electronics

MDPI

Article

CAE-Net: Cross-Modal Attention Enhancement Network for RGB-T Salient Object Detection

Chengtao Lv [1], Bin Wan [1], Xiaofei Zhou [1,*], Yaoqi Sun [1,2], Ji Hu [1,2], Jiyong Zhang [1] and Chenggang Yan [1]

[1] School of Automation, Hangzhou Dianzi University, Hangzhou 310018, China
[2] Lishui Institute of Hangzhou Dianzi University, Lishui 323000, China
* Correspondence: zxforchid@outlook.com

Abstract: RGB salient object detection (SOD) performs poorly in low-contrast and complex background scenes. Fortunately, the thermal infrared image can capture the heat distribution of scenes as complementary information to the RGB image, so the RGB-T SOD has recently attracted more and more attention. Many researchers have committed to accelerating the development of RGB-T SOD, but some problems still remain to be solved. For example, the defective sample and interfering information contained in the RGB or thermal image hinder the model from learning proper saliency features, meanwhile the low-level features with noisy information result in incomplete salient objects or false positive detection. To solve these problems, we design a cross-modal attention enhancement network (CAE-Net). First, we concretely design a cross-modal fusion (CMF) module to fuse cross-modal features, where the cross-attention unit (CAU) is employed to enhance the two modal features, and channel attention is used to dynamically weigh and fuse the two modal features. Then, we design the joint-modality decoder (JMD) to fuse cross-level features, where the low-level features are purified by higher level features, and multi-scale features are sufficiently integrated. Besides, we add two single-modality decoder (SMD) branches to preserve more modality-specific information. Finally, we employ a multi-stream fusion (MSF) module to fuse three decoders' features. Comprehensive experiments are conducted on three RGB-T datasets, and the results show that our CAE-Net is comparable to the other methods.

Keywords: salient object detection; multi-stream fusion; cross-attention unit; cross-modal fusion; single-/joint-modality decoder

Citation: Lv, C.; Wan, B.; Zhou, X.; Sun, Y.; Hu, J.; Zhang, J.; Yan, C. CAE-Net: Cross-Modal Attention Enhancement Network for RGB-T Salient Object Detection. *Electronics* **2023**, *12*, 953. https://doi.org/10.3390/electronics12040953

Academic Editor: Oscar Deniz Suarez

Received: 5 January 2023
Revised: 8 February 2023
Accepted: 12 February 2023
Published: 14 February 2023

1. Introduction

Salient object detection (SOD) attempts to imitate the human's attention mechanism, which can discover the most attractive objects in the image at first glance, to segment out the saliency objects in the image. SOD can be applied in many downstream computer vision tasks, such as object tracking [1], image quality assessment [2], scene classification [3], image fusion [4], and so on. Due to its superior performance in downstream tasks, SOD has received more and more attention in recent years.

The RGB SOD has been studied for many years. In the beginning, researchers proposed many traditional methods, which involve designing handcrafted features to estimate saliency maps. These methods cannot explore the high-level semantic information contained in the image, so it leads to unsatisfactory results. Benefiting from powerful feature representation ability, convolutional neural networks (CNNs) [5] are receiving more and more attention in computer vision applications. Particularly, when the fully convolutional networks (FCN) [6] and Unet [7] were proposed in image segmentation tasks, researchers gradually turned to embracing the deep learning-based method in SOD. Many works have been proposed in SOD. For example, to take into account the long range correlation of deep features between different positions, many works [8,9] employed ASPP [10], RFB [11], or PPM [12] modules. By using these modules, the context information of salient objects can

be fully exploited. Similarly, Pang [13] employed multi-branch features interaction to fully explore multi-scale information. Besides, edge features were also explicitly explored by many works to portray sharp boundaries of salient objects [14,15]. Though great progress has been made in recent years, RGB SOD suffers interference from low-contrast or complex background images, resulting in a poor quality saliency map. With the development of sensor technology, we can easily afford the expenditure of depth or thermal sensors. The depth image provides a description of the spatial arrangement information of the scene. By introducing the depth information, we can easily distinguish objects with different depths. However, due to the vulnerability of depth sensors to environmental changes, low-quality depth maps exist in RGB-D datasets, resulting in the decline performance of RGB-D SOD. Different from the depth information, the thermal infrared image depicts the radiated heat information of objects in the scene, so it can help us easily distinguish salient objects.

RGB-T SOD faces the problems of multi-modal feature fusion. Previous works have explored cross-modal complementary information. In [16], the multi-interactive block is designed to fuse the previous layer's decoded features with two modal features, respectively, which are afterwards concatenated to perform cross-modal fusion. In [17], the context-guided cross-modality fusion is designed to fuse two modal features using element-wise multiplication and addition at each level, and then they are fed into a stacked refinement network to decode them. Nevertheless, direct concatenation or element-wise addition/multiplication cannot fully explore the complementary information between two modal features. Besides, there are some poor quality examples in the RGB or thermal infrared image, as shown in Figure 1. If we indiscriminately concatenate or add two modal features together, the bad quality samples will mislead the saliency model, resulting in incorrect prediction results. Therefore, we need to carefully design a module to appropriately merge two modal features. In addition, similar to RGB SOD, many works have been committed to exploring multi-scale information embedded in deep features. For example, in [17], the surrounding and global context unit was proposed to capture context information. Considering that each level feature contains different scale information, where high-level features contain more semantic and holistic information, and low-level features contain more detail and local information. Properly aggregating the cross-level features and simultaneously reducing the noise impact are worth further investigating.

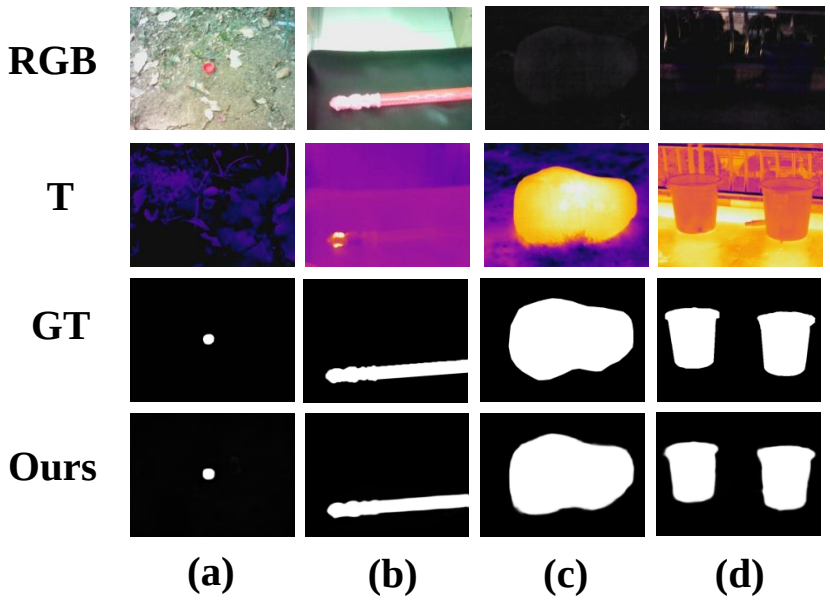

Figure 1. Some bad quality examples of RGB or thermal infrared images. (**a**,**b**) are two samples with bad quality thermal images, and (**c**,**d**) are two samples with bad quality RGB images. GT denotes groundtruth, and ours indicates the saliency maps predicted by our proposed method.

To solve these problems mentioned above, we propose a novel cross-modal attention enhancement network (CAE-Net) for RGB-T salient object detection, which is shown in Figure 2. Benefiting from three key components (i.e., cross-modal fusion (CMF), single-/joint-modality decoder (SMD/JMD), and multi-stream fusion (MSF)), the CAE-Net can fully exploit cross-modal information and suitably fuse them. Besides, it can adequately aggregate cross-level features in a gradually refined manner. Concretely, to fuse cross-modal features, we design a cross-modal fusion (CMF) module, where the cross-attention unit (CAU) is constructed to enhance the one modal feature using the attention from another modal feature, and then we employ channel attention to adaptively emphasize the significant modal features and restrain the deficient modal features. Then, to preferably fuse the cross-level features, we design the joint-modality decoder (JMD), where high-level features refine low-level features to suppress noisy information and sufficiently gather multi-scale features. Besides, we add two independent single-modality decoder (SMD) branches to preserve more modality-specific information [18] contained in the RGB and thermal image, respectively. Finally, we design the multi-stream fusion (MSF) module to fully explore complementary information between different decoder branches. With our elaborate design, the proposed model can better explore complementary information between cross-modal features and appropriately aggregate cross-level features.

Overall, we summarize the main contributions of our paper as follows:

1. We propose a novel RGB-T salient object detection model, called a cross-modal attention enhancement network (CAE-Net), which consists of the cross-modal fusion (CMF), the single-/joint-modality decoder (SMD/JMD), and multi-stream fusion (MSF).
2. To fuse the cross-modal features, we design a cross-modal fusion (CMF) module, where the cross-attention unit (CAU) is employed to filter incompatible information, and the channel attention is used to emphasize the significant modal features.
3. To fuse cross-level features, we design the joint-modality decoder (JMD) module, where the multi-scale features are extracted and aggregated, and noisy information is filtered. Besides, two independent single-modality decoder (SMD) branches are employed to preserve more modality-specific information.
4. To fully explore the complementary information between different decoder branches, we design a multi-stream fusion (MSF) module.

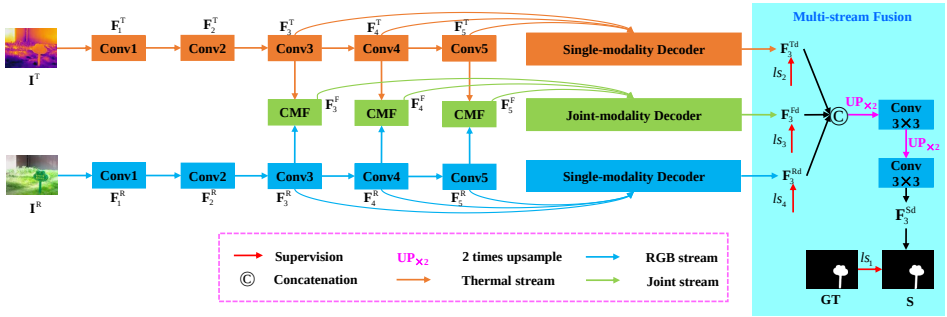

Figure 2. The overall architecture of our model's cross-modal attention enhancement network (CAE-Net). Firstly, we use double stream encoder to extract multi-level features of RGB image $\mathbf{I}^R$ and thermal infrared image $\mathbf{I}^T$, respectively, producing five level-deep features $\{\mathbf{F}_i^R, \mathbf{F}_i^T\}_{(i=1,\cdots,5)}$ for them. Then, we design a cross-modal fusion (CMF) module, which consists of cross-attention unit (CAU) and channel attention weighted fusion, to fuse two modal deep features, obtaining the fused features $\{\mathbf{F}_i^F\}_{(i=3,4,5)}$. After that, we design the joint-modality decoder (JMD) to fuse cross-level features and obtain decoded feature $\mathbf{F}_3^{Fd}$. We also add two independent single-modality decoder (SMD) branches to preserve more modality-specific information, obtaining decoded features $\mathbf{F}_3^{Rd}$ and $\mathbf{F}_3^{Td}$, respectively. Finally, we design a multi-stream fusion (MSF) module to fully fuse complementary information between different decoder branches and obtain the final fused feature $\mathbf{F}_3^{Sd}$. $\mathbf{S}$ is the final saliency map, which is obtained by applying one 1×1 convolution on $\mathbf{F}_3^{Sd}$. Here, the supervision loss of $\mathbf{S}$ and intermediate features are denoted as $ls_{i(i=1,\cdots,4)}$, which are marked with a red arrow in this figure.

