ADNet: An Asymmetric Dual-Stream Network for RGB-T Salient Object Detection

Yaqun Fang
State Key Laboratory for Novel Software Technology, Nanjing University
Nanjing, China
fangyq@smail.nju.edu.cn

Ruichao Hou
State Key Laboratory for Novel Software Technology, Nanjing University
Nanjing, China
rc_hou@smail.nju.edu.cn

Jia Bei*
State Key Laboratory for Novel Software Technology, Nanjing University
Nanjing, China
beijia@nju.edu.cn

Tongwei Ren
State Key Laboratory for Novel Software Technology, Nanjing University
Nanjing, China
rentw@smail.nju.edu.cn

Gangshan Wu
State Key Laboratory for Novel Software Technology, Nanjing University
Nanjing, China
gswu@nju.edu.cn

ABSTRACT

RGB-Thermal salient object detection (RGB-T SOD) aims to locate salient objects in images that include both RGB and thermal information. Previous approaches often suggest designing a symmetric network structure to tackle the challenge of dealing with low-quality RGB or thermal images. However, we contend that RGB and thermal modalities possess different numbers of channels and disparities in information density. In this paper, we propose a novel asymmetric dual-stream network (ADNet). Specifically, we leverage an asymmetric backbone to extract four stages of RGB features and four stages of thermal features. To enable effective interaction among low-level features in the first two stages, we introduce the Channel-Spatial Interaction (CSI) module. In the last two stages, deep features are enhanced using the Self-Attention Enhancement (SAE) module. Experimental results on the VT5000, VT1000, and VT821 datasets attest to the superior performance of our proposed ADNet compared to state-of-the-art methods.

CCS CONCEPTS

• **Computing methodologies** → **Object detection**; **Image segmentation**; *Artificial intelligence*; *Computer vision*.

KEYWORDS

RGB-T salient object detection, multi-head self attention, cross attention, cross-modal fusion.

*Corresponding author.

MMAsia '23, December 6–8, 2023, Tainan, Taiwan

ACM ISBN 979-8-4007-0205-1/23/12…$15.00
https://doi.org/10.1145/3595916.3626440

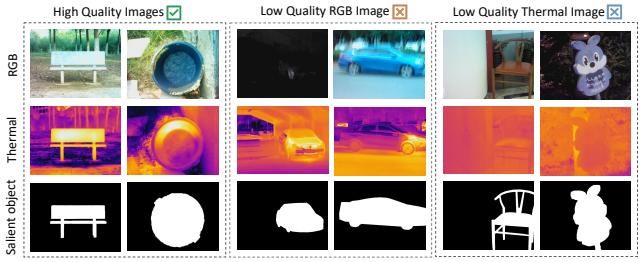

Figure 1: Examples of special cases in the RGB-T SOD dataset including low-quality RGB images caused by insufficient lighting, blurring, and low-quality thermal images caused by similar temperatures.

ACM Reference Format:
Yaqun Fang, Ruichao Hou, Jia Bei, Tongwei Ren, and Gangshan Wu. 2023. ADNet: An Asymmetric Dual-Stream Network for RGB-T Salient Object Detection. In *ACM Multimedia Asia 2023 (MMAsia '23), December 6–8, 2023, Tainan, Taiwan.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3595916.3626440

1 INTRODUCTION

RGB-Thermal salient object detection (RGB-T SOD) aims to detect and precisely segment salient objects present in visible and thermal image pairs [42]. It finds wide applications in automatic cropping [34], autonomous driving [1], and semantic segmentation [8]. Early studies primarily focused on processing RGB images [9, 27] or RGB-depth data [7, 25]. Thermal images exhibit insensitivity to lighting conditions, making them well-suited for challenging scenes, including nighttime and those with complex backgrounds [10]. Figure 1 illustrates the presence of specific instances of low-quality RGB or thermal images in the current RGB-T SOD dataset. The RGB and thermal modalities often exhibit complementary characteristics.

