf{G}_1$ and $\mathbf{G}_2$ have the same architecture, and we use Swin-Transformer (Liu et al., 2021b) as our encoder to extract hierarchical features with different scales from RGBT image pairs, which denoted as $\{\mathbf{F}_i^{G1}\}_{i=1}^4$ and $\{\mathbf{F}_i^{G2}\}_{i=1}^4$. Different from convolutional neural networks (CNN), Swin-Transformer has a strong feature extraction ability for hierarchical feature. In particular, self-attention is calculated in each moving window to make use of local information, which effectively reduces the calculation amount and improves efficiency. In the salient object detection task, high-level features provide global contextual information which is helpful to locate foreground objects, while low-level features contain much detailed information which is beneficial to refine boundaries of salient objects. Both of them all play important roles to segment salient objects. So, we combine the multi-level features in the decoder to aggregate global contextual and detailed information. The process can be formulated as:

$$f_i = \begin{cases} BConv(Up(\mathbf{F}_{i+1})), & i = 3 \\ BConv(Up(f_{i+1})) + \mathbf{F}_i, & i = 1, 2 \end{cases} \tag{4}$$

$$f_{out} = BConv(f_1)$$

where f_i is the output feature from the upper layer, $\mathbf{F}_i$ is extracted feature in the same layer. *BConv* is a 3×3 convolutional block followed by a batch normalization layer and a ReLU activation function. Note that, simple addition operation does not involve parameters in the network, but the calculation progress is efficient. Furthermore, the pixel-wise addition can find the common objects among the input features, it is helpful to aggregate these features.

3.3. Adversarial Learning Discriminator

In our work, we adopt the adversarial learning to encourage the generators to cooperate with each other to achieve excellent performance for multi-modal SOD. Before the discriminator ($\mathbf{D}$), $\mathbf{G}_1$, $\mathbf{G}_2$, and $\mathbf{G}_3$ are designed to improve Recall, Precision, and F-measure, respectively. So if there is no interaction between them, the network will lose integrity. Therefore, $\mathbf{D}$ is proposed to realize information exchange between the generators in the collaborative generator and make the saliency maps generated by them close to the ground truth. In this way, $\mathbf{G}_1$, $\mathbf{G}_2$ and $\mathbf{G}_3$ can help the other two generators improve performance when boosting their own index (i.e., Recall, Precision, and F-measure). Specifically, $\mathbf{S}_1$, $\mathbf{S}_2$, and $\mathbf{S}_{final}$ are predicted from $\mathbf{G}_1$, $\mathbf{G}_2$ and $\mathbf{G}_3$ together with the ground truth, then we feed these saliency maps into the discriminator to let the discriminator distinguish the predictions and the ground truth. With training iterations, the three generators are able to produce better saliency maps in a cooperative manner so that the discriminator has difficulty in distinguishing them from the ground truth.

The concrete architecture of $\mathbf{D}$ is expressed in Table 1. We just design three pairs of convolutional blocks, three max-pooling layers, and three fully-connected layers in $\mathbf{D}$. The fully-connected layers are activated by the tanh function, the convolution layers are activated by the ReLU function and the result is activated by the sigmoid function. The inputs of $\mathbf{D}$ are four segmentation maps together with the original inputs, $[\mathbf{S}_1, \mathbf{I}_{input}]$, $[\mathbf{S}_2, \mathbf{I}_{input}]$, $[\mathbf{S}_{final}, \mathbf{I}_{input}]$ and $[\mathbf{Y}, \mathbf{I}_{\mathbf{input}}]$, where $\mathbf{S}_1$, $\mathbf{S}_2$ and $\mathbf{S}_{final}$ are the predicted maps of three generators, $\mathbf{Y}$ is the ground truth, $\mathbf{I}_{input}$ is the average value of the original RGB image and thermal image. In detail, we concatenate each saliency map and their average original inputs to form 4-channel feature maps with the size of 384*384*4. The output of $\mathbf{D}$ is four sets of confidence scores, each of which represents the probability that the corresponding saliency map belongs to the three generators and the ground truth. By adversarial learning, three generators cooperate with each other to produce accurate saliency maps. Meanwhile, the generators try their best to predict saliency maps close to the ground truth, making it difficult for the discriminator to distinguish the prediction from the ground truth.

3.4. Co-attention Module

As a cross-modal task, noise is inevitably introduced in the fusion of RGB modality and thermal infrared modality. However, many previous methods fuse cross-modal features without considering filtering noises. Hence, we propose a co-attention module (COAM) embedded into $\mathbf{G}_1$ and $\mathbf{G}_2$ to progressively constrain the noises and highlight common salient objects in different modalities, which makes our network concentrate on representative features. The co-attention module is based on channel attention and spatial attention, which can suppress the noise in features from the channel and spatial dimensions, respectively. The details of the COAM are shown in Fig. 2, we first explore the role of each channel feature for a single modality, and then further mine the channel relationships across modalities. After

Table 1: The details of Discriminator

Layer	kernel	activation	out-channels
Conv1	3×3	ReLU	32
Conv2	3×3	ReLU	32
Max-pooling	2×2	−−	32
Conv3	3×3	ReLU	64
Conv4	3×3	ReLU	64
Max-pooling	2×2	−−	64
Conv5	3×3	ReLU	64
Conv6	3×3	ReLU	64
Max-pooling	2×2	−−	64
Fc7	−−	Tanh	128
Fc8	−−	Tanh	64
Fc9	−−	sigmoid	4

that, we further explore the spatial relations in the fused cross-modal features. Concretely, given the multi-modal input features (f_{RGB}, f_T), in which f_{RGB} and f_T represent RGB and thermal features respectively in the co-attention module of Fig. 2, we first perform channel attention (Hu et al., 2018) on each set of features, then the channel attention map of each modality is calculated. It can be expressed as:

$$CA_k = \sigma(M(f_k(i, j))) \tag{5}$$

where σ is the sigmoid function, M is the global max-pooling operation in spatial dimensions for all positions, $k \in \{RGB, T\}$ and (i, j) is the position of pixels. In order to explore the channel-wise relevance of the cross-modal features, we add the two channel attention maps in pixel-wise and calculate the common channel attention results.

$$f_{fuse} = (f_{RGB} + f_T) \odot (CA_{RGB} + CA_T) \tag{6}$$

where $\odot$ is pixel-wise multiplication. We also fuse RGB modality and thermal modality via this operation. We apply this cross-modality channel attention operation to find the common important channels between RGB features and the thermal features and then fuse them.After that, we continue to explore the informative spatial feature representations in the fusion feature f_{fuse}. We feed the fusion feature into spatial attention mechanism (Fei et al., 2017) and estimate its pixel-wise confidence map $SA_{fuse} = \sigma(Conv([M(f_{fuse}(i, j)), A(f_{fuse}(i, j))]))$, A is the global average-pooling operation. After that, we will gain the spatial attention results, represented as:

$$f_{out} = \phi(Conv(f_{fuse} \odot SA_{fuse})) \tag{7}$$

where $Conv$ is the 3×3 convolution operation, ϕ is the ReLU activation function. By spatial attention mechanism, we further highlight the salient objects in spatial level. So far, such cross-modality co-attention mechanism can model the channel-wise and space-wise relevance of multi-modal features and adaptively select informative channel-spatial features. Therefore, each generator can fuse the representative multi-modal features by the co-attention module and help our ALCG achieve quite good performances.

3.5. Loss Function

Our model is optimized by a combination of three loss functions: the adversarial loss $L_a(\mathbf{G}, \mathbf{D})$, the subgenerator loss $L_{sg}(\mathbf{S}_i, \mathbf{Y})$ and the consistency similarity loss $L_{cs}(\mathbf{S}_1, \mathbf{S}_2)$. The whole loss function G^* can be expressed as follows:

$$G^* = L_a(\mathbf{G}, \mathbf{D}) + L_{sg}(\mathbf{S}_i, \mathbf{Y}) + L_{cs}(\mathbf{S}_1, \mathbf{S}_2) \tag{8}$$

where G and D are generator and discriminator respectively. $\mathbf{S}_i \in \{\mathbf{S}_1, \mathbf{S}_2, \mathbf{S}_{final}\}$ represents the saliency maps generated by the corresponding generators (i.e., $\mathbf{G}_1$, $\mathbf{G}_2$ and $\mathbf{G}_3$). $\mathbf{Y}$ is the ground truth. We will describe these loss functions in detail in the following sections.

3.5.1. The Adversarial Loss

Since conditional GAN (Isola et al., 2016) is utilized to build our model, an adversarial loss $L_a(\mathbf{G}, \mathbf{D})$ is introduced to optimize the generator and the discriminator in the training phase. It promotes the collaborative generator to segment more refined results and improves the discrimination ability of the discriminator, so that the predicted saliency map is close to the ground truth. This process can be expressed by:

$$L_a(\mathbf{G}, \mathbf{D}) = {}_{X_{input},Y}[log\mathbf{D}(X_{input}, Y)] + \sum_{i=1}^{3} {}_{X_{input}}[log(1 - \mathbf{D}(X_{input}, \mathbf{G}_i(X_{input})))] \tag{9}$$

where X_{input} is the mean of original input RGB image and thermal image.