We organize the remaining part of this paper as follows. We briefly conclude the related works of salient object detection in Section 2. In Section 3, we describe the proposed model in detail. In Section 4, we show the comprehensive experiments and detailed analyses. Finally, this article is concluded in Section 5.

2. Related Works

In recent years, a large number of works have been proposed for salient object detection. Here, we briefly introduce RGB saliency models, RGB-D saliency models, and RGB-T saliency models.

2.1. RGB Salient Object Detection

In the beginning, researchers employed hand-crafted features and a variety of prior knowledge to determine saliency. For instance, the center-surrounding discrepancy mechanism [19] was employed to distinguish salient objects. Afterward, traditional machine-learning models were developed. In [20], multiple types of features were combined, which consist of multiscale contrast, spatial color distribution, and center-surrounded histogram, by learning conditional random field. In [21], the saliency score is predicted by fusing a multi-level regional feature vector through supervised learning. The convolutional neural network (CNNs) [5] has been widely used in many applications due to its powerful representation learning ability. Particularly, when Unet [7] and fully convolutional networks (FCN) [6] are proposed in image segmentation tasks, CNN-based models dominated in saliency detection. For example, Wu et al. [8] designed a cascaded partial decoder, where low-level features are refined by initial saliency maps, which are predicted by exploiting high-level features. Besides, many researchers have tried their best to recover boundary details of saliency maps [15]. In [22], a boundary-aware loss function and refinement module are used to depict boundaries and purify coarse prediction maps, which effectively cause the boundaries to be clearer. In [14], fine detail saliency maps are predicted by integrating salient object features and edge features, which are produced by exploiting global features and edge features. Wan et al. [23] designed a deeper feature extraction module to enhance the deep feature representation, in which a bidirectional feature extraction unit is designed. Liu et al. [9] employed a parallel multiscale pooling to capture different scale objects. Xu et al. [24] proposed a center-pooling algorithm, where the receptive field is dynamically modified, to take into account the different importance of different regions. In [25], dense attention mechanisms were employed in the decoder to guide the low-level features concentrated on the defect regions.

Though researchers have great progressed RGB saliency detection, complex scenes, such as clutter background and low contrast, will degrade the performance RGB saliency models.

2.2. RGB-D Salient Object Detection

In recent years, we can easily obtain the depth information of scenes with the development of hardware such as laser scanner and Kinect. With the help of a depth map, the challenge of complex scenes for saliency models can be overcome via understanding spatial layout cues. Many researchers have worked to promote the progress of it. The final saliency map is produced by employing the center-dark channel map in [26]. Recently, many CNN-based models have been proposed. For example, in [27], the residual connection is used to fuse the RGB and depth complementary information. The author combined depth features with multi-scale features to single out salient objects. Wang et al. [28] designed two streams to generate saliency maps for depth and RGB, respectively. Then, the switch map, which is learned by the saliency fusion module, fuses two saliency maps. In [29], RGB is processed by the master network, the depth becomes a full exploit because of the sub-network, and the depth-based features are incorporated into the master network. The two modal high-level features, including the depth features and RGB features, are fused by a selective self-mutual attention module in [30], and the depth decoder features are fused into RGB branch by introducing the residual fusion module. Multi-level features

are fused by a densely cooperative fusion (DCF), and collaborative features are learned by joint learning (JL) in [31]. In [32], attention maps were generated from depth cues to intensify salient regions. Besides, in [33], the multi-modal features are fused by employing a cross-modality feature modulation, which consists of spatial selection and channel selection. Wen et al. [34] designed a bi-directional gated pooling module to strengthen the multi-scale information, and gated-based selection to optimize cross-level information. Generally, the encouraging performance is presented by existing RGB-D saliency models, but inaccurate depth maps still degrade their performance.

2.3. RGB-T Salient Object Detection

The thermal infrared image can provide temperature field distribution of scenes, so it plays a positive role when the depth map cannot differentiate salient objects and backgrounds. In the beginning, traditional methods were proposed. In [35], the reliability was described for each modality by introducing a weight, and the weight was integrated into a graph-based manifold ranking method to achieve the adaptive fusion of different source data. Tu et al. [36] segmented RGB and thermal images into multi-scale superpixels. Then, these superpixels were used as graph nodes, and the manifold ranking was performed to obtain saliency maps. In [37], superpixels were used as graph nodes, and then the hierarchical features were used to learn graph affinity and node saliency. With the development of CNNs, deep learning-based methods were broadly employed. Zhang et al. [38] employed multi-branch group fusion to fuse the cross-modal features and designed a joint-attention guided bi-direction message passing to integrate multi-level features. In [39], feature representations were explored and integrated using cross-modal multi-stage fusion. Then, the bi-directional multi-scale decoder was proposed to learn the combination of multi-level fused features. Tu et al. [16] built a dual decoder to conduct interactions of global contexts, two modalities, and multi-level features. Huo et al. [17] established the context-guided cross-modality fusion to explore the complementary information of two modalities, and the features were refined using a stacked refinement network by spatial and semantic information interaction. In [40], multi-level features were extracted and aggregated with the attention mechanisms, and edge loss was used to portray boundaries.

Although much work has been performed on RGB-T SOD, there are still many problems that have not been fully explored. The majority of RGB-T SOD models employ concatenation or element-wise addition/multiplication to fuse the cross-modal features, but these fusion methods do not take into account the distinct significance of two modal features, leading to suboptimal results. Moreover, by employing vanilla Unet to decode cross-level features, the saliency models cannot sufficiently excavate the global context information embedded in deep features, and it is easily interfered by noise in low-level features. To solve these problems, we propose a novel cross-modal attention enhancement network (CAE-Net), where the cross-modal complementary information is fully explored and fused and the cross-level features are effectively aggregated.

3. The Proposed Method

In this section, the architecture of our proposed cross-modal attention enhancement network (CAE-Net) is introduced in Section 3.1. The cross-modal fusion (CMF) and single-/joint-modality decoder (SMD/JMD) are described in Sections 3.2 and 3.3, respectively. We present the multi-stream fusion (MSF) in Section 3.4. The loss functions are illustrated in Section 3.5.

3.1. Architecture Overview

The architecture of the proposed cross-modal attention enhancement network (CAE-Net) is shown in Figure 2. Firstly, we use a double stream encoder to extract the multi-level features of the RGB image $\mathbf{I}^R$ and thermal infrared image $\mathbf{I}^T$, respectively. Here, we use VGG16 [41] as the backbone of the encoder, where we specially remove the last pooling layer and three fully connected layers of it. After deploying the encoder, we can obtain

five level-deep features $\{\mathbf{F}_i^R, \mathbf{F}_i^T\}_{(i=1,\cdots,5)}$ for two modal inputs, respectively, and their resolution are 1, 1/2, 1/4, 1/8, and 1/16 of the original input image, respectively. Then, we design a cross-modal fusion (CMF), which consists of a cross-attention unit (CAU) and channel attention weighted fusion, to adequately explore the cross-modal complementary information, obtaining the fused features $\{\mathbf{F}_i^F\}_{(i=3,4,5)}$. After that, we design the joint-modality decoder (JMD) to fuse the cross-level features, obtaining decoded feature $\mathbf{F}_3^{Fd}$. The JMD can effectively extract multi-scale information and filter the noisy information in the low-level features. Furthermore, we add two independent single-modality decoder (SMD) branches to preserve more modality-specific information, obtaining decoded features $\mathbf{F}_3^{Rd}$ and $\mathbf{F}_3^{Td}$, respectively. Finally, we design a multi-stream fusion (MSF) module to fully fuse the complementary information between different decoder branches, obtaining the final fused feature $\mathbf{F}_3^{Sd}$. Then, one 1×1 convolution followed by a sigmoid function is applied on $\mathbf{F}_3^{Sd}$ to generate the final saliency map $\mathbf{S}$.

3.2. Cross-Modal Fusion

Digging out complementary information between two modal features is a major problem in RGB-T SOD. Here, we design the cross-modal fusion (CMF) module shown in Figure 3 to tackle this problem. The majority of existing methods just simply concatenate or element-wise add two modal features together. However, these methods cannot avoid the performance degradation caused by the misleading information in two modal inputs (i.e., the low-quality input image and the noisy information). Hence, we employ an attention mechanism to suppress the noisy information contained in two modal features. Different from frequently used self-attention, we design the cross-attention unit (CAU-R/CAU-T) shown in Figure 3 to filter one modal feature using the attention generated from another modal feature, where it can help enhance the shared features of two modalities. Concretely, using CAU-R as an example, we separately feed the thermal features $\mathbf{F}_i^T$ into a channel attention [42] and spatial attention [43] module to produce channel attention and spatial attention values of $\mathbf{F}_i^T$, respectively. Then, we sequentially multiply the RGB features $\mathbf{F}_i^R$ with these two attention values. To avoid the RGB features being diluted by bad quality thermal samples, we introduce the residual connection for $\mathbf{F}_i^R$. Following this way, we obtain cross-attention enhanced RGB features $\mathbf{F}_i^{Re}$. Similar to CAU-R, we also deploy a CAU-T to enhance the thermal features $\mathbf{F}_i^T$. The whole process is formulated as follows,

$$\begin{cases} \mathbf{F}_i^{Re} = \mathbf{F}_i^R \oplus (\mathbf{F}_i^R \odot CA(\mathbf{F}_i^T) \odot SA(\mathbf{F}_i^T)) \\ \mathbf{F}_i^{Te} = \mathbf{F}_i^T \oplus (\mathbf{F}_i^T \odot CA(\mathbf{F}_i^R) \odot SA(\mathbf{F}_i^R)) \end{cases}, \tag{1}$$

$$\begin{cases} CA(\mathbf{F}) = \sigma(MLP(Relu(MLP(GMP_s(\mathbf{F}))))) \\ SA(\mathbf{F}) = \sigma(Conv_{7\times7}(GMP_c(\mathbf{F}))) \end{cases}, \tag{2}$$

where CA and SA are channel attention and spatial attention, respectively, GMP_s is global max pooling along the spatial dimension, GMP_c is global max pooling along the channel dimension, $Relu$ is nonlinear activation function, MLP is fully connected layer, σ is activation function, and $Conv_{7\times7}$ is convolution layer with 7×7 kernel. More details of channel attention and spatial attention can be found in [42,43]. $\odot$ is element-wise multiplication and $\oplus$ is element-wise addition. $\mathbf{F}_i^{Re}$ and $\mathbf{F}_i^{Te}$ are the enhanced RGB and thermal features, respectively.