The initial RGB-T SOD methods predominantly rely on CNN [10, 33], but their performance is subpar. As the Transformer [31]

has evolved, the RGB-T SOD method based on it has emerged gradually [12, 15]. The Transformer-based RGB-T SOD method significantly enhances detection accuracy, yet it often exhibits a large number of computational parameters, hindering its deployment on edge or embedded devices. Consequently, the Transformer-based SOD methods should prioritize lightweight design.

Previous research in the field of RGB-T SOD acknowledges the equal significance of the thermal mode compared to the RGB modality [12, 15], and advocates the use of a dual-stream symmetrical structure to address potential damage in RGB images or corruption in thermal images. However, there exists a noticeable information density gap between the RGB and thermal modes. The thermal mode should serve as an auxiliary and complementary modality to the RGB mode. Consequently, employing symmetric networks would result in inefficient utilization of computing resources. Recent studies on the RGB-depth task have proposed asymmetric networks [13]. These networks employ a Transformer model in the RGB mode and a lightweight CNN model in the depth mode, simultaneously preserving detection effectiveness and reducing model complexity. Due to the presence of RGB image corruptions in the RGB-T dataset, such as extreme darkness or blurring, employing a lightweight CNN structure would severely impair detection performance. Consequently, achieving effective fusion becomes challenging, and in cases where RGB image information is incomplete, the thermal modes exhibit a deficiency in capturing global features, leading to inadequate detection of salient regions.

In this paper, we propose an asymmetric network ADNet based on lightweight Transformer. Initially, RGB features are extracted using an undisclosed backbone architecture inspired by Swin-B and Mobilevit. Subsequently, interaction with the low-level features is performed using the Channel-Spatial Interaction (CSI) module, which incorporates channel attention and spatial attention mechanisms. Deep features are subsequently enhanced using the Self-Attention Enhancement (SAE) module. Finally, the predicted salient objects are obtained through the decoder. Experiments show that our method preserves the performance of the results while reducing the model complexity.

In summary, we make three main contributions: (1) We introduce the first asymmetric network for RGB-T salient object detection. Experimental results demonstrate that our method achieves superior performance, reducing the number of parameters by approximately 46% and the computational load by around 40%. (2) We introduce a CSI module for low-level features, enabling the model to better leverage the CNN's capability to emphasize local features. Additionally, we present an SAE module for enhancing deep features, improving attention on salient regions by enhancing global features in both the RGB branch and the thermal branch.

2 RELATED WORK

2.1 RGB-T SOD

Due to their illumination invariance, thermal images are suitable for object detection in complex scenes with dim light and cluttered backgrounds. Early RGB-T SOD methods mainly used machine learning techniques to learn cross-modal feature representations [18, 29]. Wang *et al.* [32] created the first RGB-T SOD benchmark dataset named VT821 and proposed a graph-based multi-task manifold ranking algorithm to fuse RGB and thermal data. Tu *et al.* [30] adopted a multi-mode multi-scale manifold ranking and cooperative graph learning algorithm to achieve cross-modal salient object detection. They also built a more challenging dataset, VT1000.

With the development of deep learning, Tu *et al.* [28] contributed a large dataset VT5000 and proposed a baseline model that combines CNN and attention mechanisms. Tu *et al.* [26] proposed a dual-decoder with multi-interactions to integrate multi-stage interactions of bimodal and global context. Huo *et al.* [10] proposed a context-guided stacked refinement network to fuse the two modalities, which used light CNN as backbone. Cong *et al.* [2] introduced a global illumination estimation module to predict the global illuminance score of the image, so as to regulate the role played by the two modalities. Ma *et al.* [17] proposed a spatial complementary fusion module to explore the com plementary local regions between RGB-T images. With the development of Transformer, Liu *et al.* [15] proposed an RGB-T SOD method based on Swin Transformer. Feature extraction and feature interaction fusion are two key issues in the RGB-T SOD task.