3.5.2. Subgenerator Loss

$\mathbf{G}_1$, $\mathbf{G}_2$ and $\mathbf{G}_3$ are designed to improve Recall, Precision and F-measure, respectively, while generating their own saliency maps $S_i \in \{S_1, S_2, S_{final}\}$. According to the Eq. 2, Recall can be improved by decreasing the false negative (i.e., FN), which is calculated from salient pixels misclassified as background. Gavin the saliency map S, FN is calculate as $FN = \|(S - Y) \odot Y\|_2^2$. Eq. 1 indicates that Precision can be improved by decreasing FP, which is decided by the background pixels segmented as foreground. FP is formulated as $FP = \|(S - Y) \odot (1 - Y)\|_2^2$. Therefore, for $\mathbf{G}_1$ and $\mathbf{G}_2$, we define the following loss functions to boost Recall and Precision separately:

$$\begin{aligned} L_{sg}(S_1, Y) &= \frac{1}{n} \textstyle\sum_{i=1}^{n} \lambda_1 FN_i + FP_i \\ L_{sg}(S_2, Y) &= \frac{1}{n} \textstyle\sum_{i=1}^{n} FN_i + \lambda_2 FP_i \end{aligned} \tag{10}$$

where n is the number of input samples, i represents i-th image in the training set. λ_1 and λ_2 are the weighting parameters to balance FN and FP , which promotes $\mathbf{G}_1$ to focus on optimizing FN and $\mathbf{G}_2$ to mainly optimize FP. In our model, λ_1 and λ_2 are set as 10 and 1, respectively, and the corresponding ablation experiments will be presented in Sec. 4.5.

If we take the combination of S_1 and S_2 as the final salient map directly, it would be difficult for the network to take advantage of both $\mathbf{G}_1$ and $\mathbf{G}_2$ due to their differences, which will lead to inaccurate classification between foreground and background. So $\mathbf{G}_3$ is designed to integrate the outputs of $\mathbf{G}_1$ and $\mathbf{G}_2$ and predict the final saliency map S_{final}, which is dedicated to improve F-measure. Therefore, we compute FLoss (Zhao et al., 2019) between S_{final} and the ground truth Y to achieve the improvement. In particular, given the final predicted saliency maps $S_{final} = \{S_{final_i} | i = 1, ..., N\}$ and the ground truth $Y = \{Y_i | i = 1, ..., N\}$, where N is the number of total pixels. The process can be expressed as following:

$$L_{sg}(S_{final}, Y) = 1 - \frac{(1 + \beta^2) \times TP}{\beta^2 \times (TP + FN) + (TP + FP)} \tag{11}$$

where $TP = \sum_{i=1}^{N} S_{final_i} \cdot Y_i$, $FP = \sum_{i=1}^{N} S_{final_i} \cdot (1 - Y_i)$, and $FN = \sum_{i=1}^{N} (1 - S_{final_i}) \cdot Y_i$. Note that the calculation of FP and FN is the same as (Zhao et al., 2019), which is different from the Eq. 10. Since $\mathbf{G}_3$ takes advantage of the output features of $\mathbf{G}_1$ and $\mathbf{G}_2$, Recall and Precision can be integrated and further improved from a global perspective by the FLoss on S_{final}. As shown in Eq. 3, F-measure is calculated on the basis of Recall and Precision, so it also can be improved since both Recall and Precision achieve a delicate balance. Therefore, by applying FLoss to $\mathbf{S}_1$, $\mathbf{S}_2$ and $\mathbf{S}_{final}$, Recall and Precision can be improved from local and global perspectives, respectively, and F-measure is thereby also improved. Therefore, $\mathbf{G}_1$, $\mathbf{G}_2$ and $\mathbf{G}_3$ can be optimized in a cooperative manner.

3.5.3. Consistency Similarity Loss

In order to further form a cooperative generator, we introduce a consistency similarity loss function to make $\mathbf{G}_1$ and $\mathbf{G}_2$ cooperate with each other through the mutual constraint of $\mathbf{S}_1$ and $\mathbf{S}_2$. Although $\mathbf{G}_1$ and $\mathbf{G}_2$ perform different tasks, their segmentation results should be as similar as possible to the ground truth. Therefore, predicted saliency maps $\mathbf{S}_1$ and $\mathbf{S}_2$ should also be similar with each other in theory. Though the above loss functions force the saliency

maps $\mathbf{S}_1$ and $\mathbf{S}_2$ to be close to the ground truth and alleviate their differences to some extent, there is still a large gap between them. So we introduce the consistency similar loss to narrow the gap between $\mathbf{S}_1$ and $\mathbf{S}_2$, as follows:

$$L_{cs}(\mathbf{S}_1, \mathbf{S}_2) = \frac{1}{w \times h \times c} \|\mathbf{S}_1 - \mathbf{S}_2\|_2^2 \tag{12}$$

where w, h and c are width, height and channels of the feature maps. In this way, the model bias towards Recall or Precision can be eliminated by further eliminating the differences between $\mathbf{S}_1$ and $\mathbf{S}_2$. In addition, $\mathbf{S}_3$, which specializes in F-measure, can also be improved.

4. Experiments

In this section, we elaborate on the details of our experiments. We first introduce three RGBT datasets we adopted in Sec. 4.1. The experiment setup and the evaluation criteria for our method are described in Sec. 4.2 and Sec. 4.3 separately. Then, we conduct the comparison experiments and ablation study on RGBT datasets to demonstrate the effectiveness of our method in Sec. 4.4 and Sec. 4.5, respectively. In addition, to make a comparison with the latest RGBD SOD methods, more experiments are conducted on six RGBD datasets in Sec. 4.6.

4.1. RGBT Datasets

We utilize three publicly available RGBT SOD datasets, namely VT821 (Li et al., 2018), VT1000 (Tu et al., 2020) and VT5000 (Tu et al., 2022b) for the following RGBT SOD experiments. These datasets consist of 821, 1000 and 5000 aligned visible and thermal infrared image pairs and corresponding binary segmentation masks respectively. VT821 is the first RGBT SOD dataset. It consists of 821 RGB and thermal infrared image pairs, which are added artificial noises to make the dataset more challenging. VT1000 contains 1000 paired images with relatively simple scenes. VT5000 contains 5000 image pairs, including various objects and complex scenes. Among them, 2500 image pairs are divided into training set and the remaining 2500 are used as testing set. All of the datasets have 11 challenges to test the effectiveness of different methods, as follows: bad weather (BW), big salient object (BSO), similar appearance (SA), small salient object (SSO), multiple salient object (MSO), cross image boundary (CIB), low illumination (LI), center bias (CB), and out of focus (OF), thermal crossover (TC) and image clutter (IC). Similar with MIDD (Tu et al., 2021) and ADF (Tu et al., 2022b), we utilize 2500 paired RGBT images in the training set of VT5000 to train our model, and utilize VT821, VT1000 and the testing set of VT5000 to evaluate the performance of methods.

4.2. Experiment Setup

We implement our network with Pytorch and train it by a single Titan XP GPU. For the inputs, all image pairs are resized to 384×384. We set λ_1=10 , λ_2=1 and β^2 =0.3, and test different values of them in the part of ablation study. During training, we use the adaptive moment estimation (Adam) (Kingma and Ba, 2014) to optimize the generators and the discriminator with batch size of 2 for 100 epochs. The initial learning rates of the generators and the discriminator are set to 1e-4 for and 1e-5, which are multiplied by 0.2 every 25 epochs.

4.3. Evaluation Criteria

We use F-measure (Achanta et al., 2009), S-measure (Fan et al., 2017), E-measure (Fan et al., 2018), mean absolute error (MAE) (Perazzi et al., 2012), wfm (Margolin et al., 2014) and Precision-Recall (PR) curve to evaluate the performance of our method and other methods from different perspectives. The above evaluation criteria are widely used in multimodal SOD. Among them, PR curve is an index for comprehensive evaluation of Recall and Precision. In detail, the saliency maps are binarized using a threshold, and then the Precision and Recall are calculated using the ground truth. By setting different thresholds and calculating multiple sets of Precision and Recall values, the PR curve is finally obtained. F-measure (Fm) is a weighted harmonic metric of both Precision and Recall, which is not dominated by a single one of them. Therefore, we utilize Fm as our primary evaluation metric to demonstrate the effectiveness of our method.

Table 2: Performance comparison with 15 methods on three RGBT datasets. The best scores are highlighted in **BOLD**.