After refining the two modal features, we attempt to appropriately fuse them. The existing methods indiscriminately fuse two modal features using concatenation or element-wise addition, but they do not take into account the different importance of two modal features. When encountering a bad quality sample, it will present a failure saliency prediction. With the help of channel attention, we can explicitly estimate the dynamic importance of RGB feature $\mathbf{F}_i^{Re}$ and thermal feature $\mathbf{F}_i^{Te}$. Concretely, we concatenate these two features along the channel dimension, and then we feed them into the channel attention module to obtain a channel-wise importance weight for indicating which modal feature is more

valuable. After that, we multiply this weight with the concatenated features, and then we employ a 1×1 convolution to reduce the channel number of concatenated features. The above calculation process is expressed as follows,

$$\mathbf{F}_i^F = Conv_{1\times1}(cat(\mathbf{F}_i^{Re}, \mathbf{F}_i^{Te}) \odot CA(cat(\mathbf{F}_i^{Re}, \mathbf{F}_i^{Te}))) \quad , \tag{3}$$

where *cat* means concatenation operation, and $Conv_{1\times1}$ means a 1×1 convolution and a BN layer [44]. $\mathbf{F}_i^F$ means the fused features of two modalities at the i-th level.

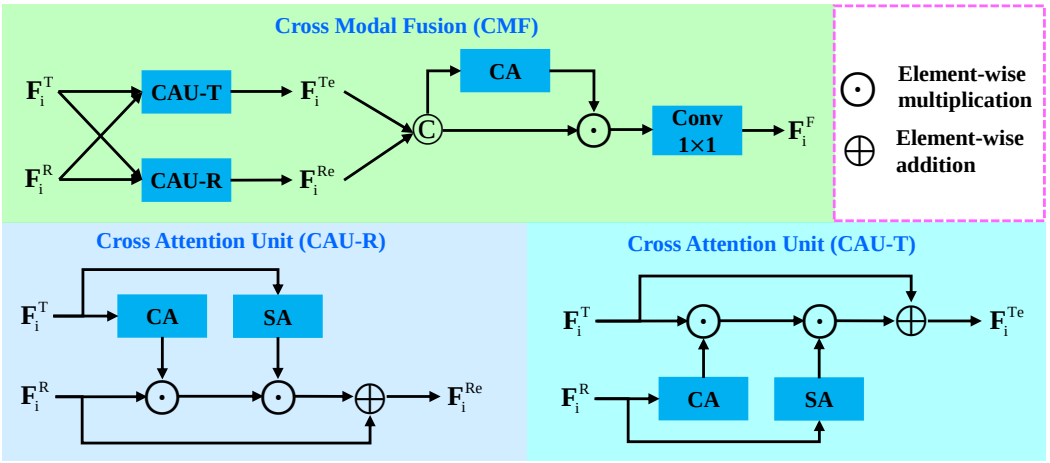

Figure 3. The architecture of the cross-modal fusion (CMF).

3.3. Single-/Joint-Modality DECODER

The Unet [7] has been widely used in SOD research. However, considering that the low-level features contain a lot of noisy information, directly concatenating low-level encoder features with decoder features is not a optimal method. Under the guidance of high-level features, we can filter the noisy information contained in low-level features. Furthermore, multi-scale modules (PPM [12], ASPP [10], and RFB [11]) have been proved to be powerful in context information extraction. Different from [8], we use the RFB in the feature decoding phase. This is because, after concatenating the encoder feature with the previous layer decoder feature, the RFB can learn a more accurate and robust feature representation. In addition, considering that only one joint-modality decoder (JMD) may put more bias on one of the two modal features, we also add two single-modality decoder (SMD) branches to preserve more specific information in two modal features. Namely, the SMD can help each modal encoder extract effective and specific information. Concretely, using the SMD shown in Figure 4 as an example, firstly, we fed the fifth level feature $\mathbf{F}_5^R$ into RFB [11] to capture global context information, thus obtaining the decoded feature $\mathbf{F}_5^{Rd}$. Then, we multiply the fourth level encoder feature $\mathbf{F}_4^R$ with $\mathbf{F}_5^{Rd}$ to filter the noisy information in the low-level feature. Next, we concatenate the filtered feature with $\mathbf{F}_5^{Rd}$ and feed it into RFB to obtain $\mathbf{F}_4^{Rd}$, which is enriched with multi-scale information. The third level decoder is similar to the above process. However, it should be noted that, in the third level feature decoding process, we also added one skip connection from $\mathbf{F}_5^{Rd}$ to avoid the high-level feature being diluted. The above calculation processes are formulated as,

$$\begin{cases} \mathbf{F}_5^{Rd} = RFB(\mathbf{F}_5^R) \\ \mathbf{F}_4^{Rd} = RFB(cat(\mathbf{F}_4^R \odot Conv_{3\times3}(UP_{\times2}(\mathbf{F}_5^{Rd})), Conv_{3\times3}(UP_{\times2}(\mathbf{F}_5^{Rd})))) \\ \mathbf{F}_3^{Rd} = RFB(cat(\mathbf{F}_3^R \odot Conv_{3\times3}(UP_{\times2}(\mathbf{F}_4^{Rd})), Conv_{3\times3}(UP_{\times2}(\mathbf{F}_4^{Rd})), Conv_{3\times3}((UP_{\times4}(\mathbf{F}_5^{Rd}))))) \end{cases} , \tag{4}$$

where RFB means the RFB module and $\mathbf{F}_i^{Rd}$ means the i-th level decoded features. $Conv_{3\times3}$ denotes a 3×3 convolution followed by a BN layer. $UP_{\times2}$ and $UP_{\times4}$ means 2 and 4 times bilinear interpolation upsampling, respectively. Our JMD is similar to SMD, but we replace the RFB operation in SMD with the context module (CM) shown in Figure 4, where we

employ two parallel branches with RFB and Nonlocal [45] operation to further enhance the global context information. Notably, before feeding the feature into the Nonlocal module, we employ a 1×1 convolution to compress the feature channel into 64 to reduce the computation cost of the Nonlocal operation. The above calculation processes are formulated as,

$$\begin{cases} \mathbf{F}_5^{Fd} = CM(\mathbf{F}_5^F) \\ \mathbf{F}_4^{Fd} = CM(cat(\mathbf{F}_4^F \odot Conv_{3\times3}(UP_{\times2}(\mathbf{F}_5^{Fd})), Conv_{3\times3}(UP_{\times2}(\mathbf{F}_5^{Fd})))) \\ \mathbf{F}_3^{Fd} = CM(cat(\mathbf{F}_3^F \odot Conv_{3\times3}(UP_{\times2}(\mathbf{F}_4^{Fd})), Conv_{3\times3}(UP_{\times2}(\mathbf{F}_4^{Fd})), Conv_{3\times3}(UP_{\times4}(\mathbf{F}_5^{Fd})))) \end{cases}, \tag{5}$$

where CM is context module shown in Figure 4, and it can be formulated as,

$$\mathbf{F}_{out}^{CM} = RFB(\mathbf{F}_{in}^{CM}) \oplus Nonlocal(Conv_{1\times1}(\mathbf{F}_{in}^{CM})), \tag{6}$$

where *Nonlocal* means Nonlocal operation.

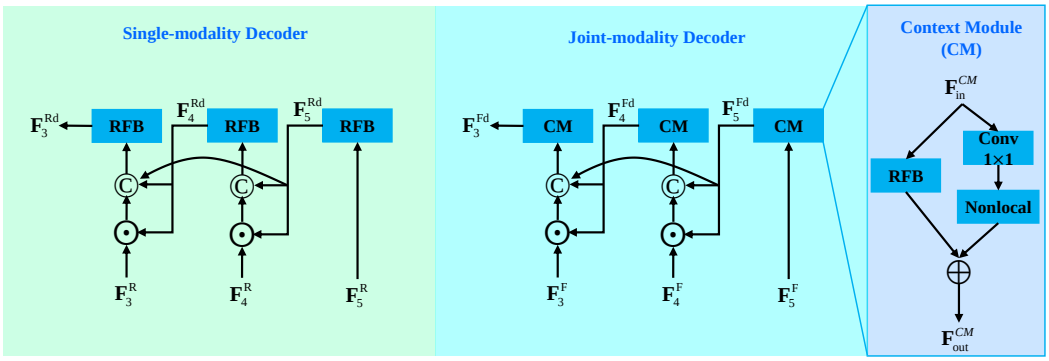

Figure 4. The architecture of the single-modality decoder (SMD) and joint-modality decoder (JMD).

3.4. Multi-Stream Fusion (MSF)

If we only use the joint-modality decoder output $\mathbf{F}_3^{Fd}$ as the final saliency results, it may lose some distinctive information contained in RGB or thermal modality. Based on this observation, we again aggregate three branches of decoded features, as shown in Figure 2. We firstly concatenate these three decoded features $\mathbf{F}_3^{Rd}$, $\mathbf{F}_3^{Td}$, and $\mathbf{F}_3^{Fd}$ together. Then, we upsample the resulting features two times and employ a 3×3 convolution to enhance the upsampling features, and we repeat this operation again, obtaining the final saliency features $\mathbf{F}_3^{Sd}$. The above calculating processes are formulated as,

$$\mathbf{F}_3^{Sd} = Conv_{3\times3}(UP_{\times2}(Conv_{3\times3}(UP_{\times2}(cat(\mathbf{F}_3^{Rd}, \mathbf{F}_3^{Td}, \mathbf{F}_3^{Fd}))))), \tag{7}$$

where $Conv_{3\times3}$ means a 3×3 convolution layer and a BN layer. Finally, we employ a 1×1 convolution toward $\mathbf{F}_3^{Sd}$, which is followed by a sigmoid function, obtaining the final saliency map $\mathbf{S}$. This process is formulated as,

$$\mathbf{S} = \sigma(Conv_{1\times1}(\mathbf{F}_3^{Sd})), \tag{8}$$

where σ is the sigmoid activation function; furthermore, we employ deep supervision [46] in our model, as shown in Figure 2, where $\mathbf{F}_3^{Rd}$, $\mathbf{F}_3^{Td}$, and $\mathbf{F}_3^{Fd}$ are also fed into a 1×1 convolutional layer followed by the sigmoid activation function to predict the saliency results, respectively. Their losses, which are marked as $\{ls_i\}_{i=2,3,4}$, are calculated between the saliency results and GT.