2.2 Vision Transformer for SOD

The breakthrough progress of Transformers [31] in NLP encouraged researchers to apply them to computer vision. Dosovitskiy *et al.* [4] first proposed the ViT model based on Transformers for large-scale supervised image classification. Liu *et al.* [14] proposed Swin Transformer with shifted window operations and hierarchical design, achieving good performance on various image/video tasks. To address the high computational cost of Transformer, Sachin *et al.* [21] proposed Mobilevit, which fuses MobileNet and ViT, achieving outstanding performance on lightweight networks.

Many researchers have also applied Transformers to SOD tasks. For example, Ren *et al.* [24] and Zhu *et al.* [44] applied an encoder based purely on Transformer for single-modal SOD, while some studies [41] used a PVT [35] and CNN hybrid structure for feature extraction and salient map prediction. Some studies [16] separately modeled RGB and depth features based on a pure Transformer architecture. Liu *et al.* [15] used the Swin Transformer [14] encoder to extract features from RGB and thermal/depth images, and then achieved cross-modal fusion for SOD with edge guidance. Pang *et al.* [22] proposed a cross-modal view-mixed transformer to construct a top-down transformer-based information propagation path. Transformer-based methods achieve better performance due to their ability to capture global features, but at the cost of significantly higher parameters and computations.

3 METHOD

3.1 Overview

The framework of our proposed model is shown in Figure 2. It consists of an asymmetric dual-stream backbone as the feature extraction module, a Channel-Spatial Interaction module to fuse the low-level features, a Self-Attention Enhancement module to enhance the deep features, and finally a decoder to give the salient object prediction.

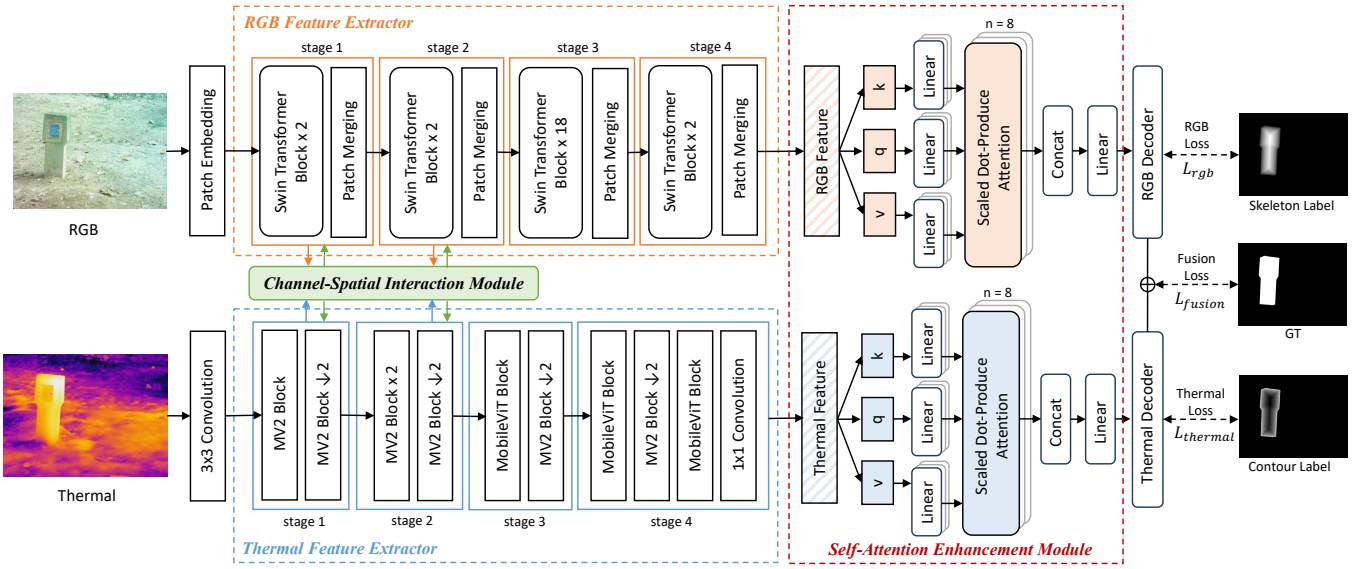

Figure 2: The framework of the proposed ADNet, including RGB feature extractor, Thermal feature extractor, Channel-Spatial Interaction module and Self-Attention Enhancement module.