Methods	VT821					VT1000					VT5000				
	Em	Sm	Fm	MAE	wfm	Em	Sm	Fm	MAE	wfm	Em	Sm	Fm	MAE	wfm
DMRA (Piao et al., 2019)	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 (Liu et al., 2020)	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
A2dele (Piao et al., 2020)	0.651	0.617	0.569	0.061	0.505	0.796	0.759	0.758	0.057	0.696	0.746	0.689	0.662	0.059	0.587
BBSNet (Fan et al., 2020b)	0.876	0.868	0.768	0.045	0.741	0.915	0.923	0.855	0.027	0.845	0.896	0.882	0.786	0.040	0.770
MTMR (Li et al., 2018)	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 (Tu et al., 2019)	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 (Tu et al., 2020)	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 (Tu et al., 2022b)	0.842	0.810	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
MIDD (Tu et al., 2021)	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
APNet (Zhou et al., 2022b)	0.907	0.867	0.816	0.034	0.792	0.938	0.921	0.883	0.021	0.883	0.914	0.875	0.820	0.035	0.806
ECFFNet (Zhou et al., 2022a)	0.902	0.877	0.810	0.034	0.801	0.930	0.923	0.876	0.021	0.885	0.906	0.874	0.807	0.038	0.807
CSRNet (Huo et al., 2022a)	0.909	0.885	0.845	0.038	0.821	0.925	0.918	0.877	0.024	0.878	0.905	0.868	0.811	0.042	0.796
CGFNet (Wang et al., 2022)	0.912	0.881	0.845	0.038	0.829	0.944	0.923	0.906	0.023	0.900	0.922	0.883	0.851	0.035	0.831
MIA_DPD (Liang et al., 2022)	0.850	0.844	0.741	0.070	0.720	0.926	0.924	0.868	0.025	0.864	0.893	0.879	0.793	0.040	0.780
OSRNet (Huo et al., 2022b)	0.896	0.875	0.814	0.043	0.801	0.935	0.926	0.892	0.022	0.891	0.908	0.875	0.832	0.040	0.807
DCNet (Tu et al., 2022a)	0.913	0.877	0.841	0.033	0.822	0.949	0.923	0.911	0.021	0.902	0.921	0.872	0.847	0.035	0.819
CCFENet (Liao et al., 2022)	0.925	**0.900**	0.857	**0.027**	**0.852**	0.946	0.934	0.906	0.018	0.910	0.932	0.896	0.859	0.030	0.849
Ours	**0.928**	0.888	**0.867**	**0.027**	0.845	**0.957**	**0.936**	**0.926**	**0.015**	**0.922**	**0.945**	**0.899**	**0.886**	**0.027**	**0.865**

MAE is designed to evaluate the discrepancy between the predicted saliency maps and the ground truth:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|S_i - Y_i| \tag{13}$$

where N is the total number of pixels in the feature map, S_i is the predicted segmentation map and Y_i is the ground truth.

S-measure (Sm) is proposed to calculate the similarity between the predicted map and the ground truth:

$$S_m = \alpha \times S_o + (1-\alpha) \times S_r \tag{14}$$

where S_o is the object-aware structural similarity, S_r is the region-aware structural similarity and α is set to 0.5, as suggested in (Fan et al., 2017).

E-measure (Em) is an enhanced-alignment measure which jointly captures the image-level mean values and pixel-wise values. We use above metrics to make a comprehensive evaluation of our method, demonstrating our method is effective and reasonable.

4.4. Comparisons with the State-of-the-art methods

We compare our method with 17 existing methods, including four deep learning RGBD SOD methods (i.e., S2MA (Liu et al., 2020), BBSnet (Fan et al., 2020b), A2dele (Piao et al., 2020) and DMRA (Piao et al., 2019)), and three traditional RGBT SOD methods (i.e., MTMR (Li et al., 2018), M3S-NIR (Tu et al., 2019) and SGDL (Tu et al., 2020)), and ten deep learning based RGBT SOD methods (i.e., ADF (Tu et al., 2022b), MIDD (Tu et al., 2021), APNet Zhou et al. (2022b), ECFFNet (Zhou et al., 2022a), CSRNet (Huo et al., 2022a), CGFNet (Wang et al., 2022), MIA_DPD (Liang et al., 2022), OSRNet (Huo et al., 2022b), DCNet (Tu et al., 2022a), CCFENet (Liao et al., 2022)). Different from deep learning based RGBD and RGBT SOD methods, our proposed method separates the model into three subnetworks which aim to improve Precision, Recall and Fm score of the saliency maps respectively. Furthermore, we use a simpler fusion method than these comparative methods. Besides, our generators are end-to-end framework without any post-processing.

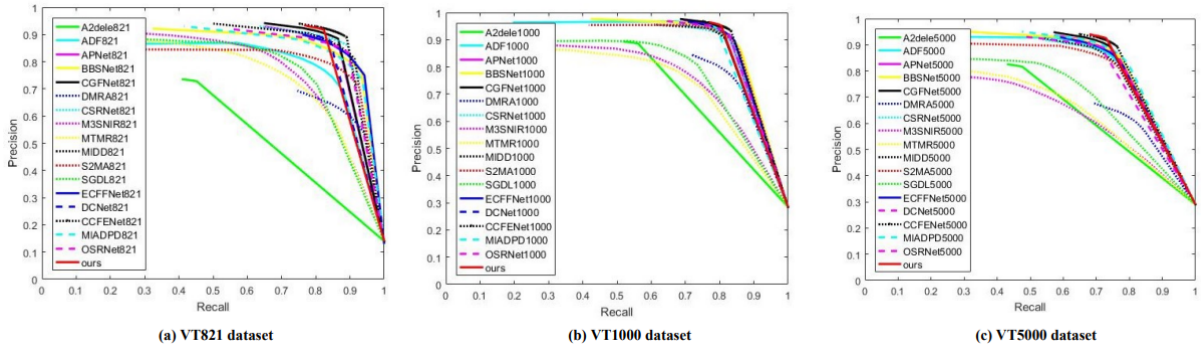

Figure 3: PR curves comparison with other 17 methods. From left to right, PR curve in VT821, VT1000, and VT5000 testing set. We use different color represent different method.

4.4.1. Quantitative Evaluation

First of all, quantitative results are shown in Table 2. The comparisons between predicted maps of our method and the other 15 methods on three RGBT testing sets are recorded clearly. We calculate 5 evaluation metrics which have been already introduced, and our method outperforms all of the other methods,including RGBT SOD methods and RGBD SOD methods, on all RGBT datasets. Certainly, because of the shortness of feature representations in the manual features, three traditional RGBT SOD methods are inferior to the deep learning-based methods.

Secondly, to compare with RGBD SOD methods, we retrain the methods on RGBT datasets and gain saliency maps. The compared RGBD SOD methods are existing state-of-the-art. DMRA (Piao et al., 2019) uses a middle fusion strategy to fuse the cross-modality features, and it is effective in dealing with RGBD SOD problems. A2dele (Piao et al., 2020) designs a lightweight model to resolve the RGBD SOD challenge and also has a good performance. BBSNet (Fan et al., 2020b) uses a depth-enhanced method and designs the high-level features as a teacher network to guide the low-level features. S2MA (Liu et al., 2020) is the latest method for RGBD SOD. In Table 2, because BBSNet and S2MA all design complex fusion strategies, we can find that they perform well on RGBT datasets. To sum up, RGBD SOD has developed more maturely than RGBT SOD, we can graft the saliency detection method of RGBD SOD to RGBT SOD tasks, but it may be not decent. Like analysis in MIDD (Tu et al., 2021), depth is usually used as an auxiliary modality to enhance RGB modality in RGBD SOD tasks, while RGBT SOD tasks jointly use thermal maps to infer the saliency maps, RGB image and thermal image play equivalent important role to segment foreground objects.

Finally, we compare our method with thirteen RGBT SOD methods. Because traditional methods have limitation in generalization and robustness, we mainly compare our results with the deep learning methods with complex topological structure. Fm is the weighted harmonic metric of precision and recall, so we choose it as our primary metrics method. As shown in Table 2, we improve Fm 1% on VT821 dataset, improve Fm 2%, wfm 1.2% on VT1000 dataset, improve Fm 2.7%, wfm 1.6% on VT5000 dataset than sub-optimal method (i.e., CCFENet). Besides, Compared with the sub-optimal method, our method achieves an average improvement of 8.8% with the MAE metric on the three datasets. It demonstrates that our method can balance the Precision and the Recall. Besides, our method decreases the MAE on three datasets obviously. As we can observe from Table 2, the deep learning-based methods outperform the traditional methods, which is mainly attributed to the design of different fusion modules. However, these methods fall into a performance bottleneck because none of them can optimize Precision and Recall simultaneously, which directly determines the accuracy of saliency prediction. In contrast, our method optimizes Precision and Recall separately and enables them to reach a trade-off, while outperforming these methods. In addition, the performance of the RGBD SOD method decays on the RGBT dataset, while our method achieves impressive performance on both the RGBD and the RGBT datasets, which demonstrates the robustness of our method.