3.5. Loss Functions

We adopt the hybrid loss [22] to supervise our model's CAE-Net,

$$\begin{cases} ls = \ell_{bce} + \ell_{ssim} + \ell_{iou} \\ \ell_{bce} = -\frac{1}{N}\sum_{i=1}^{N}\left[\mathbf{G}(i)\log\mathbf{S}(i) + \left(1-\mathbf{G}(i)\right)\log\left(1-\mathbf{S}(i)\right)\right] \\ \ell_{ssim} = 1 - \frac{(2\mu_x\mu_y+C_1)(2\sigma_{xy}+C_2)}{(\mu_x^2+\mu_y^2+C_1)(\sigma_x^2+\sigma_y^2+C_2)} \\ \ell_{iou} = 1 - \frac{\sum_{i=1}^{N}\mathbf{G}(i)\mathbf{S}(i)}{\sum_{i=1}^{N}\left[\mathbf{G}(i)+\mathbf{S}(i)-\mathbf{G}(i)\mathbf{S}(i)\right]} \end{cases}, \tag{9}$$

where ℓ_{bce}, ℓ_{ssim}, and ℓ_{iou} are binary cross-entropy loss [47], SSIM loss [48], and IoU loss [49], respectively. $\mathbf{G}$ and $\mathbf{S}$ mean the groundtruth and saliency map, respectively. N indicates the number of total pixels in the image, i means the i-th pixel. For SSIM loss, the image is cropped m patches, and μ_x, μ_y, σ_x, and σ_y are the mean and standard deviations of GT and predictions, respectively. σ_{xy} is the covariance of them. C_1 and C_2 are set to 0.01^2 and 0.03^2 by default.

Finally, the total loss ls_{total} of the proposed CAE-Net can be defined as,

$$ls_{total} = \sum_{i=1}^{4} ls_i, \tag{10}$$

where ls_i are shown in Figure 2 and calculated using Equation (9).

4. Experiments

In this section, the datasets and implementation details are presented in Section 4.1. The evaluation metrics are described in Section 4.2. In Section 4.3, our model is quantitatively and qualitatively compared with 18 state-of-the-art models. The ablation studies are shown in Section 4.4. Finally, we analyze the scalability of our model on RGB-D datasets in Section 4.5.

4.1. Datasets and Implementation Details

To evaluate the performance of the proposed CAE-Net, we employ three widely used RGB-T datasets, including VT821 [35], VT1000 [37], and VT5000 [40]. VT821 contains 821 RGB-T image pairs. VT1000 includes 1000 RGB-T image pairs. VT5000 includes 5000 RGB-T image pairs.

For a fair comparison, we follow the setting in [16], where 2500 samples from VT5000 are chosen as the training set. The remaining datasets are treated as testing datasets. To avoid overfitting, we augmented the training datasets using random flipping.

We implement our model by using the PyTorch toolbox [50], and our PC is equipped with one RTX2080Ti GPU. We resize the input image to 224×224 before training. The encoder of RGB and thermal branches are initialized using pretrained VGG16 [41]. We train our model by using the Adam optimizer, where the initial learning rate is set to 1×10^{-4}. Additionally, the batchsize is 14, and the total training epoch is 250. We decrease the learning rate to 1×10^{-5} after 200 epochs.

4.2. Evaluation Metrics

In this paper, we compare our CAE-Net with 18 state-of-the-art models in terms of four widely used SOD metrics, including mean absolute error (MAE), F-measure (F_β) [51], E-measure (E_ξ) [52], and structure-measure (S_α) [48].

4.2.1. *MAE*

The mean absolute error (MAE) is expressed as follows,

$$MAE = \frac{1}{W \times H}\sum_{i=1}^{W}\sum_{j=1}^{H}|\mathbf{S}(i,j) - \mathbf{G}(i,j)|, \tag{11}$$

where $\mathbf{G}(i,j)$ and $\mathbf{S}(i,j)$ denotes the groundtruth and the predicted saliency map, respectively.

4.2.2. F_β

The F-measure (F_β) is a weighted harmonic mean of recall and precision, which is formulated as,

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

where β^2 is set to 0.3 referring to [51].

4.2.3. E_ξ

The E-measure (E_ξ) is a metric that evaluates global and local similarities between the groundtruth and the predicted saliency map. Concretely, it is formulated as,

$$E_\xi = \frac{1}{W \times H}\sum_{i=1}^{W}\sum_{j=1}^{H}\varphi\big(\mathbf{S}(i,j), \mathbf{G}(i,j)\big), \tag{13}$$

where φ indicates the enhanced alignment matrix.

4.2.4. S_α

Structure-measure (S_α) is employed to evaluate the structure similarities between salient objects in the groundtruth and the predicted saliency map,

$$S_\alpha = \alpha\mathbf{S}_o + (1-\alpha)\mathbf{S}_r, \tag{14}$$

where $\mathbf{S}_r$ and $\mathbf{S}_o$ mean region-aware and object-aware structural similarity, respectively, and α is set to 0.5, referring to [48].

4.3. Comparison with State-of-the-Arts

Our model is compared with 18 state-of-the-art saliency models, which are split into three groups, including the RGB, RGB-D, and RGB-T models. They are five RGB saliency models, including PoolNet [9], R3Net [53], BASNet [22], EGNet [14], and CPD [8]; five RGB-D saliency models, including DMRA [27], S2MA [30], AFNet [28], JLDCF [31], and PDNet [29]; and eight RGB-T saliency models, including M3S-NIR [36], MTMR [35], SGDL [37], FMCF [38], ADF [40], MMNet [39], MIDD [16], and CSRNet [17]. For fair comparison, the prediction maps of the RGB-T models are provided by the authors of the original paper. However, for the RGB and RGB-D models, the authors did not provide their prediction maps on RGB-T datasets, so we run the officially released codes to retrain and test them. It is worth noting that the authors [16] have tested part of the RGB and RGB-D models in this repository (https://github.com/lz118/Multi-interactive-Dual-decoder, accessed on 1 December 2022), so we directly use them. To ensure a fair comparison, we evaluate the saliency maps of all the models using the same one as the publicly available evaluation toolbox (https://github.com/lartpang/PySODMetrics, accessed on 1 December 2022).

4.3.1. Quantitative Comparison

We present PR curves and F-measure curves in Figure 5. For PR curves, our model is closest to the upright corner compared with other models. Except for VT821, our model is slightly inferior to CSRNet. For F-measure curves, our model outperforms other models on VT5000 and V1000. Namely, it locates the top position in the figure on these two datasets, but it is comparable to CSRNet on VT821. In addition, the quantitative comparison results, including MAE, F_β, E_ξ, and S_α, are presented in Table 1, where the adaptive F-measure and adaptive E-measure are reported. As can be seen from Table 1, our model outperforms most models on three datasets, except for VT821, our model ranks as second order with regard to F_β and S_α. To be specific, the traditional RGB-T methods M3S-NIR, MTMR, and SGDL perform poorly. This demonstrates the powerful representation learning ability of CNNs. Besides, our model surpasses the best RGB method CPD and RGB-D method PDNet by a

large margin. This result indicates that our carefully designed model is effective. Compared to the competitive RGB-T model CSRNet, our model advances the MAE, F_β, E_ξ, and S_α by 10.0%, 1.7%, 1.2%, and 1.4% on VT5000, respectively.

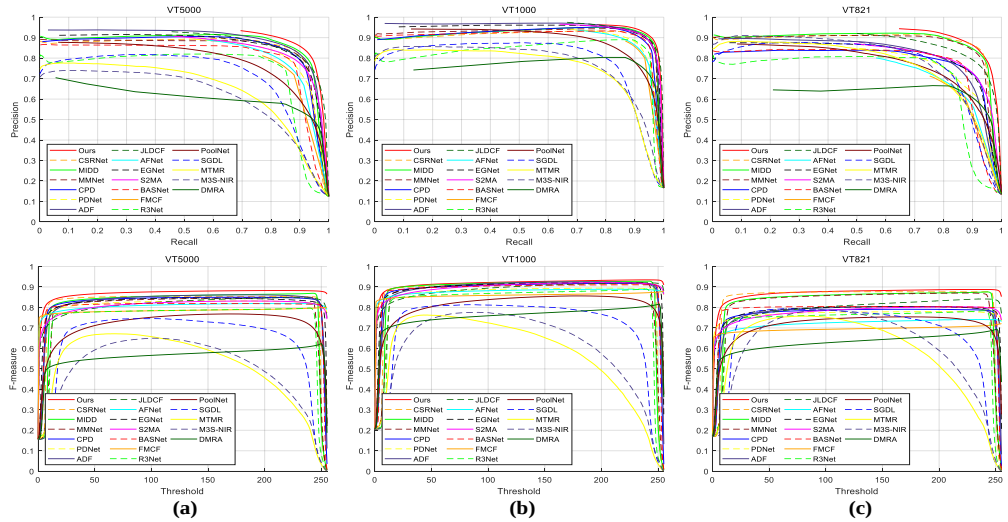

Figure 5. PR and F-measure curves of different models. (**a**) Results on the VT5000 dataset. (**b**) Results on the VT1000 dataset. (**c**) Results on the VT821 dataset.

Table 1. Quantitative comparisons with 18 models on three RGB-T datasets. The top three results are marked with red, green, and blue color in each column. ↑ and ↓ denote that the larger value is better and the smaller value is better, respectively. * denotes tradition method, and others are deep learning method.