Table 1: The output feature map sizes of different stages for RGB and thermal modalities, where stage 0 represents the feature map before the feature extractor.

Stage	Size of F_{rgb}	Size of $F_{thermal}$	Size of F_{fusion}
stage 0	[128, 96, 96]	[16, 192, 192]	/
stage 1	[128, 96, 96]	[64, 96, 96]	[64, 96, 96]
stage 2	[256, 48, 48]	[96, 48, 48]	[64, 48, 48]
stage 3	[512, 24, 24]	[128, 24, 24]	[64, 24, 24]
stage 4	[1024, 12, 12]	[640, 12, 12]	[64, 12, 12]

3.2 Asymmetric Dual-stream Backbone

Previous studies utilize a dual-stream backbone for extracting features from both RGB and thermal modalities. Through the analysis of feature maps at each stage, we observe that an asymmetric structure effectively attends to both local and global features. Figure 3 illustrates the feature maps of the dual-stream Swin-B in the top two rows, the feature maps of dual-stream Mobilevit in the bottom two rows, and the feature maps extracted by our proposed asymmetric dual-stream backbone in the middle two rows. The Mobilevit-based backbone predominantly emphasizes low-level features, the Swin-B-based backbone primarily focuses on high-level features, whereas our asymmetric backbone effectively captures both low-level and high-level features.

RGB Feature Extractor. We use a pretrained Swin-B as the backbone for the RGB modality, taking inputs of size 384×384. It first splits the input into a series of patches and then proceeds through four stages where the number of channels is doubled and the feature map size is halved at each stage. Through the RGB modality backbone, we can obtain a set of five feature maps with different resolutions that $F_{rgb} = \{F^i_{rgb} | i = 0, 1, 2, 3, 4\}$.

Thermal Feature Extractor. The thermal modality utilizes a pretrained Mobilevit model as its backbone, with inputs consisting of infrared images of the same size as the RGB modality. Initially, it performs a 3×3 convolution to extract local features from the image. Subsequently, following the RGB branch, the backbone is divided into four stages, with the number of channels doubling and the resolution halving at each stage, as depicted in Table 1. Through the thermal modality backbone, we also obtain a set of five feature maps with different sizes $F_t = \{F^i_t | i = 0, 1, 2, 3, 4\}$.

The features of stage 1 to stage 4 from both the RGB and thermal modalities are resized to 64 channels using a 1×1 convolution, and subsequently inputted into the CSI module.

3.3 Channel-Spatial Interaction Module

Based on our analysis in Figure 3, it is observed that CNN is better at extracting low-level features, while Transformer is better at extracting high-level features. As a result, we propose a method to combine the strengths of both structures by incorporating feature interaction attention layers between the extracted feature layers to enhance the interaction of RGB and thermal modes.

For designing the modality interaction, we employ both the channel attention mechanism and the spatial attention mechanism [38]. We take the features extracted in stages 1 to 4 of the two modalities respectively, denoting two groups of features as $F_{rgb} = \{F^i_{rgb} | i = 1, 2, 3, 4\}$ and $F_t = \{F^i_t | i = 1, 2, 3, 4\}$.