4.4.2. Qualitative Evaluation

The visual comparison is shown in Fig. 4. It provides predicted saliency maps from our method and other eight RGBT methods. Our visual comparisons are conducted for simple scenes and complex scenes. It is easy to find

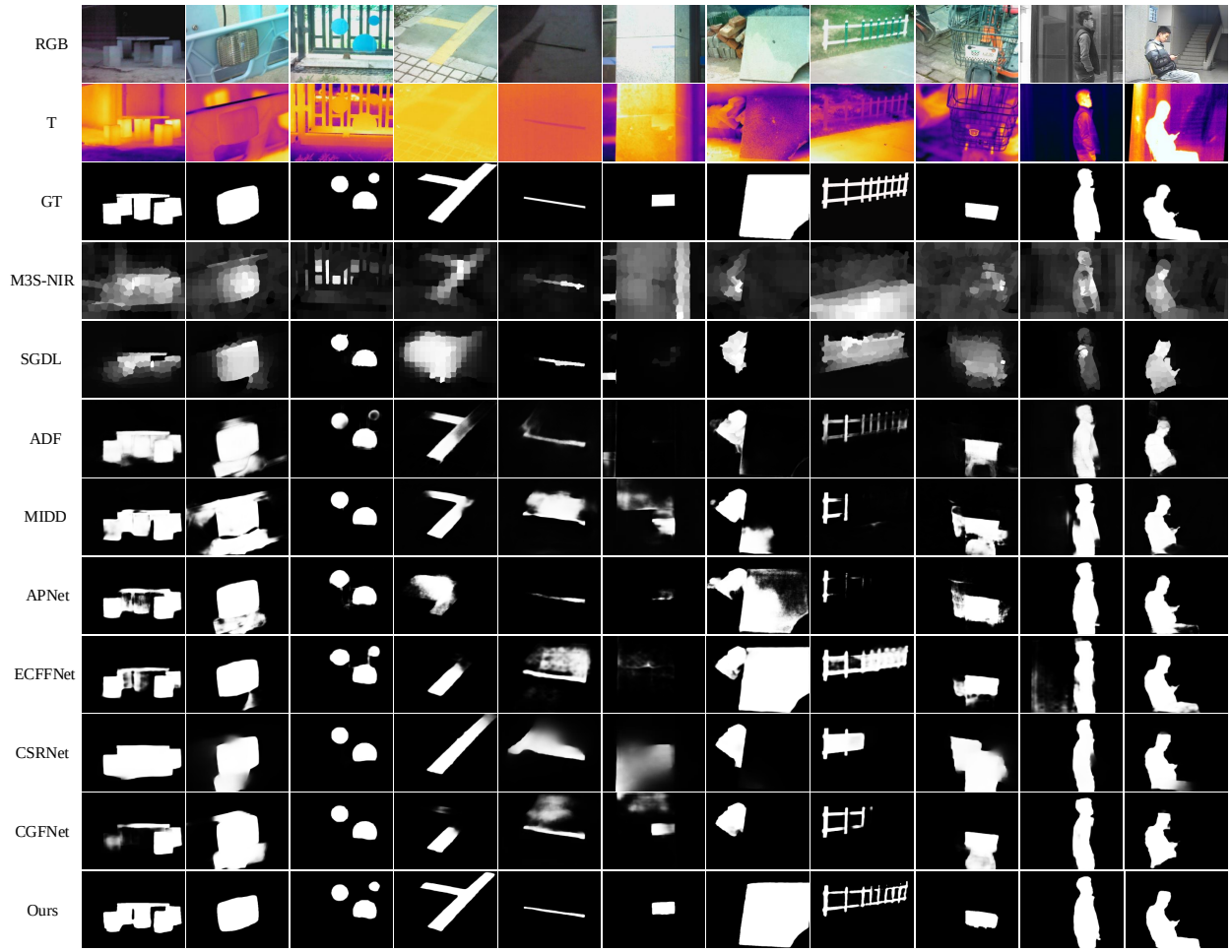

Figure 4: Qualitative comparison of our method with other 8 RGBT SOD methods. We select 11 RGBT image pairs with diverse challenges to compare the quality of the predicted maps.

that RGBT methods achieve good performances for low lightness, shadows, and complex scene challenges. When one of the modalities has poor quality, the other modality plays a more important role in providing representative information. Certainly, our method is trained to take advantage of informative features from two modalities as much as possible. Although salient objects have twisted shape such as 8th columns in Fig. 4, our method can explore the useful information which is helpful to segment objects. Compared with other methods, our method segments salient regions more accurately in scenes with complex challenges, such as multiple targets (i.e., columns 1, 3), heat map failure (i.e., column 4), and background clutter (i.e., columns 3, 9). This is mainly due to our proposed collaborative generator, which optimizes FP and FN separately and makes them get a balance, reducing prediction errors and thus directly improving detection performance. Besides, as shown in Fig. 3, we compare PR curve with other 17 methods on three RGBT datasets. It is easy to find that our method has lower but reasonable Recall values, but has achieved competitive Precision values on all datasets. Combined with the quantitative comparison in Table 2, our method achieves the highest Fm on all datasets, which demonstrates our method makes the best balance between Precision and Recall at the pixel level.

4.4.3. Challenge-based Quantitative Evaluation

We make a concrete comparison with other 17 methods(i.e., 4 RGBD SOD methods and 13 RGBT SOD methods) on 11 challenges attributes and 2 modality quality annotations(bad RGB modality and bad thermal infrared modality),

Table 3: Performance comparison of 15 methods on 11 cha (Wang et al., 2022)llenges and 2 modality quality annotations. The best scores are highlighted in **BOLD**.

Methods	*BSO*	*CB*	*CIB*	*IC*	*LI*	*MSO*	*OF*	*SSO*	*SA*	*TC*	*BW*	*bRGB*	*bT*
DMRA (Piao et al., 2019)	0.688	0.538	0.626	0.526	0.585	0.548	0.580	0.259	0.522	0.474	0.498	0.551	0.461
S2MA (Liu et al., 2020)	0.844	0.781	0.831	0.758	0.830	0.772	0.808	0.680	0.777	0.769	0.741	0.784	0.767
A2dele (Piao et al., 2020)	0.756	0.701	0.730	0.706	0.626	0.678	0.684	0.548	0.659	0.708	0.615	0.626	0.707
BBSNet (Fan et al., 2020b)	0.870	0.829	0.857	0.800	0.841	0.812	0.827	0.741	0.824	0.800	0.776	0.812	0.794
MTMR (Li et al., 2018)	0.489	0.470	0.421	0.449	0.547	0.495	0.571	0.634	0.538	0.463	0.492	0.531	0.454
M3S-NIR (Tu et al., 2019)	0.499	0.464	0.450	0.443	0.556	0.469	0.567	0.508	0.504	0.441	0.512	0.506	0.431
SGDL (Tu et al., 2020)	0.676	0.658	0.622	0.625	0.659	0.660	0.685	0.710	0.592	0.621	0.588	0.597	0.607
ADF (Tu et al., 2022b)	0.852	0.816	0.833	0.791	0.834	0.803	0.805	0.743	0.798	0.797	0.775	0.786	0.789
MIDD (Tu et al., 2021)	0.872	0.838	0.851	0.798	0.853	0.815	0.828	0.743	0.815	0.809	0.772	0.812	0.796
APNet (Zhou et al., 2022b)	0.883	0.843	0.868	0.813	0.871	0.825	0.845	0.753	0.846	0.821	0.809	0.8358	0.806
ECFFNet (Zhou et al., 2022a)	0.875	0.831	0.862	0.802	0.848	0.823	0.821	0.723	0.809	0.808	0.756	0.810	0.798
CSRNet (Huo et al., 2022a)	0.867	0.812	0.834	0.790	0.855	0.815	0.837	0.759	0.799	0.801	0.760	0.824	0.792
CGFNet (Wang et al., 2022)	0.888	0.868	0.873	0.841	0.878	0.844	0.861	0.804	0.858	0.854	0.817	0.844	0.846
MIA_DPD (Liang et al., 2022)	0.877	0.828	0.870	0.797	0.851	0.811	0.832	0.710	0.823	0.804	0.781	0.807	0.795
OSRNet (Huo et al., 2022b)	0.880	0.842	0.858	0.811	0.855	0.832	0.841	0.753	0.814	0.820	0.813	0.827	0.809
DCNet (Tu et al., 2022a)	0.887	0.853	0.865	0.821	0.869	0.835	0.850	0.767	0.846	0.839	0.816	0.833	0.834
CCFENet (Liao et al., 2022)	0.896	0.870	0.888	0.844	0.882	0.853	0.867	0.832	0.870	0.867	0.823	0.854	0.863
Ours	**0.913**	**0.885**	**0.898**	**0.866**	**0.892**	**0.872**	**0.892**	**0.835**	**0.872**	**0.879**	**0.859**	**0.873**	**0.880**

which are provided by the VT5000 test set. We compute the mean F-score as the metric index which can mainly present the performance on challenges. As shown in Table 3, our method achieves top performance than other SOD methods, showing the advantage in balancing Precision and Recall. Furthermore, our method has a strong ability to enhance salient objects and suppress background noises. We achieve 1.7%, 1.5%, 1%, 2.2%, 1%, 1.9% , 2.5%, 0.3%, 0.2%, 1.2%, 3.6%, 1.9% and 1.7% gains on these challenges compared with the sub-optimal results. On the other hand, SSO, IC, and BW, are considered as the most difficult challenges, all of the comparison methods have lower scores on these challenges. But our method also achieves great performance on these challenges. To resolve the SSO problem, we need to design a particular method in the future.