	Dataset	**VT5000**				**VT1000**				**VT821**			
	Metric	$MAE\downarrow$	$F_\beta\uparrow$	$E_\xi\uparrow$	$S_\alpha\uparrow$	$MAE\downarrow$	$F_\beta\uparrow$	$E_\xi\uparrow$	$S_\alpha\uparrow$	$MAE\downarrow$	$F_\beta\uparrow$	$E_\xi\uparrow$	$S_\alpha\uparrow$
RGB	PoolNet [9]	0.0805	0.6431	0.8089	0.7881	0.063	0.7503	0.8552	0.8485	0.0828	0.6518	0.811	0.7884
	R3Net [53]	0.0588	0.7283	0.8618	0.8128	0.0369	0.8325	0.9191	0.8865	0.0809	0.6815	0.8165	0.7823
	BASNet [22]	0.0542	0.764	0.8793	0.8385	0.0304	0.848	0.9244	0.9084	0.0673	0.7354	0.857	0.8228
	EGNet [14]	0.0528	0.7741	0.8885	0.8526	0.0339	0.8474	0.9226	0.9093	0.0661	0.7256	0.8583	0.829
	CPD [8]	0.0465	0.7859	0.8965	0.8547	0.0312	0.8617	0.9307	0.9071	0.0795	0.7173	0.8474	0.8185
RGB-D	DMRA [27]	0.1845	0.5273	0.6869	0.6589	0.1241	0.7151	0.8197	0.7836	0.2165	0.5772	0.7144	0.6663
	S2MA [30]	0.0533	0.7432	0.8703	0.8535	0.0297	0.848	0.9286	0.9182	0.098	0.7092	0.8376	0.8112
	AFNet [28]	0.0503	0.7488	0.8794	0.8323	0.0328	0.8382	0.9226	0.8891	0.0687	0.6616	0.8212	0.7787
	JLDCF [31]	0.0503	0.7391	0.8639	0.8615	0.0299	0.8291	0.9145	0.9127	0.0756	0.7265	0.8486	0.8389
	PDNet [29]	0.0474	0.7612	0.8836	0.845	0.0327	0.8362	0.9212	0.8974	0.0566	0.7126	0.8587	0.8099
RGB-T	M3S-NIR * [36]	0.168	0.5752	0.7818	0.6527	0.1454	0.7167	0.8281	0.7263	0.1397	0.7339	0.8607	0.7238
	MTMR * [35]	0.1143	0.5952	0.7948	0.6808	0.1194	0.7136	0.8356	0.7063	0.1083	0.662	0.8142	0.7258
	SGDL * [37]	0.0886	0.6712	0.8241	0.7517	0.0896	0.7626	0.857	0.7878	0.0849	0.7292	0.8472	0.7666
	FMCF [38]	0.0556	0.7326	0.8672	0.813	0.037	0.822	0.916	0.8723	0.0808	0.6405	0.8035	0.7596
	ADF [40]	0.0483	0.7774	0.891	0.8636	0.0339	0.8462	0.9222	0.9094	0.0765	0.7158	0.8442	0.8106
	MMNet [39]	0.0433	0.7823	0.8903	0.8639	0.0275	0.8607	0.9284	0.9173	0.04	0.7958	0.8931	0.8749
	MIDD [16]	0.0433	0.7994	0.8988	0.8679	0.0271	0.88	0.942	0.9155	0.0446	0.8032	0.8975	0.8712
	CSRNet [17]	0.0417	0.8092	0.9068	0.8676	0.0242	0.8751	0.9392	0.9183	0.0376	**0.829**	0.9116	**0.8848**
	Ours	**0.0375**	**0.8233**	**0.9185**	**0.8802**	**0.0232**	**0.8813**	**0.9491**	**0.9234**	**0.0359**	0.8201	**0.9159**	0.8837

4.3.2. Complexity Analysis

In Table 2, we report the number of parameters and floating-point operations per second (FLOPs) of the compared models. We also visualize the accuracy corresponding to FLOPs in Figure 6, where the area of the circle denotes the relative size of the parameter quantities. The model located at the top-left position achieves a better trade-off between

the accuracy and model complexity. We can see that the lightweight model CSRNet has the fewest parameters and FLOPs, while ranking second in terms of the F_β score. Our model has a moderate number of parameters (38.8 M) and fewer FLOPs (47.1 G), while ranking first in terms of the F_β score. From Figure 6, we can see that our model is located at the top and the second left position. It shows that our model achieves a better trade-off between accuracy and model complexity.

Table 2. The comparisons of model complexity between different models. Here, "↓" means that the smaller the better.

Models	PoolNet [9]	R3Net [53]	BASNet [22]	EGNet [14]	CPD [8]	DMRA [27]	S2MA [30]	JLDCF [31]	ADF [40]	MIDD [16]	CSRNet [17]	Ours
Param (M) ↓	53.6	56.1	87.1	108.1	29.2	59.7	86.7	143.5	83.1	50	1	38.8
FLOPs (G) ↓	123.4	47.5	97.7	291.9	59.4	120.9	141.1	211.1	247.2	114.6	4.4	47.1

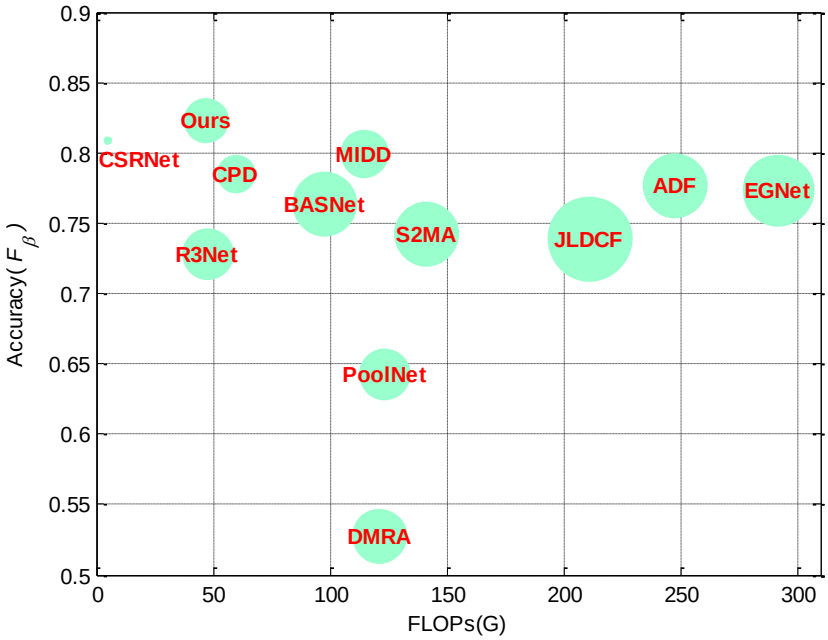

Figure 6. The accuracy and complexity of each model. The horizontal axis indicates FLOPs, while the vertical axis indicates the accuracy. Here, we measure the accuracy by F_β score on VT5000. The area of circle represents the relative size of parameter quantity of each model. The model with top-left position means the better trade-off between accuracy and FLOPs.

4.3.3. Qualitative Comparison

We show the qualitative results in Figure 7, where some representative samples, containing bad quality thermal images and small objects (the 1st row), bad quality RGB images (the 8th row), low-contrast RGB images (the 5th row), multiple objects (the 6th row and the 8th row), and vimineous object (the 10th row), are displayed. Concretely, in Figure 7 (first row and eighth row), even though the bad quality thermal image or RGB image exists, our method can highlight the salient objects without being disturbed by the bad quality sample. In the fifth row, our model can detect the bulb with the help of the thermal image, but other models are interfered by the low-contrast RGB image. In the sixth and eighth row, our model can detect two salient objects, but other models either detect only one object or detect objects with blurry boundaries. Especially in the first and sixth row, the salient objects are small, but our model can also detect them. In the 10th row, the vimineous stick can be integrally detected by our model. Generally, it can be found that, compared with other models, our model can detect small objects with less noise and can adaptively mitigate the distraction from low-quality samples.

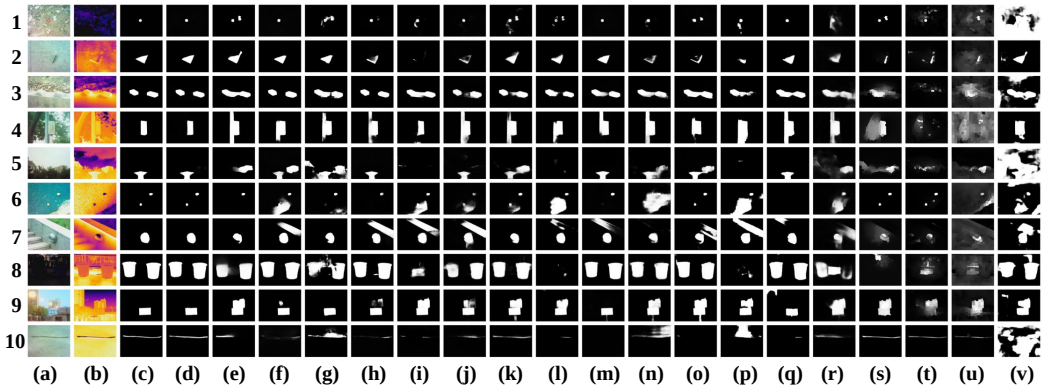

Figure 7. Visual comparison of saliency maps. (**a**) RGB image. (**b**) Thermal infrared image. (**c**) Groundtruth. (**d**) Ours. (**e**) CSRNet [17]. (**f**) MIDD [16]. (**g**) MMNet [39]. (**h**) CPD [8]. (**i**) PDNet [29]. (**j**) ADF [40]. (**k**) JLDCF [31]. (**l**) AFNet [28]. (**m**) EGNet [14]. (**n**) S2MA [30]. (**o**) BASNet [22]. (**p**) FMCF [38]. (**q**) R3Net [53]. (**r**) PoolNet [9]. (**s**) SGDL [37]. (**t**) MTMR [35]. (**u**) M3S-NIR [36]. (**v**) DMRA [27].

4.4. Ablation Studies

To demonstrate the effectiveness of each component in the proposed CAE-Net, we conduct several ablation experiments, including the effectiveness of CMF, the effectiveness of SMD/JMD, the effectiveness of MSF, the effectiveness of backbone, and the effectiveness of loss functions. We provide the quantitative results in Table 3 and the visualization results in Figures 8 and 9.

Table 3. Ablation studies are implemented on three datasets, where the best result is marked with red color in each column. Here, "↓" means that the smaller the better.

	Ablation Study	VT5000				VT1000				VT821			
		$MAE\downarrow$	$F_\beta\uparrow$	$E_\xi\uparrow$	$S_\alpha\uparrow$	$MAE\downarrow$	$F_\beta\uparrow$	$E_\xi\uparrow$	$S_\alpha\uparrow$	$MAE\downarrow$	$F_\beta\uparrow$	$E_\xi\uparrow$	$S_\alpha\uparrow$
No.1	CI	0.043	0.7955	0.9013	0.8639	0.0275	0.8647	0.9388	0.9111	0.046	0.7894	0.8884	0.8595
	w/o CMF	0.039	0.8096	0.9091	0.873	0.0259	0.8682	0.9393	0.9132	0.0393	0.7889	0.8931	0.8653
	Self	0.0385	0.8155	0.9113	0.8763	0.0252	0.8733	0.9413	0.9176	0.0381	0.7977	0.8968	0.8712
No.2	Unet	0.0445	0.7679	0.8883	0.8516	0.0319	0.8426	0.9223	0.8989	0.0484	0.7412	0.8673	0.8433
	w/o SMD	0.0386	0.8144	0.9103	0.8775	0.0233	0.8753	0.9411	0.9196	0.0387	0.8066	0.9023	0.8754
	w/o RFB	0.04	0.7995	0.9	0.8704	0.0281	0.8634	0.9295	0.9128	0.0406	0.7871	0.8882	0.8673
	w/o Nonlocal	0.0386	0.8153	0.9123	0.8735	0.0257	0.8771	0.9455	0.9178	0.0381	0.801	0.902	0.8698
No.3	Only-J	0.0382	0.8126	0.9106	0.8746	0.0246	0.8757	0.944	0.9188	0.0391	0.7971	0.8957	0.8688
	Only-R	0.0442	0.7937	0.9001	0.8601	0.0278	0.8638	0.9338	0.9088	0.0575	0.7524	0.8704	0.84
	Only-T	0.0509	0.7644	0.8904	0.8335	0.0384	0.8354	0.9238	0.8816	0.0541	0.7424	0.8699	0.8176
	Both-Avg	0.0395	0.797	0.9002	0.8731	0.0271	0.8614	0.9304	0.9134	0.0378	0.7912	0.8936	0.8731
No.4	Res50	0.0523	0.7486	0.8772	0.8361	0.0372	0.8277	0.9121	0.8866	0.0445	0.7796	0.8917	0.8501
	PS	0.039	0.8098	0.9074	0.8738	0.0243	0.8781	0.9458	0.9193	0.0373	0.7987	0.8971	0.8738
No.5	bce	0.0393	0.7918	0.8981	0.8771	0.0268	0.8553	0.9299	0.9175	0.041	0.78	0.8861	0.8719
	bce+IoU	0.0381	0.8138	0.9089	**0.8805**	0.0235	0.8737	0.9391	0.921	0.0383	0.8059	0.8992	0.8767
	bce+SSIM	0.0385	0.8041	0.9072	0.8781	0.0258	0.8595	0.9327	0.9206	0.0386	0.7884	0.8919	0.8738
	Ours	**0.0375**	**0.8233**	**0.9185**	0.8802	**0.0232**	**0.8813**	**0.9491**	**0.9234**	**0.0359**	**0.8201**	**0.9159**	**0.8837**