Figure 4 illustrates that, at each layer i, the feature maps of the RGB and thermal branches are initially concatenated as F^i_{fusion}:

$$F^i_{fusion} = Concat(F^i_{rgb}, F^i_t). \tag{1}$$

Next, the channel attention of the fused modality is computed:

$$CA^i = Sigmoid(MLP(P_{avg}(F^i_{fusion})) + MLP(P_{max}(F^i_{fusion}))), \tag{2}$$

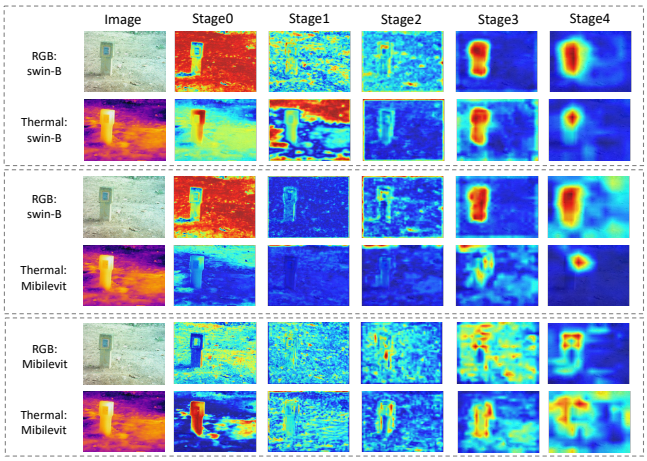

Figure 3: Visualization of input images and features maps of 4 stages using different backbones.

where CA^i means channel attention operation, $i \in \{1, 2, 3, 4\}$, $Sigmoid(\cdot)$ means denotes the sigmoid activation function, $MLP(\cdot)$ is a two-layer perceptron. $P_{avg}(\cdot)$ and $P_{max}(\cdot)$ represent the global max pooling and average pooling, respectively.

Subsequently, we calculate the spatial attention SA^irgb and SA^it for each individual modality F^irgb and F^it, respectively:

$$SA^i_{rgb} = Sigmoid(Conv^{7\times7}(Concat(P_{avg}(F^i_{rgb}), P_{max}(F^i_{rgb})))), \tag{3}$$

$$SA^i_t = Sigmoid(Conv^{7\times7}(Concat(P_{avg}(F^i_t), P_{max}(F^i_t)))), \tag{4}$$

where $Sigmoid(\cdot)$ denotes the sigmoid activation function, $Conv^{7\times7}$ means convolution operation with the kernel size 7×7, $P_{avg}(\cdot)$ and $P_{max}(\cdot)$ represent the global max pooling and average pooling.

Thirdly, we combine the channel attention and spatial attention to derive the fused attention for the two modalities using the following procedure:

$$Att^i_{rgb} = SA^i_t + Conv^{3\times3}(F^i_{fusion} \times CA^i), \tag{5}$$

$$Att^i_t = SA^i_{rgb} + Conv^{3\times3}(F^i_{fusion} \times CA^i), \tag{6}$$

where SA^i_t and SA^i_{rgb} means the spatial attention, CA^i_t and CA^i_{rgb} means the channel attention, Att^i_t and Att^i_{rgb} denotes the fused attention.

Lastly, we add Att^i_{rgb} and Att^i_t to the original feature maps, obtaining features that incorporate information from the interaction.

3.4 Self-Attention Enhancement Module

Deep features, in contrast to shallow features, exhibit a heightened focus on the semantic information within the image, enabling them to further prioritize salient regions. We propose to enhance feature maps by employing multi-head self-attention (MHSA) [31] for deep features. We define the multi-head attention mechanism $MultiHeadAtt(q, k, v)$ as follows:

$$MultiHeadAtt(q, k, v) = Concat(Head_1, ..., Head_h)\omega_o, \tag{7}$$

where q, k and v are the same feature map as input, $\{Head_1, ..., Head_h\}$ means to split the input into several heads, here we take the head as 8. ω_o represents the matrix of the final linear transformation. Then, we calculate attention $Head_i$ for each head:

$$Head_i = Attention(Q_i, K_i, V_i) = Softmax\{\frac{Q_iK_i^T}{\sqrt{d_k}}\} \cdot V_i, \tag{8}$$

where T represents the matrix transpose operation, d_k is a scaling factor, here equal to dim_{model}/num_{heads}, d_{model} is equal to the number of channels of the input feature map 64, and num_{heads} is equal to 8, Q_i, K_i and V_i is conputed by:

$$Q_i = W^i_q \cdot q, \tag{9}$$

$$K_i = W^i_k \cdot k, \tag{10}$$

$$V_i = W^i_v \cdot v, \tag{11}$$

where q, k and v are the input feature map, W^i_q, W^i_k, W^i_v correspond to linear transformation matrix of query, key and value.

In this module, we aim to use an 8-head attention mechanism for the features extracted from the input in stage4 and stage5, namely, $\{F^4_{rgb}, F^5_{rgb}\}$ and $\{F^4_t, F^5_t\}$. Specifically, we first decompose Q, K, and V into h heads, and then calculate attentions for each h. We then concatenate and project them to obtain the final output. By adding multi-head self-attention to the original feature map, we can enhance the feature of h heads, and calculate attentions of each h. Next, we concatenate and practice the final projection to obtain the final output. We can obtain an enhanced feature representation by adding multi-head self-attention to the original feature map.

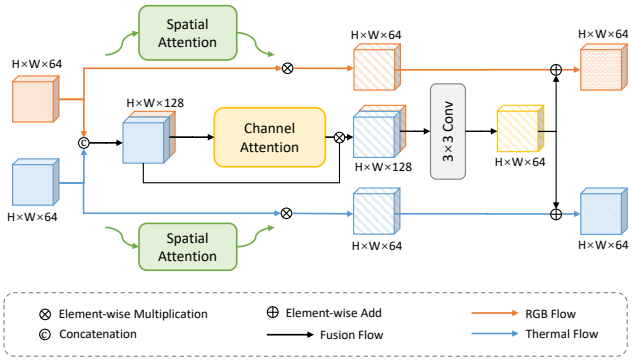

Figure 4: The design of Channel-Spatial Interaction module.

3.5 Decoder

Based on previous work [37], we adopt the residual block structure to construct a convolutional decoder. In total, the model comprises 4 residual blocks, with each block consisting of a 3×3 convolutional layer, 64 output channels, a BatchNorm layer, and a ReLU activation function. The output of each residual block is upsampled using bilinear interpolation to match the input size of the subsequent residual block.

This upsampling decoder gradually recovers high-dimensional feature maps through multi-scale feature extraction and upsampling. We send two sets of transformed feature maps $\{\overline{F}^i_{rgb} | i = 1, 2, 3, 4\}$ and $\{\overline{F}^i_t | i = 1, 2, 3, 4\}$ into two decoders, and then the output of the decoder is concatenated to form a fused result.

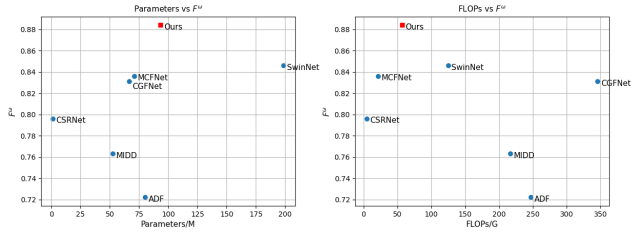

Figure 5: Comparison of Parameters, FLOPs and F^ω of our method with SOTA methods.

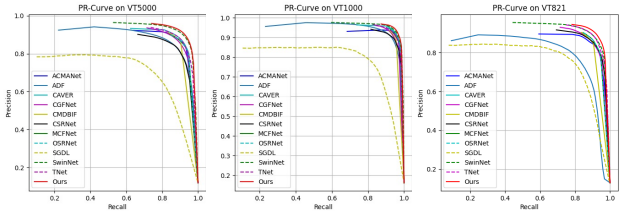

Figure 6: P-R curves comparison of different models on three RGB-T datasets.