4.4.4. Precision-Recall Quantitative Evaluation

We quantify the values of Precision and Recall of our method and previous RGBT SOD methods on VT821 and VT1000 in Table 4. The experimental results show that the values of the Precision of our method have reached the optimal. Our method outperforms the suboptimal method (i.e., CSRNet) by 2.8% and 4.8% on the two datasets, respectively. It is easy to find that our method gains the highest Fm on both datasets with the highest Precision and rational Recall. It demonstrates our primary idea of balancing Precision and Recall in the predicted maps is achieved. However, in contrast to our method, other methods may not achieve this ideal situation. For example, on VT821, ECFFNet (Zhou et al., 2022a) and MIDD (Tu et al., 2021) have achieved the highest Recall but lower Precision, which illustrates that these methods detect some background noises as the salient area. And on VT1000, MTMR (Li et al., 2018) and SGDL (Tu et al., 2020) have achieved high Precision but the lowest Recall, illustrating that these methods predict foreground objects inaccurately. All of these examples also explain why their Fm values are lower than our method and indicate that balance between Precision and Recall is important for the MSOD task.

4.5. Ablation Study

We design an ablation study to demonstrate the effect of each component in our network. Concretely, we will testify the effectiveness of the co-attention module, the adversarial learning, and the consistency similarity loss sequentially. The details of the ablation study are expressed in Table 5.

Table 4: Comparison of our method against other RGBT SOD methods, *Pre* and *Rec* represent Precision and Recall respectively. The best scores are highlighted in **BOLD**.

Methods	VT821			VT1000		
	Pre	*Rec*	*Fm*	*Pre*	*Rec*	*Fm*
MTMR (Li et al., 2018)	0.716	0.713	0.662	0.809	0.610	0.715
SGDL (Tu et al., 2020)	0.794	0.724	0.730	0.854	0.650	0.764
ADF (Tu et al., 2022b)	0.767	0.811	0.716	0.900	0.811	0.847
MIDD (Tu et al., 2021)	0.841	0.877	0.804	0.910	0.809	0.882
ECFFNet (Zhou et al., 2022a)	0.847	**0.879**	0.810	0.908	**0.832**	0.876
CSRNet (Huo et al., 2022a)	0.894	0.870	0.845	0.912	0.816	0.877
Ours	**0.919**	0.818	**0.867**	**0.956**	0.800	**0.926**

Table 5: Details of ablation study, w/o means disable the corresponding component.

Variant	VT821				VT1000				VT5000			
	Fm	*MAE*	*wfm*	*Sm*	*Fm*	*MAE*	*wfm*	*Sm*	*Fm*	*MAE*	*wfm*	*Sm*
Single generator	0.807	0.043	0.749	0.872	0.873	0.024	0.865	0.930	0.825	0.036	0.790	0.891
Double generators	0.814	0.041	0.756	0.881	0.884	0.021	0.871	0.935	0.833	0.034	0.796	0.895
w/o L_{cs}	0.856	0.030	0.835	0.884	0.920	0.017	0.915	0.933	0.879	0.029	0.858	0.895
w/o COAM	0.853	0.029	0.825	0.871	0.925	0.016	0.920	0.935	0.880	0.029	0.855	0.892
w/o GAN	0.854	0.032	0.840	0.887	0.914	0.017	0.911	0.932	0.865	0.032	0.846	0.889
Ours	**0.867**	**0.027**	**0.845**	**0.888**	**0.926**	**0.015**	**0.922**	**0.936**	**0.886**	**0.027**	**0.865**	**0.899**

Firstly, 'Single generator' represents that we use a single generator network $\mathbf{G}_1$ as our baseline without adversarial learning. Similarly, 'Double generators' represents that we apply two generators network($\mathbf{G}_1$, $\mathbf{G}_2$) and combine their output to predict saliency maps. We directly calculate the cross-entropy loss between the ground truth and their predicted maps. Comparing the outputs of our baseline with the MIDD results (Tu et al., 2021), it is easy to find that our method has comparable performance, which demonstrates our features extraction network and the data process are reasonable. In particular, by comparing the results of our method shown in last row with the results with single generator shown in first row in Table 5, the proposed collaborative generator brings an average 50.9% improvement with the MAE evaluation metric for the three datasets and gains consistency with other metrics as well. Then, validations for each component in our method are as following.

Effectiveness of the L_{cs}. Consistency similarity loss between S_1 and S_2 is used to narrow their segmentation gap. The third row in Table 5 expresses its importance. We compare the values of 'with' and 'without' consistency similarity loss, and find that the Fm is improved by 1.1% on VT821, 0.6% on VT1000, and 0.7% on VT5000 testing set. It demonstrates that consistency similarity loss is effective to narrow the gap between two generators' predicted maps and urge both of them close to the ground truth as much as possible.

Effectiveness of the co-attention module. To validate the effectiveness of the cross-modality fusion module for multimodal saliency detection, we compare the proposed fusion strategy COAM with only pixel-wised addition operation between features of two modalities. As shown in Table 5, we can easily find that the our method with the co-attention module achieves better performances when fusing RGB and thermal features. Besides, we visualize the RGB features and the thermal features before entering the co-attention module and the features after the integration with COAM. As shown in Fig. 5, it can be seen that the COAM can effectively suppress the background noise and further highlight the foreground object by fusing RGB features and thermal features.

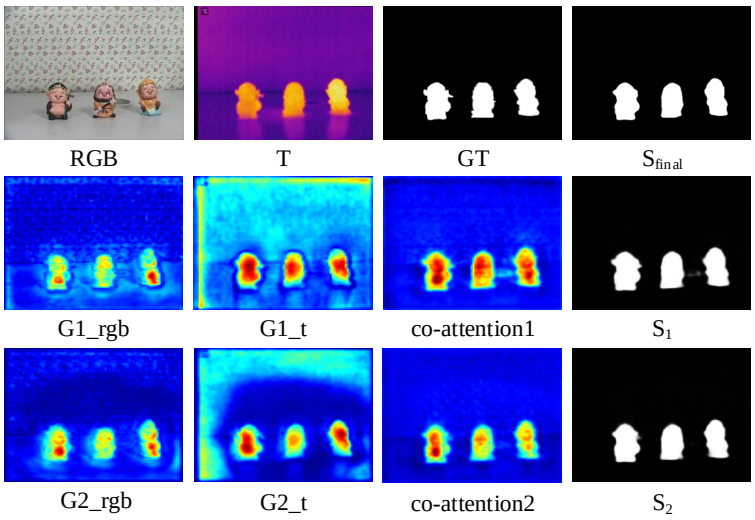

Figure 5: The visualization of the co-attention module. The first row, four images are input RGB image, thermal image, GT and S_{final} sequentially. The second row, four images are output RGB feature map and thermal feature map by the encoder-decoder network in the generator1, the refined feature map by the co-attention module and the predicted map from generator1. The third row is similar with second row, expressing the same progress of generator2.

Table 6: Different sittings of $\lambda 1$ and $\lambda 2$.

$\lambda 1$	$\lambda 2$	Fm	MAE	wfm	$\lambda 1$	$\lambda 2$	Fm	MAE	wfm
10	1	**0.886**	**0.027**	**0.865**	10	10	0.879	**0.027**	0.862
50	1	0.881	0.028	0.861	10	5	**0.886**	0.028	0.860
100	1	0.876	0.029	0.855	10	0.5	0.884	**0.027**	**0.865**

Effectiveness of the adversarial learning. To verify the importance of adversarial learning, we remove the discriminator network and directly use the existing collaborative generator network and the generator loss functions to predict the saliency maps. As shown in Table 5, 'w/o GAN' represents this operation. The experimental results show that without adversarial learning, four metric indexes of the predicted maps on three testing datasets have declined. It is also proved that three generators and discriminators compete with each other in the form of mini-max game. Discriminator classifies the different input maps which include three generators' segmentation maps and the ground truth, three generators predict saliency maps close to the ground truth in their own ways. Finally, the whole network achieves the best performance.

Different settings of λ_1 and λ_2. We test different values of λ_1 and λ_2 in generator loss functions and perform corresponding experiments on VT5000 testing set. We choose the best one, λ_1=10 and λ_2=1 to optimize our model. The results with different settings of λ_1 and λ_2 are shown as Table 6. By observing the results, we find that the model achieves better performance when λ_1 and λ_2 have some difference, but the difference should not be too large.