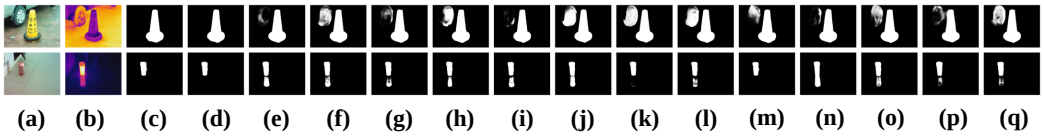

Figure 8. Visual comparison of different ablation models. (**a**) RGB image. (**b**) Thermal infrared image. (**c**) Groundtruth. (**d**) Ours. (**e**) CI. (**f**) *w/o* CMF. (**g**) Self. (**h**) Unet. (**i**) *w/o* SMD. (**j**) *w/o* RFB. (**k**) *w/o* Nonlocal. (**l**) Only-J. (**m**) Only-R. (**n**) Only-T. (**o**) Both-Avg. (**p**) Res50. (**q**) PS.

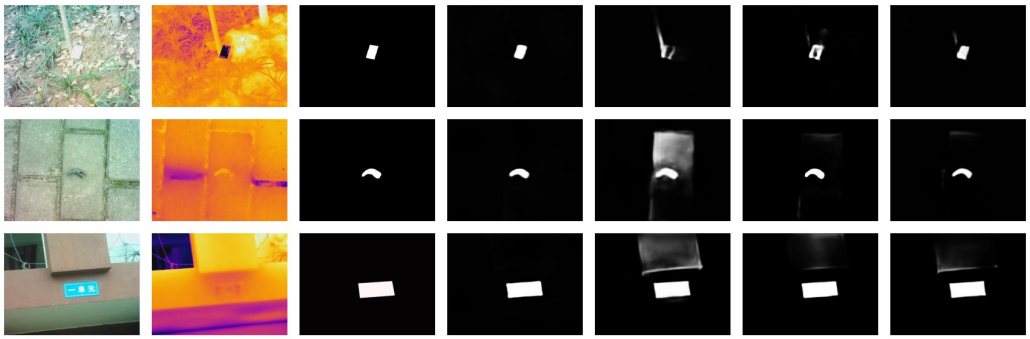

Figure 9. Visual comparison of different ablation losses. (**a**) RGB image. (**b**) Thermal infrared image. (**c**) Groundtruth. (**d**) Ours. (**e**) bce. (**f**) bce+IoU. (**g**) bce+SSIM.

4.4.1. Effectiveness of cross-Modal Fusion (CMF)

In order to verify the effectiveness of feature fusion in the middle layer, we conduct comparative experiments by concatenating two modal features at the input stage, which is abbreviated as "CI" in Table 3 (No.1). Concretely, we directly concatenate two modal input images $\mathbf{I}^R$ and $\mathbf{I}^T$ along the channel dimension at the beginning stage, and then feed it into the single branch saliency prediction network (i.e., the bottom stream in Figure 2). From Table 3 we can see that our model enhances the MAE by 12.7% on VT5000. It demonstrates the effectiveness of fusing features at the intermediate level. The visual results shown in Figure 8e also prove the same conclusion. This is because the early fusion scheme (i.e., concatenating two inputs) fails to fully explore deep complementary cues between two modal inputs. Next, we verify the effectiveness of the CMF module by removing it, shown in Table 3 (No.1 *w/o* CMF). Namely, we replace the CMF module by simply concatenating two modal features $\mathbf{F}_i^R$ and $\mathbf{F}_i^T$ together along the channel, which is followed by a 3×3 convolution layer to produce fusion features $\mathbf{F}_i^F$, and other parts are kept the same with our full model. Compared to this variant, our model elevates the MAE, F_β, E_ξ, and S_α by 3.8%, 1.6%, 1.0%, and 0.8% on VT5000, respectively. As can be seen from Figure 8f, the model "*w/o* CMF" cannot suppress the background noise. This proves that the design of the CMF is beneficial. The reason is that the CMF can suppress the noisy information in two modal features with the help of an attention module. To verify the effectiveness of cross attention in CMF, we replace it with self-attention, which is abbreviated as "Self" in Table 3 (No.1). That is, in CAU-R, we employ CA and SA of RGB feature $\mathbf{F}_i^R$ to enhance itself, but not CA and SA of thermal feature $\mathbf{F}_i^T$, and, in CAU-T, thermal feature $\mathbf{F}_i^T$ also employs attention from itself. Compared to this variant, our model elevates the MAE, F_β, E_ξ, and S_α by 2.5%, 0.9%, 0.7%, and 0.4% on VT5000, respectively. From Figure 8g, we can see that the ablation model "Self" is easily affected by background noise. This suggests that the cross-attention can highlight the shared information and suppress distracting information in another modal features.

4.4.2. Effectiveness of Single-/Joint-Modality Decoder (SMD/JMD)

To verify the effectiveness of SMD and JMD, we perform an ablation experiment by removing both of them, which is shown in Table 3 (No.2 Unet). Concretely, we use three simple Unet [7] structures to fuse the cross-level features of the three branches, respectively,

where cross-level features $\mathbf{F}_5^X$, $\mathbf{F}_4^X$, and $\mathbf{F}_3^X$ are concatenated, followed by a 3×3 convolution to fuse them layer by layer. Compared to this variant, our model can improve MAE, F_β, E_ξ, and S_α by 15.7%, 7.2%, 3.3%, and 3.3% on VT5000, respectively. As can be seen from Figure 8h, the ablation model "Unet" displays poor prediction results. This is because simple Unet cannot capture long-range context information and filter cross-level interfering information. Besides, we remove two SMDs, retaining only JMD, which is abbreviated as w/o SMD. Specifically, two single-modality decoders for the RGB branch and thermal branch are removed, only retaining one joint-modality decoder for the joint branch, so the multi-stream fusion module is also removed. The saliency maps are predicted on $\mathbf{F}_3^{Fd}$. Compared to this variant, our model elevates the MAE, F_β, E_ξ, and S_α by 2.8%, 1.0%, 0.9%, and 0.3% on VT5000, respectively. As shown in Figure 8i, it can be seen that the ablation model "w/o SMD" is easily affected by the inverted reflection of the cup in the first row. The reason is that the SMD can help two encoders extract more modality-specific information, and then the cross-modal features contain more valuable information to be fused. We further verify the effectiveness of RFB in SMD/JMD (Table 3 No.2 w/o RFB). That is, in SMD and JMD, the RFB module is replaced by a 3×3 convolution, while the Nonlocal branch in CM remains unchanged. Compared to this setting, our model enhances the MAE, F_β, E_ξ, and S_α by 6.2%, 2.9%, 2.0%, and 1.1% on VT5000, respectively. We also show the visual comparison in Figure 8j. The reason is that the RFB can effectively capture the long-range context information, which is more beneficial to depict the salient objects. Besides, we verify the effectiveness of the Nonlocal branch in the context module (No.2 w/o Nonlocal). Namely, we remove the Nonlocal branch in CM and only keep the RFB branch; at this time, the CM module is identical to the RFB. Compared to this setting, our model improves the MAE, F_β, E_ξ, and S_α by 2.8%, 0.9%, 0.6%, and 0.7% on VT5000, respectively. As can be seen from Figure 8k, the ablation model "w/o Nonlocal" is disturbed by the vehicle wheel, which is prominent in thermal image. This proves that the Nonlocal module is effective in CM because it can capture long-range relationships between different pixel positions.

4.4.3. Effectiveness of Multi-Stream Fusion (MSF)

To verify the validity of the MSF, we remove it and retrain the variant under the supervision of ls_2, ls_3, and ls_4. In this ablation model, there are three saliency outputs corresponding to features $\mathbf{F}_3^{Fd}$, $\mathbf{F}_3^{Rd}$, and $\mathbf{F}_3^{Td}$, so we evaluate their different contributions. First, we evaluate the contribution of joint-modality decoder branch (i.e., the middle stream in Figure 2), which is denoted as "Only-J" in Table 3 (No.3). That is, the saliency map is predicted on $\mathbf{F}_3^{Fd}$. Compared to this variant, our model elevates the MAE, F_β, E_ξ, and S_α by 1.8%, 1.3%, 0.8%, and 0.6% on VT5000, respectively. Second, we evaluate the contribution of RGB branch (i.e., the bottom stream in Figure 2), with the saliency map predicted on $\mathbf{F}_3^{Rd}$, which is marked as "Only-R" in Table 3. Third, we evaluate the contribution of the thermal branch (i.e., the top stream in Figure 2), and the saliency maps are predicted on $\mathbf{F}_3^{Td}$, which is marked as "Only-T" in Table 3. We can see that the RGB branch provides more contributions than the thermal branch on VT5000 with $MAE(\downarrow)$ 0.0442 vs. 0.0509. However, our model largely outperforms the single RGB branch or single thermal branch. This shows that single modal information is deficient. By fusing two modal features together (i.e., Only-J), the performance is boosted, but is still inferior to our full model. Finally, we average three saliency predictions of $\mathbf{F}_3^{Fd}$, $\mathbf{F}_3^{Rd}$, and $\mathbf{F}_3^{Td}$, which is labeled as "Both-Avg". It turns out that simply averaging the three predictions will not yield better results. However, our model with MSF can further explore the complementary relationship between three branches by fusing them at the feature level with two 3×3 convolution layers. The visual comparisons shown in Figure 8l–o also consistently prove the effectiveness of the MSF module.