3.6 Loss Function

The results of the RGB branch are obtained from the RGB decoder, the results of the thermal branch are obtained from the thermal decoder, and the two results are concatenated to obtain the fused result. Inspired by previous work [12, 37], we compute these three loss functions to measure the gap between the predicted results and the true results:

$$\mathcal{L}_{rgb} = \mathcal{L}_{BCE} + \mathcal{L}_{SSIM}, \tag{12}$$

$$\mathcal{L}_{thermal} = \mathcal{L}_{BCE} + \mathcal{L}_{SSIM}, \tag{13}$$

$$\mathcal{L}_{fusion} = \mathcal{L}_{BCE} + \mathcal{L}_{SSIM} + \mathcal{L}_{IoU}, \tag{14}$$

where BCE loss stands for binary cross entropy loss [3], SSIM loss represents structural similarity index measure [36], IoU loss denotes intersection over union loss [20]. The total loss function consists of three components:

$$\mathcal{L} = \mathcal{L}_{rgb} + \mathcal{L}_{thermal} + \mathcal{L}_{fusion}, \tag{15}$$

where $\mathcal{L}_{rgb}$,$\mathcal{L}_{thermal}$ and $\mathcal{L}_{fusion}$ represent RGB loss, thermal loss and fusion loss, respectively.

4 EXPERIMENT

4.1 Datasets and Metrics

Our method is evaluated on three publicly available datasets using seven commonly employed metrics.

Datasets. We evaluate our proposed method on three publicly available RGB-T SOD datasets, including VT821 [32], VT1000 [30], and VT5000 [28]. Thermal images in VT821 pose specific challenges during manual registration by forming empty regions, which we address by adding noise to some of the visible images. In VT5000, the authors labeled 11 challenging scenes based on factors such as object size, lighting conditions, center bias, number of prominent objects, and background quality. Our model is trained using 2,500 different image pairs from VT5000, while the remaining image pairs, along with VT821 and VT1000, are utilized for testing.

Metrics. We adopt widely used metrics to evaluate the performance of our model and the SOTA RGB-T SOD model. They are precision-recall (PR) curve, the mean F-measure (F_{avg}) [43], max F-measure (F_{max}) [43], weighted F-measure (F^ω) [19], mean absolute error (MAE) [23], E-measure (E_m) [6], and S-measure (S_m) [5].

4.2 Implementation Details

The RGB and thermal input images are resized to a resolution of 384×384 pixels. During the training phase, we apply various strategies to augment all the training image pairs of RGB-thermal for data augmentation. Our networks are trained using the SGD optimizer with a batch size of 8, an initial learning rate of 0.025, and a momentum of 0.9. The model is trained for 100 epochs with a learning rate decay of 5e-4. Both the training and testing of our model are conducted using PyTorch on an NVIDIA RTX 3090 GPU.

4.3 Ablation Studies

Asymmetric Backbone. As shown in Table 3, we compare the symmetric SwinTransformer based network, the symmetric Mobilevit based network and our proposed asymmetric ADNet. From the comparison results, we can see that our proposed asymmetric network performs well, with much higher performance than the symmetric Mobilevit network and even surpassing the structure of the symmetric Swin Transformer.

Feature Fusion. Table 4 demonstrates that both interaction and augmentation operations on features yield improvements, with the combined interaction and augmentation yielding the most favorable results. These findings validate the effectiveness of our proposed feature manipulation and highlight the crucial role played by information sharing between RGB and thermal modes in enhancing the effectiveness of saliency detection.

Model Complexity. In addition, we explore the complexity of the model, as shown in Table 5. We measure the complexity of the model by counting the number of parameters and the number of computations. According to the experimental results, the introduction of Mobilevit can effectively reduce the model complexity, and the feature interaction module and feature enhancement model do not introduce too many parameters and computations.