Different settings of β^2. The β^2 balances the bias between precision and precision. For salient object detection, β^2 in F-measure is usually set to 0.3 (Achanta et al., 2009). Therefore, we set β^2 to 0.3. We also test different values of β^2 in generator loss functions on three RGBT testing set. The results show our method is insensitive to this parameter, and our model achieves best overall results when β^2 is set to 0.3. For example, the performance with MAE on three RGBT SOD datasets do not vary by more than 0.002. The results are shown in Table 7.

Different computation methods for FP and FN. We follow the computation method in MDvsFA-cGAN (Wang et al., 2019) for obtaining FP and FN as shown in Eq.10 to balance miss detection and false alarming. Ad we use the FmLoss (Zhao et al., 2019) to improve F-measure, the computation method for FP and FN shown in Eq.11 is consistent with that work (Zhao et al., 2019). We measure FN, FP from two aspects of L2 norm and summation.

Table 7: Different sittings of β^2

β^2	VT821				VT1000				VT5000			
	Fm	MAE	wfm	Sm	Fm	MAE	wfm	Sm	Fm	MAE	wfm	Sm
0.3	**0.867**	**0.027**	0.845	0.888	**0.926**	**0.015**	**0.922**	**0.936**	**0.886**	0.027	0.865	0.899
0.5	0.859	0.028	0.845	0.892	0.919	0.016	0.919	0.935	0.880	0.027	0.866	0.901
0.8	0.861	0.028	0.841	0.887	0.919	0.016	0.918	0.934	0.881	0.027	0.862	0.897
1.1	0.855	**0.027**	**0.847**	**0.894**	0.911	0.016	0.914	0.932	0.876	**0.026**	**0.867**	**0.902**
1.4	0.850	0.028	0.841	0.890	0.908	0.017	0.911	0.931	0.869	0.027	0.861	0.900

Table 8: FP and FN with different computation rules.

Methods	VT821				VT1000				VT5000			
	Fm	MAE	wfm	Em	Fm	MAE	wfm	Em	Fm	MAE	wfm	Em
MDvsFA-cGAN (Wang et al., 2019)	0.832	0.034	0.804	0.913	0.892	0.019	0.891	0.938	0.845	0.031	0.825	0.926
FLoss(Zhao et al., 2019)	0.856	0.029	0.836	0.926	0.914	0.017	0.910	0.951	0.868	0.031	0.841	0.931
Ours	**0.867**	**0.027**	**0.845**	**0.928**	**0.926**	**0.015**	**0.922**	**0.957**	**0.886**	**0.027**	**0.865**	**0.945**

Considering these two kinds of perspectives together can optimize FP and FN and make them reach a balance. We test the computation method for FP and FN in Eq. 10 and Eq. 11 both using MDvsFA-cGAN or both using FmLoss. The results are shown in Table 8. Our method achieves best performances.

4.6. Comparisons with RGBD SOD Methods

To verify the generalization ability of our method in the RGBD SOD task, we compare it with the latest RGBD SOD methods.

4.6.1. Datasets and Evaluation Metrics

We test six public RGBD SOD datasets, which include NJU2K (Ju et al., 2014), NLPR (Peng et al., 2014), DES (135 paired images) (Cheng et al., 2014), SIP (Fan et al., 2020a), STERE (Niu et al., 2012), and ReDWeb-S(Liu et al., 2021a). NJUD and NLPR consist of 1985 and 1000 RGBD image pairs separately, containing many different scenes. DES is a small dataset containing 135 simple RGBD images. SIP is a dataset with 929 high-quality samples which are designed for salient person detection. STERE contains 1000 RGBD image pairs, which are collected from the Internet with coarse depth quality. ReDWeb-S obtains 3179 images with high-quality depth maps. We employ Sm, Em, Fm, and MAE for quantitative evaluations. Fm is also a primary metric, which is the weighted harmonic mean of Precision and Recall.

4.6.2. Experiment Setup

We choose 1485 paired images in NJU2K and 700 paired images in NLPR as the training set, which is widely used in RGBD SOD task. The remaining images are used for testing. We compute the mean and standard deviation of all the images in the training set and use them to normalize the original inputs. The other experimental parameter settings are the same as the previous experimental settings in RGBT SOD task.

4.6.3. Compare with the State-of-the-art Methods

On RGBD SOD datasets, our method also achieves comparable performance. We compare our method with other 14 existing RGBD SOD methods, including D3Net (Fan et al., 2020a), ICNet (Li et al., 2020), DCMF (Chen et al., 2020), SSF (Zhang et al., 2020a), S2MA (Liu et al., 2020), A2dele (Piao et al., 2020), CoNet (Ji et al., 2020), DANet (Zhao et al., 2020), DSA2F (Sun et al., 2021), MMNet (Gao et al., 2022), CDNet (Jin et al., 2021),

Table 9: Performance comparison with 14 methods on six RGBD datasets. The best scores are highlighted in **BOLD**.

Methods	NLPR				NJU2K				DES				SIP				STERE				ReDWeb-S			
	Sm	*Em*	*Fm*	*MAE*	*Sm*	*Em*	*Fm*	*MAE*	*Sm*	*Em*	*Fm*	*MAE*	*Sm*	*Em*	*Fm*	*MAE*	*Sm*	*Em*	*Fm*	*MAE*	*Sm*	*Em*	*Fm*	*MAE*
D3Net (Fan et al., 2020a)	0.912	0.944	0.861	0.030	0.901	0.914	0.865	0.046	0.898	0.951	0.870	0.031	0.860	0.902	0.835	0.063	0.899	0.920	0.859	0.046	0.689	0.742	0.664	0.149
ICNet(Li et al., 2020)	0.923	0.944	0.870	0.028	0.894	0.905	0.868	0.052	0.920	0.959	0.889	0.027	0.854	0.899	0.836	0.069	0.903	0.915	0.865	0.045	-	-	-	-
DCMF (Chen et al., 2020)	0.900	0.933	0.839	0.035	0.889	0.897	0.859	0.052	0.877	0.923	0.820	0.040	0.859	0.898	0.819	0.068	0.883	0.904	0.841	0.054	0.675	0.742	0.653	0.160
SSF (Zhang et al., 2020a)	0.914	0.949	0.875	0.026	0.899	0.913	0.886	0.043	0.905	0.948	0.876	0.025	0.868	0.911	0.851	0.056	0.887	0.921	0.867	0.046	0.595	0.684	0.559	0.189
S2MA2(Liu et al., 2020)	0.915	0.938	0.853	0.030	0.894	0.896	0.865	0.053	0.941	0.974	0.906	0.021	0.872	0.911	0.854	0.057	0.890	0.907	0.855	0.051	0.711	0.781	0.696	0.139
A2dele (Piao et al., 2020)	0.896	0.945	0.878	0.028	0.869	0.897	0.874	0.051	0.885	0.922	0.865	0.028	0.826	0.892	0.825	0.070	0.878	0.915	0.874	0.044	0.641	0.672	0.603	0.160
CoNet2 (Ji et al., 2020)	0.908	0.934	0.846	0.031	0.895	0.912	0.872	0.046	0.911	0.945	0.861	0.027	0.858	0.909	0.842	0.063	0.905	0.927	0.884	0.037	0.696	0.762	0.688	0.147
DANet (Zhao et al., 2020)	0.920	0.951	0.875	0.027	0.899	0.908	0.871	0.045	0.924	0.968	0.899	0.023	0.875	0.914	0.855	0.054	0.901	0.921	0.868	0.04	0.693	0.753	0.684	0.142
DSA2F (Sun et al., 2021)	0.918	0.950	0.892	0.024	0.904	0.922	0.898	0.039	0.916	0.955	0.901	0.023	0.862	0.908	0.865	0.057	0.897	0.927	0.893	0.039	-	-	-	-
MMNet (Gao et al., 2022)	0.925	0.950	0.889	0.024	0.911	0.919	0.900	0.038	0.830	0.893	0.746	0.058	0.836	0.882	0.839	0.075	0.891	0.924	0.880	0.045	-	-	-	-
CDNet (Jin et al., 2021)	0.902	0.935	0.848	0.032	0.885	0.911	0.866	0.048	0.875	0.921	0.839	0.034	0.823	0.880	0.805	0.076	0.896	0.922	0.873	0.042	0.693	0.733	0.684	0.137
RD3D (Chen et al., 2021)	0.930	0.958	0.892	0.022	**0.916**	0.918	0.901	0.036	**0.935**	**0.975**	0.917	0.019	0.885	0.920	0.874	0.048	**0.911**	0.927	0.886	0.037	0.689	0.742	0.664	0.149
HAINet2 (Li et al., 2021)	0.924	0.957	0.897	0.024	0.912	0.922	0.900	0.038	**0.935**	0.974	0.924	**0.018**	0.880	0.919	0.875	0.053	0.907	0.925	0.885	0.040	**0.724**	**0.766**	0.713	0.132
DCF (Ji et al., 2021)	0.923	0.957	0.890	0.022	0.912	**0.924**	0.902	**0.035**	-	-	-	-	0.875	0.920	0.875	0.052	0.902	0.929	0.884	0.039	0.709	0.755	0.710	0.135
Ours	**0.932**	**0.966**	**0.925**	**0.018**	0.892	0.923	**0.909**	0.041	0.923	0.960	**0.927**	0.019	**0.889**	**0.928**	**0.911**	**0.043**	0.910	**0.940**	**0.907**	**0.033**	0.713	0.739	**0.722**	**0.124**