4.4.4. Effectiveness of Backbone

In Table 3, (No.4), we verify the effectiveness of backbone of the encoder. Firstly, we replace VGG16 with ResNet50 [54] as backbone of the encoder for two modal inputs $\mathbf{I}^R$

and $\mathbf{I}^T$, which is abbreviated as "Res50". Compared to this variant, our model elevates the MAE, F_β, E_ξ and S_α by 28.2%, 9.9%, 4.7%, 5.2% on VT5000, respectively. From Figure 8p we can see that, the model "Res50" can only predict the inferior saliency results. This proves that our model is not compatible with ResNet50. Secondly, we share the parameters of two encoders for RGB and thermal branches, which is abbreviated as "PS". That is, the Conv1–Conv5 of the RGB branch share the same parameters as Conv1–Conv5 of the thermal branch. Compared to this variant, our model elevates the MAE, F_β, E_ξ, and S_α by 3.8%, 1.6%, 1.2%, and 0.7% on VT5000, respectively. The visual results are shown in Figure 8q. The results show that two parameter independent encoders can learn more diverse feature representations for each modality, respectively.

4.4.5. Effectiveness of Loss Functions

In Table 3, (No.5), we verify the effectiveness of loss functions. Firstly, we only use the bce loss ℓ_{bce} in the training process. Compared to this setting, our model elevates the MAE, F_β, E_ξ, and S_α by 4.5%, 3.9%, 2.2%, and 0.3% on VT5000, respectively. Secondly, we combine the bce with IoU loss. Namely, simultaneously employing ℓ_{bce} and ℓ_{iou} to train our model. Compared to only employing bce loss, this variant elevates the MAE, F_β, E_ξ, and S_α by 3.0%, 2.7%, 1.2%, and 0.3% on VT5000, respectively. Thirdly, we combine the bce with SSIM loss. Namely, simultaneously employing ℓ_{bce} and ℓ_{ssim} to train our model. Compared to only employing bce loss, this variant elevates the MAE, F_β, E_ξ, and S_α by 2.0%, 1.5%, 1.0%, and 0.1% on VT5000, respectively. Compared to bce+IoU and bce+SSIM, our model can elevate the MAE by 1.5% and 2.5%, respectively. As can be seen from Figure 9, our full model shows the superiority in all cases. The results show that either IoU or SSIM loss can help the model learn more helpful information. Furthermore, by simultaneously employing bce, IoU, and SSIM losses, our model presents the best results.

4.5. Scalability Analysis

We also verify the adaptation of our CAE-Net on four RGB-D datasets, including NJU2K (1985 image pairs) [55], NLPR (1000 image pairs) [56], STERE (1000 image pairs) [57], and DUT (1200 image pairs) [27]. Following previous work settings [58,59], 1485 images from the NJU2K dataset and 700 images from the NLPR dataset are used for training, when testing our model on NJU2K, NLPR, and STERE. Additionally, as in the widely adopted training strategy in [60,61], an additional 800 image pairs from DUT are used for training, when testing our model on DUT.

We provide the quantitative results of 10 SOTA RGB-D methods in Table 4, including JLDCF [31], DCMF [62], SSF [63], DANet [61], A2dele [60], DMRA [27], ICNet [64], S2MA [30], AFNet [28], and CPFP [65]. There are some methods, for which their codes are not available or for which the authors do not provide the saliency results, where we mark them with symbol "−" in Table 4. From the quantitative comparisons, we can see that our CAE-Net is comparable to these SOTA RGB-D methods. In general, our model ranks in the top three on most datasets, except on STERE in terms of S_α, where our model ranks fourth. Specifically, our model enhances MAE by 8.6% and 2.9% on NJU2K and STERE, respectively. These quantitative results show that our model can be successfully adapted to RGB-D datasets, demonstrating favorable generation ability of our model.

Table 4. Quantitative comparisons with 10 methods on four RGB-D datasets. The top three results are marked with red, green, blue color in each column. ↑ and ↓ denote that the larger value is better and smaller value is better, respectively. The symbol "−" denotes that their saliency results are not available.

Datasets	Metric	CPFP [65]	AFNet [28]	S2MA [30]	ICNet [64]	DMRA [27]	A2dele [60]	DANet [61]	SSF [63]	DCMF [62]	JLDCF [31]	Ours
NJU2K	$MAE\downarrow$	0.0534	0.0533	0.0533	0.052	0.051	0.0509	0.0464	0.0435	0.0427	0.0415	**0.0379**
	$F_\beta\uparrow$	0.8364	0.8672	0.8646	0.8676	0.8701	0.8709	0.8763	0.8827	0.8804	0.8841	**0.9007**
	$E_\xi\uparrow$	0.9002	0.9188	0.9163	0.9127	0.92	0.916	0.926	0.9335	0.9246	0.9347	**0.9409**
	$S_\alpha\uparrow$	0.8777	0.8801	0.8942	0.8939	0.8859	0.871	0.8969	0.8984	**0.9125**	0.9025	0.9074
NLPR	$MAE\downarrow$	0.036	0.033	0.03	0.0284	0.0315	0.0286	0.0285	0.0267	0.029	**0.0219**	0.0221
	$F_\beta\uparrow$	0.8189	0.8203	0.8479	0.865	0.8494	0.87	0.8662	0.8672	0.849	0.8732	**0.897**
	$E_\xi\uparrow$	0.9227	0.9306	0.9407	0.9435	0.94	0.9441	0.9478	0.949	0.9381	0.9539	**0.9606**
	$S_\alpha\uparrow$	0.8874	0.8994	0.9145	0.9215	0.8986	0.898	0.9137	0.9135	0.921	0.9239	**0.9241**
STERE	$MAE\downarrow$	0.0514	0.0472	0.0508	0.0447	0.0477	0.0432	0.0476	0.0448	0.0427	0.0404	**0.0392**
	$F_\beta\uparrow$	0.8296	0.8718	0.8545	0.8642	0.8658	0.8808	0.8581	0.878	0.8659	0.8688	**0.8824**
	$E_\xi\uparrow$	0.9071	0.9337	0.9254	0.9256	0.9332	0.9348	0.9263	0.9342	0.9298	0.9368	**0.9403**
	$S_\alpha\uparrow$	0.8793	0.8914	0.8904	0.9025	0.8856	0.887	0.8922	0.8928	**0.9097**	0.9029	0.8993
DUT	$MAE\downarrow$	–	–	0.044	0.0722	0.0478	0.0427	0.0467	**0.034**	0.0351	0.043	0.035
	$F_\beta\uparrow$	–	–	0.8847	0.8298	0.8831	0.8901	0.8836	0.9129	0.9057	0.8827	**0.916**
	$E_\xi\uparrow$	–	–	0.9349	0.9012	0.9301	0.9296	0.929	**0.9514**	0.9505	0.9375	0.9498
	$S_\alpha\uparrow$	–	–	0.903	0.8524	0.8889	0.8869	0.8894	0.9159	**0.9279**	0.9055	0.9141

5. Conclusions

In this paper, we propose a cross-modal attention enhancement network (CAE-Net), which consists of cross-modal fusion (CMF), a single-/joint-modality decoder (SMD/JMD), and multi-stream fusion (MSF), to accurately detect the salient objects. Firstly, we design the cross-modal fusion (CMF) to fuse cross-modal features, where a cross-attention unit (CAU) is employed to refine two modal features, and channel attention weighted fusion is used to merge two modal features. The CMF can effectively enhance features and reduce disturbance from bad quality samples. Then, we design the joint-modality decoder (JMD) to fuse cross-level features, where the low-level features are purified using high-level decoded features. The JMD effectively filter noise in low-level features and capture wider context information. Besides, we add two single-modality decoder (SMD) branches to preserve more modality-specific information. Finally, we employ multi-stream fusion (MSF) to fuse three branches of decoded features. The MSF can further aggregate effective information in three decoder branches. Extensive experiments are performed on three public datasets, and the results show that our model CAE-Net is comparable to 18 state-of-the-art saliency models.

Author Contributions: Conceptualization, C.L. and B.W.; methodology, C.L. and X.Z.; software, C.L.; validation, Y.S. and J.H.; investigation, X.Z.; resources, Y.S. and J.H.; writing—original draft preparation, C.L.; writing—review and editing, B.W. and X.Z.; supervision, X.Z. and C.Y.; project administration, J.Z. and C.Y.; funding acquisition, J.Z. and C.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Key Research and Development Program of China under Grants 2020YFB1406604; the Fundamental Research Funds for the Provincial Universities of Zhejiang under Grants GK229909299001-009; the National Natural Science Foundation of China under Grants 62271180, 62171002, 61901145, U21B2024, 61931008, 62071415, 61972123, 62001146; the "Pioneer" and "Leading Goose" R&D Program of Zhejiang Province (2022C01068); the Zhejiang Province Nature Science Foundation of China under Grants LR17F030006, LY19F030022, LZ22F020003; the Hangzhou Dianzi University (HDU) and the China Electronics Corporation DATA (CECDATA) Joint Research Center of Big Data Technologies under Grants KYH063120009; the 111 Project under

Grants D17019; and the Fundamental Research Funds for the Provincial Universities of Zhejiang under Grants GK219909299001-407.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Liu, J.; Gong, S.; Guan, W.; Li, B.; Li, H.; Liu, J. Tracking and Localization based on Multi-angle Vision for Underwater Target. *Electronics* **2020**, *9*, 1871. [CrossRef]
2. Tang, L.; Sun, K.; Huang, S.; Wang, G.; Jiang, K. Quality Assessment of View Synthesis Based on Visual Saliency and Texture Naturalness. *Electronics* **2022**, *11*, 1384. [CrossRef]
3. Ji, L.; Hu, X.; Wang, M. Saliency Preprocessing Locality-Constrained Linear Coding for Remote Sensing Scene Classification. *Electronics* **2018**, *7*, 169. [CrossRef]
4. Duan, C.; Liu, Y.; Xing, C.; Wang, Z. Infrared and Visible Image Fusion Using Truncated Huber Penalty Function Smoothing and Visual Saliency Based Threshold Optimization. *Electronics* **2022**, *11*, 33. [CrossRef]
5. LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [CrossRef]
6. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
7. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; pp. 234–241.
8. Wu, Z.; Su, L.; Huang, Q. Cascaded partial decoder for fast and accurate salient object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3907–3916.
9. Liu, J.J.; Hou, Q.; Cheng, M.M.; Feng, J.; Jiang, J. A simple pooling-based design for real-time salient object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 3917–3926.
10. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 834–848. [CrossRef] [PubMed]
11. Liu, S.; Huang, D. Receptive field block net for accurate and fast object detection. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 385–400.
12. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890.
13. Pang, Y.; Zhao, X.; Zhang, L.; Lu, H. Multi-scale interactive network for salient object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 9413–9422.
14. Zhao, J.X.; Liu, J.J.; Fan, D.P.; Cao, Y.; Yang, J.; Cheng, M.M. EGNet: Edge guidance network for salient object detection. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 8779–8788.
15. Zhou, X.; Shen, K.; Liu, Z.; Gong, C.; Zhang, J.; Yan, C. Edge-aware multiscale feature integration network for salient object detection in optical remote sensing images. *IEEE Trans. Geosci. Remote. Sens.* **2021**, *60*, 1–19. [CrossRef]
16. Tu, Z.; Li, Z.; Li, C.; Lang, Y.; Tang, J. Multi-Interactive dual-decoder for RGB-Thermal salient object detection. *IEEE Trans. Image Process.* **2021**, *30*, 5678–5691. [CrossRef] [PubMed]
17. Huo, F.; Zhu, X.; Zhang, L.; Liu, Q.; Shu, Y. Efficient Context-Guided Stacked Refinement Network for RGB-T Salient Object Detection. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *32*, 3111–3124. [CrossRef]
18. Zhou, T.; Fu, H.; Chen, G.; Zhou, Y.; Fan, D.P.; Shao, L. Specificity-preserving rgb-d saliency detection. In Proceedings of the IEEE International Conference on Computer Vision, Virtual Conference, 10 March 2021; pp. 4681–4691.
19. Itti, L.; Koch, C.; Niebur, E. A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* **1998**, *20*, 1254–1259. [CrossRef]
20. Liu, T.; Yuan, Z.; Sun, J.; Wang, J.; Zheng, N.; Tang, X.; Shum, H.Y. Learning to detect a salient object. *IEEE Trans. Pattern Anal. Mach. Intell.* **2010**, *33*, 353–367.
21. Jiang, H.; Wang, J.; Yuan, Z.; Wu, Y.; Zheng, N.; Li, S. Salient object detection: A discriminative regional feature integration approach. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Portland, OR, USA, 23–28 June 2013; pp. 2083–2090.
22. Qin, X.; Zhang, Z.; Huang, C.; Gao, C.; Dehghan, M.; Jagersand, M. Basnet: Boundary-aware salient object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 7479–7489.
23. Wan, B.; Zhou, X.; Zheng, B.; Sun, Y.; Zhang, J.; Yan, C. Deeper feature integration network for salient object detection of strip steel surface defects. *J. Electron. Imaging* **2022**, *31*, 023013. [CrossRef]

24. Xu, C.; Liu, X.; Zhao, W. Salient object detection network with center pooling and distance-weighted affinity loss function. *J. Electron. Imaging* **2022**, *31*, 023008. [CrossRef]
25. Zhou, X.; Fang, H.; Liu, Z.; Zheng, B.; Sun, Y.; Zhang, J.; Yan, C. Dense Attention-guided Cascaded Network for Salient Object Detection of Strip Steel Surface Defects. *IEEE Trans. Instrum. Meas.* **2021**, *71*, 1–14. [CrossRef]
26. Zhu, C.; Li, G.; Wang, W.; Wang, R. An innovative salient object detection using center-dark channel prior. In Proceedings of the 2017 IEEE International Conference on Computer Vision Workshops (ICCVW), Venice, Italy, 22–29 October 2017; pp. 1509–1515.
27. Piao, Y.; Ji, W.; Li, J.; Zhang, M.; Lu, H. Depth-induced multi-scale recurrent attention network for saliency detection. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 7254–7263.
28. Wang, N.; Gong, X. Adaptive fusion for RGB-D salient object detection. *IEEE Access* **2019**, *7*, 55277–55284. [CrossRef]
29. Zhu, C.; Cai, X.; Huang, K.; Li, T.H.; Li, G. PDNet: Prior-model guided depth-enhanced network for salient object detection. In Proceedings of the 2019 IEEE International Conference on Multimedia and Expo (ICME), Shanghai, China, 8–12 July 2019; pp. 199–204.
30. Liu, N.; Zhang, N.; Han, J. Learning selective self-mutual attention for RGB-D saliency detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, 13–19 June 2020; pp. 13753–13762.
31. Fu, K.; Fan, D.P.; Ji, G.P.; Zhao, Q. JL-DCF: Joint learning and densely-cooperative fusion framework for RGB-D salient object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 3052–3062.
32. Zhou, X.; Li, G.; Gong, C.; Liu, Z.; Zhang, J. Attention-guided RGBD saliency detection using appearance information. *Image Vis. Comput.* **2020**, *95*, 103888. [CrossRef]
33. Li, C.; Cong, R.; Piao, Y.; Xu, Q.; Loy, C.C. RGB-D salient object detection with cross-modality modulation and selection. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 225–241.
34. Wen, H.; Yan, C.; Zhou, X.; Cong, R.; Sun, Y.; Zheng, B.; Zhang, J.; Bao, Y.; Ding, G. Dynamic selective network for RGB-D salient object detection. *IEEE Trans. Image Process.* **2021**, *30*, 9179–9192. [CrossRef]
35. Wang, G.; Li, C.; Ma, Y.; Zheng, A.; Tang, J.; Luo, B. RGB-T saliency detection benchmark: Dataset, baselines, analysis and a novel approach. In Proceedings of the 13th Conference on Image and Graphics Technologies and Applications, Beijing, China, 8–10 April 2018; pp. 359–369.
36. Tu, Z.; Xia, T.; Li, C.; Lu, Y.; Tang, J. M3S-NIR: Multi-modal multi-scale noise-insensitive ranking for RGB-T saliency detection. In Proceedings of the 2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), San Jose, CA, USA, 28–30 March 2019; pp. 141–146.
37. Tu, Z.; Xia, T.; Li, C.; Wang, X.; Ma, Y.; Tang, J. RGB-T image saliency detection via collaborative graph learning. *IEEE Trans. Multimedia* **2019**, *22*, 160–173. [CrossRef]
38. Zhang, Q.; Huang, N.; Yao, L.; Zhang, D.; Shan, C.; Han, J. RGB-T salient object detection via fusing multi-level CNN features. *IEEE Trans. Image Process.* **2019**, *29*, 3321–3335. [CrossRef]
39. Gao, W.; Liao, G.; Ma, S.; Li, G.; Liang, Y.; Lin, W. Unified information fusion network for multi-modal RGB-D and RGB-T salient object detection. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *32*, 2091–2106. [CrossRef]
40. Tu, Z.; Ma, Y.; Li, Z.; Li, C.; Xu, J.; Liu, Y. RGBT salient object detection: A large-scale dataset and benchmark. *IEEE Trans. Multimedia* **2022**. [CrossRef]
41. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
42. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
43. Chen, L.; Zhang, H.; Xiao, J.; Nie, L.; Shao, J.; Liu, W.; Chua, T.S. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–6 July 2017; pp. 5659–5667.
44. Ioffe, S.; Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, International Machine Learning Society (IMLS), Lile, France, 6–11 July 2015; pp. 448–456.
45. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-local neural networks. In Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7794–7803.
46. Hou, Q.; Cheng, M.M.; Hu, X.; Borji, A.; Tu, Z.; Torr, P.H. Deeply supervised salient object detection with short connections. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–16 July 2017; pp. 3203–3212.
47. De Boer, P.T.; Kroese, D.P.; Mannor, S.; Rubinstein, R.Y. A tutorial on the cross-entropy method. *Ann. Oper. Res.* **2005**, *134*, 19–67. [CrossRef]
48. Fan, D.P.; Cheng, M.M.; Liu, Y.; Li, T.; Borji, A. Structure-measure: A new way to evaluate foreground maps. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 4548–4557.
49. Máttyus, G.; Luo, W.; Urtasun, R. Deeproadmapper: Extracting road topology from aerial images. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 3438–3446.

50. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. Pytorch: An imperative style, high-performance deep learning library. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 8026–8037.
51. Achanta, R.; Hemami, S.; Estrada, F.; Susstrunk, S. Frequency-tuned salient region detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 22–24 June 2009; pp. 1597–1604.
52. Fan, D.P.; Gong, C.; Cao, Y.; Ren, B.; Cheng, M.M.; Borji, A. Enhanced-alignment measure for binary foreground map evaluation. In Proceedings of the International Joint Conference on Artificial Intelligence, Stockholm, Sweden, 13–19 July 2018; pp. 698–704.
53. Deng, Z.; Hu, X.; Zhu, L.; Xu, X.; Qin, J.; Han, G.; Heng, P.A. R3net: Recurrent residual refinement network for saliency detection. In Proceedings of the 27th International Joint Conference on Artificial Intelligence, Stockholm, Sweden, 13–19 July 2018; pp. 684–690.
54. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
55. Ju, R.; Ge, L.; Geng, W.; Ren, T.; Wu, G. Depth saliency based on anisotropic center-surround difference. In Proceedings of the IEEE International Conference on Image Processing, Paris, France, 27–30 October 2014; pp. 1115–1119.
56. Peng, H.; Li, B.; Xiong, W.; Hu, W.; Ji, R. Rgbd salient object detection: A benchmark and algorithms. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; pp. 92–109.
57. Niu, Y.; Geng, Y.; Li, X.; Liu, F. Leveraging stereopsis for saliency analysis. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 454–461.
58. Fan, D.P.; Lin, Z.; Zhang, Z.; Zhu, M.; Cheng, M.M. Rethinking RGB-D salient object detection: Models, data sets, and large-scale benchmarks. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *32*, 2075–2089. [CrossRef] [PubMed]
59. Chen, S.; Fu, Y. Progressively guided alternate refinement network for RGB-D salient object detection. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 520–538.
60. Piao, Y.; Rong, Z.; Zhang, M.; Ren, W.; Lu, H. A2dele: Adaptive and attentive depth distiller for efficient RGB-D salient object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 9060–9069.
61. Zhao, X.; Zhang, L.; Pang, Y.; Lu, H.; Zhang, L. A single stream network for robust and real-time RGB-D salient object detection. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 646–662.
62. Chen, H.; Deng, Y.; Li, Y.; Hung, T.Y.; Lin, G. RGBD salient object detection via disentangled cross-modal fusion. *IEEE Trans. Image Process.* **2020**, *29*, 8407–8416. [CrossRef]
63. Zhang, M.; Ren, W.; Piao, Y.; Rong, Z.; Lu, H. Select, supplement and focus for RGB-D saliency detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 3472–3481.
64. Li, G.; Liu, Z.; Ling, H. ICNet: Information conversion network for RGB-D based salient object detection. *IEEE Trans. Image Process.* **2020**, *29*, 4873–4884. [CrossRef]
65. Zhao, J.X.; Cao, Y.; Fan, D.P.; Cheng, M.M.; Li, X.Y.; Zhang, L. Contrast prior and fluid pyramid integration for RGBD salient object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 18–22 June 2018; pp. 3927–3936.