RD3D (Chen et al., 2021), HAINet (Li et al., 2021), DCF (Ji et al., 2021). We directly evaluate the provided saliency maps or use available codes with provided models to predict saliency maps. The comparison results with Fm, MAE, Sm, and Em are shown in Table 9, it is easy to find that our experimental results on all RGBD datasets achieve the highest Fm and the overall results outperform other methods, which demonstrates the rationality of the design of our collaborative generator and the robustness of our method. We also conduct the experiments on a new RGBD dataset: ReDWeb-S (Liu et al., 2021a). The experimental results on this dataset show that our method achieves the highest Fm which is the primary evaluation metric in SOD. Due to the lack of samples in DES dataset and the low diversity of samples in STERE dataset, our method does not achieve the best performance on some metrics and the experimental results lack the stability. To solve this problem, we can design a special fusion method in the future. In addition, although our method shows superior performance on both RGBD and RGBT datasets, demonstrating its robustness and effect. However, it requires different data to train our model separately, which is inconvenient. In the future, we plan to improve this problem by designing the corresponding modules to make the model ease of use.

5. Conclusion

In this paper, we propose a novel method for multimodal saliency detection, named ALCG, which is based on the conditional generative adversarial network and co-attention fusion strategy. Considering two errors(ie., FN and FP) and Fm of the saliency maps, we decompose the complex multimodal SOD into three sub-tasks, improving Recall, Precision, and Fm by three generators respectively. Moreover, the co-attention module for fusing cross-modal features is helpful to explore the complementary information between the RGB modality and the thermal modality and remove the impacts of noises in features simultaneously. Experiment results show our method has superior performance against state-of-the-art methods on three RGBT SOD datasets and six RGBD SOD datasets, demonstrating the rationality and effectiveness of our proposed method for multimodal saliency detection.

6. Acknowledgement

This research is jointly supported by The University Synergy Innovation Program of Anhui Province (No.GXXT-2022-014, No.GXXT-2020-015, No.GXXT-2021-065), Natural Science Foundation of Anhui Higher Education Institution (KJ2020A0040), Natural Science Foundation of Anhui Higher Education Institution of China (No.KJ2020A0033), Anhui Provincial Natural Science foundation (No.2108085MF211), UK Engineering and Physical Sciences Research Council (EPSRC) - Grants Ref. EP/M026981/1, EP/T021063/1, EP/T024917/1.

References

Achanta, R., Hemami, S., Estrada, F., Susstrunk, S., 2009. Frequency-tuned salient region detection, in: IEEE Conference on Computer Vision and Pattern Recognition.

Achanta, R., Hemami, S., Estrada, F., Susstrunk, S., 2009. Frequency-tuned salient region detection .

Ben-Cohen, A., Klang, E., Raskin, S.P., Soffer, S., Ben-Haim, S., Konen, E., Amitai, M.M., Greenspan, H., 2019. Cross-modality synthesis from ct to pet using fcn and gan networks for improved automated lesion detection. Engineering Applications of Artificial Intelligence 78.

Chen, G., Liu, L., Hu, W., Pan, Z., 2018. Semi-supervised object detection in remote sensing images using generative adversarial networks, in: IEEE International Geoscience and Remote Sensing Symposium.

Chen, H., Deng, Y., Li, Y., Hung, T.Y., Lin, G., 2020. Rgbd salient object detection via disentangled cross-modal fusion. IEEE Transactions on Image Processing 29.

Chen, H., Li, Y., 2018. Progressively complementarity-aware fusion network for rgb-d salient object detection, in: IEEE Conference on Computer Vision and Pattern Recognition.

Chen, Q., Liu, Z., Zhang, Y., Fu, K., Zhao, Q., Du, H., 2021. Rgb-d salient object detection via 3d convolutional neural networks 35.

Cheng, Y., Fu, H., Wei, X., Xiao, J., Cao, X., 2014. Depth enhanced saliency detection method, in: Proceedings of International Conference on Internet Multimedia Computing and Service.

Cong, R., Lei, J., Fu, H., Huang, Q., Cao, X., Ling, N., 2019. Hscs: Hierarchical sparsity based co-saliency detection for rgbd images. IEEE Transactions on Multimedia 21.

Dou, Q., Ouyang, C., Chen, C., Chen, H., Heng, P.A., 2018. Unsupervised cross-modality domain adaptation of convnets for biomedical image segmentations with adversarial loss. arXiv preprint arXiv:1804.10916, 2018 .

Fan, D.P., Cheng, M.M., Liu, Y., Li, T., Borji, A., 2017. Structure-measure: A new way to evaluate foreground maps, in: IEEE International Conference on Computer Vision.

Fan, D.P., Gong, C., Cao, Y., B., R., Cheng, M.M., Borji, A., 2018. Enhanced-alignment measure for binary foreground map evaluation, in: International Joint Conference on Artificial Intelligence.

Fan, D.P., Lin, Z., Zhang, Z., Zhu, M., Cheng, M.M., 2020a. Rethinking rgb-d salient object detection: Models, data sets, and large-scale benchmarks. IEEE Transactions on Neural Networks and Learning Systems 32.

Fan, D.P., Zhai, Y., Borji, A., Yang, J., Shao, L., 2020b. Bbs-net: Rgb-d salient object detection with a bifurcated backbone strategy network, in: European Conference on Computer Vision, Springer.

Fang, H., Gupta, S., Iandola, F., Srivastava, R.K., Deng, L., Dollár, P., Gao, J., He, X., Mitchell, M., Platt, J.C., et al., 2015. From captions to visual concepts and back, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Fei, W., Jiang, M., Chen, Q., Yang, S., Tang, X., 2017. Residual attention network for image classification, in: IEEE Conference on Computer Vision and Pattern Recognition.

Fu, K., Fan, D.P., Ji, G.P., Zhao, Q., 2020. Jl-dcf: Joint learning and densely-cooperative fusion framework for rgb-d salient object detection, in: IEEE Conference on Computer Vision and Pattern Recognition.

Gao, W., Liao, G., Ma, S., Li, G., Liang, Y., Lin, W., 2022. Unified information fusion network for multi-modal rgb-d and rgb-t salient object detection. IEEE Transactions on Circuits and Systems for Video Technology 32.

Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Bing, X., Bengio, Y., 2014. Generative adversarial nets. MIT Press .

Han, J., Chen, H., Liu, N., Yan, C., Li, X., 2018. Cnns-based rgb-d saliency detection via cross-view transfer and multiview fusion. IEEE Transactions on Cybernetics 48.

Hao, Y., Wang, N., Li, J., Gao, X., 2019. Hsme: Hypersphere manifold embedding for visible thermal person re-identification. Proceedings of the AAAI Conference on Artificial Intelligence 33.

Hu, J., Shen, L., Sun, G., 2018. Squeeze-and-excitation networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Huo, F., Zhu, X., Zhang, L., Liu, Q., Shu, Y., 2022a. Efficient context-guided stacked refinement network for rgb-t salient object detection. IEEE Transactions on Circuits and Systems for Video Technology 32.

Huo, F., Zhu, X., Zhang, Q., Liu, Z., Yu, W., 2022b. Real-time one-stream semantic-guided refinement network for rgb-thermal salient object detection. IEEE Transactions on Instrumentation and Measurement 71.

Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A., 2016. Image-to-image translation with conditional adversarial networks. IEEE Conference on Computer Vision and Pattern Recognition .

Ji, W., Li, J., Yu, S., Zhang, M., Piao, Y., Yao, S., Bi, Q., Ma, K., Zheng, Y., Lu, H., et al., 2021. Calibrated rgb-d salient object detection, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.

Ji, W., Li, J., Zhang, M., Piao, Y., Lu, H., 2020. Accurate rgb-d salient object detection via collaborative learning, in: European Conference on Computer Vision, Springer.

Jiang, B., Zhou, Z., Wang, X., Tang, J., Luo, B., 2020. cmsalgan: Rgb-d salient object detection with cross-view generative adversarial networks. IEEE Transactions on Multimedia 23.

Jin, W.D., Xu, J., Han, Q., Zhang, Y., Cheng, M.M., 2021. Cdnet: Complementary depth network for rgb-d salient object detection. IEEE Transactions on Image Processing 30.

Ju, R., Ge, L., Geng, W., Ren, T., Wu, G., 2014. Depth saliency based on anisotropic center-surround difference, in: IEEE International Conference on Image Processing.

Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 .

Li, C., Hui, C., Hu, S., Liu, X., Liang, L., 2016. Learning collaborative sparse representation for grayscale-thermal tracking. IEEE Transactions on Image Processing 25.

Li, C., Liang, X., Lu, Y., Zhao, N., Tang, J., 2019. Rgb-t object tracking: Benchmark and baseline. Pattern Recognition 96.

Li, C., Wang, G., Ma, Y., Zheng, A., Luo, B., Tang, J., 2018. Rgb-t saliency detection benchmark: Dataset, baselines, analysis and a novel approach, in: Chinese Conference on Image and Graphics Technologies.

Li, G., Liu, Z., Chen, M., Bai, Z., Lin, W., Ling, H., 2021. Hierarchical alternate interaction network for rgb-d salient object detection. IEEE Transactions on Image Processing 30.

Li, G., Liu, Z., Ling, H., 2020. Icnet: Information conversion network for rgb-d based salient object detection. IEEE Transactions on Image Processing 29.

Liang, F., Duan, L., Ma, W., Qiao, Y., Cai, Z., Qing, L., 2018. Stereoscopic saliency model using contrast and depth-guided-background prior. Neurocomputing 275.
Liang, Y., Qin, G., Sun, M., Qin, J., Yan, J., Zhang, Z., 2022. Multi-modal interactive attention and dual progressive decoding network for rgb-d/t salient object detection. Neurocomputing 490.
Liao, G., Gao, W., Li, G., Wang, J., Kwong, S., 2022. Cross-collaborative fusion-encoder network for robust rgb-thermal salient object detection. IEEE Transactions on Circuits and Systems for Video Technology 32.
Liu, N., Zhang, N., Han, J., 2020. Learning selective self-mutual attention for rgb-d saliency detection, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition.
Liu, N., Zhang, N., Shao, L., Han, J., 2021a. Learning selective mutual attention and contrast for rgb-d saliency detection. IEEE Transactions on Pattern Analysis and Machine Intelligence .
Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021b. Swin transformer: Hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF International Conference on Computer Vision.
Liu, Z., Shi, S., Duan, Q., Zhang, W., Zhao, P., 2019. Salient object detection for rgb-d image by single stream recurrent convolution neural network. Neurocomputing 363.
Liu, Z., Zhang, W., Zhao, P., 2020. A cross-modal adaptive gated fusion generative adversarial network for rgb-d salient object detection. Neurocomputing 387.
Liu, Z.y., Liu, J.w., Zuo, X., Hu, M.f., 2021c. Multi-scale iterative refinement network for rgb-d salient object detection. Engineering Applications of Artificial Intelligence 106.
Mahadevan, V., Vasconcelos, N., 2009. Saliency-based discriminant tracking, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.
Margolin, R., Zelnik-Manor, L., Tal, A., 2014. How to evaluate foreground maps, in: IEEE Conference on Computer Vision and Pattern Recognition.
Niu, Y., Geng, Y., Li, X., Liu, F., 2012. Leveraging stereopsis for saliency analysis, in: IEEE Conference on Computer Vision and Pattern Recognition.
Noori, M., Mohammadi, S., Majelan, S.G., Bahri, A., Havaei, M., 2020. Dfnet: Discriminative feature extraction and integration network for salient object detection. Engineering Applications of Artificial Intelligence 89.
Pan, J., Ferrer, C.C., McGuinness, K., O'Connor, N.E., Torres, J., Sayrol, E., Giro-i Nieto, X., 2017. Salgan: Visual saliency prediction with generative adversarial networks. arXiv preprint arXiv:1701.01081 .
Pan, X., Zhao, J., Xu, J., 2021. Conditional generative adversarial network-based training sample set improvement model for the semantic segmentation of high-resolution remote sensing images. IEEE Transactions on Geoscience and Remote Sensing 59.
Peng, H., Bing, L., Xiong, W., Hu, W., Ji, R., 2014. Rgbd salient object detection: A benchmark and algorithms, in: European Conference on Computer Vision.
Perazzi, F., Kr?henb¨uhl, P., Pritch, Y., Hornung, A., 2012. Saliency filters: Contrast based filtering for salient region detection, in: IEEE Conference on Computer Vision and Pattern Recognition.
Piao, Y., Ji, W., Li, J., Zhang, M., Lu, H., 2019. Depth-induced multi-scale recurrent attention network for saliency detection, in: IEEE/CVF International Conference on Computer Vision.
Piao, Y., Rong, Z., Zhang, M., Ren, W., Lu, H., 2020. A2dele: Adaptive and attentive depth distiller for efficient rgb-d salient object detection, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition.
Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical Image Computing and Computer-assisted Intervention, Springer.
Sun, P., Zhang, W., Wang, H., Li, S., Li, X., 2021. Deep rgb-d saliency detection with depth-sensitive attention and automatic multi-modal fusion, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.
Tu, Z., Li, Z., Li, C., Lang, Y., Tang, J., 2021. Multi-interactive dual-decoder for rgb-thermal salient object detection. IEEE Transactions on Image Processing 30.
Tu, Z., Li, Z., Li, C., Tang, J., 2022a. Weakly alignment-free rgbt salient object detection with deep correlation network. IEEE Transactions on Image Processing 31.
Tu, Z., Ma, Y., Li, Z., Li, C., Xu, J., Liu, Y., 2022b. Rgbt salient object detection: A large-scale dataset and benchmark. IEEE Transactions on Multimedia .
Tu, Z., Xia, T., Li, C., Lu, Y., Tang, J., 2019. M3s-nir: Multi-modal multi-scale noise-insensitive ranking for rgb-t saliency detection, in: IEEE Conference on Multimedia Information Processing and Retrieval.
Tu, Z., Xia, T., Li, C., Wang, X., Ma, Y., Tang, J., 2020. Rgb-t image saliency detection via collaborative graph learning. IEEE Transactions on Multimedia 22.
Wang, A., Wang, M., 2017. Rgb-d salient object detection via minimum barrier distance transform and saliency fusion. IEEE Signal Processing Letters 24.
Wang, H., Zhou, L., Wang, L., 2019. Miss detection vs. false alarm: Adversarial learning for small object segmentation in infrared images, in: Proceedings of the IEEE/CVF International Conference on Computer Vision.
Wang, J., Song, K., Bao, Y., Huang, L., Yan, Y., 2022. Cgfnet: Cross-guided fusion network for rgb-t salient object detection. IEEE Transactions on Circuits and Systems for Video Technology 32.
Wang, N., Gong, X., 2019. Adaptive fusion for rgb-d salient object detection. IEEE Access 7.
Wei, J., Wang, S., Huang, Q., 2020. F3net: Fusion, feedback and focus for salient object detection. Proceedings of the AAAI Conference on Artificial Intelligence 34.
Ye, M., Zheng, W., Lan, X., Yuen, P.C., 2018. Visible thermal person re-identification via dual-constrained top-ranking, in: International Joint Conference on Artificial Intelligence.
Zhang, M., Ren, W., Piao, Y., Rong, Z., Lu, H., 2020a. Select, supplement and focus for rgb-d saliency detection, in: IEEE Conference on Computer Vision and Pattern Recognition.

Zhang, Q., Huang, N., Yao, L., Zhang, D., Shan, C., Han, J., 2020b. Rgb-t salient object detection via fusing multi-level cnn features. IEEE Transactions on Image Processing 29.

Zhang, Q., Xiao, T., Huang, N., Zhang, D., Han, J., 2020c. Revisiting feature fusion for rgb-t salient object detection. IEEE Transactions on Circuits and Systems for Video Technology 31.

Zhao, K., Gao, S., Wang, W., Cheng, M.M., 2019. Optimizing the f-measure for threshold-free salient object detection, in: IEEE/CVF International Conference on Computer Vision.

Zhao, R., Ouyang, W., Wang, X., 2013. Unsupervised salience learning for person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Zhao, X., Zhang, L., Pang, Y., Lu, H., Zhang, L., 2020. A single stream network for robust and real-time rgb-d salient object detection, in: European Conference on Computer Vision, Springer.

Zhou, W., Guo, Q., Lei, J., Yu, L., Hwang, J.N., 2022a. Ecffnet: Effective and consistent feature fusion network for rgb-t salient object detection. IEEE Transactions on Circuits and Systems for Video Technology 32.

Zhou, W., Zhu, Y., Lei, J., Wan, J., Yu, L., 2022b. Apnet: Adversarial learning assistance and perceived importance fusion network for all-day rgb-t salient object detection. IEEE Transactions on Emerging Topics in Computational Intelligence 6.

Zhu, C., Li, G., Wang, W., Wang, R., 2017. An innovative salient object detection using center-dark channel prior, in: Proceedings of the IEEE International Conference on Computer Vision Workshops.

Zhu, J.Y., Wu, J., Xu, Y., Chang, E., Tu, Z., 2014. Unsupervised object class discovery via saliency-guided multiple class learning. IEEE Transactions on Pattern Analysis and Machine Intelligence 37.