4.4 Comparison with State-of-the-Arts

To demonstrate the effectiveness of the proposed method, 14 state-of-the-art SOD methods are introduced to compare as follows: M3S-NIR [29], MTMR [32], SGDL [30], ADF [32], MIDD [26], CSRNet [10], OSRNet [11], TNet [2], MCFNet [17], CGFNet [33], CAVER [22], ACMANet [40], SwinNet [15], CMDBIF [39]. All models are trained on the VT5000 training set (2,500 images).

Quantitative Evaluation. We use the same evaluation toolkit to evaluate the prediction results of all methods as shown in Table 2, with the best metrics in bold. Figure 5 presents a comparison of Parameters, FLOPs and F^ω of our method with SOTA methods and Figure 6 is a comparison of precision-recall (PR) curves between our method and other approaches. It is evident that our proposed method surpasses the current state-of-the-art.

Table 2: Performance comparison with SOTA methods on VT5000, VT1000 and VT821, the best results are highlighted in Bold.

	VT5000						VT1000						VT821					
Methods	$F_{avg}\uparrow$	$F_{max}\uparrow$	$F^{\omega}\uparrow$	$MAE\downarrow$	$E_m\uparrow$	$S_m\uparrow$	$F_{avg}\uparrow$	$F_{max}\uparrow$	$F^{\omega}\uparrow$	$MAE\downarrow$	$E_m\uparrow$	$S_m\uparrow$	$F_{avg}\uparrow$	$F_{max}\uparrow$	$F^{\omega}\uparrow$	$MAE\downarrow$	$E_m\uparrow$	$S_m\uparrow$
M3S-NIR [29]	0.575	0.644	0.327	0.168	0.780	0.652	0.717	0.769	0.463	0.145	0.827	0.726	0.734	0.780	0.407	0.140	0.859	0.723
MTMR [32]	0.595	0.662	0.397	0.114	0.795	0.680	0.715	0.755	0.485	0.119	0.836	0.706	0.662	0.747	0.462	0.108	0.815	0.725
SGDL [30]	0.672	0.737	0.558	0.089	0.824	0.750	0.764	0.807	0.652	0.090	0.856	0.787	0.731	0.780	0.583	0.085	0.846	0.764
ADF [32]	0.778	0.863	0.722	0.048	0.891	0.864	0.847	0.923	0.804	0.034	0.921	0.910	0.717	0.804	0.627	0.077	0.843	0.810
MIDD [26]	0.801	0.871	0.763	0.043	0.897	0.867	0.882	0.926	0.856	0.027	0.933	0.915	0.805	0.874	0.760	0.045	0.895	0.871
CSRNet [10]	0.811	0.857	0.796	0.042	0.905	0.868	0.877	0.918	0.878	0.024	0.925	0.918	0.831	0.88	0.821	0.038	0.909	0.885
OSRNet [11]	0.823	0.866	0.807	0.040	0.908	0.875	0.892	0.929	0.891	0.022	0.935	0.926	0.814	0.862	0.801	0.043	0.896	0.875
TNet [2]	0.846	0.895	0.84	0.033	0.927	0.895	0.889	0.937	0.895	0.021	0.937	0.929	0.842	0.904	0.841	0.03	0.919	0.899
MCFNet [17]	0.848	0.886	0.836	0.033	0.924	0.887	0.902	0.939	0.906	0.019	0.944	0.932	0.844	0.889	0.835	0.029	0.918	0.891
CGFNet [33]	0.851	0.887	0.831	0.035	0.922	0.883	0.906	0.936	0.900	0.023	0.944	0.923	0.845	0.885	0.829	0.038	0.912	0.881
CAVER [22]	0.856	0.897	0.849	0.028	0.935	0.899	0.906	0.945	0.912	0.016	0.949	0.938	0.854	0.897	0.846	0.026	0.928	0.897
ACMANet [40]	0.858	0.89	0.823	0.033	0.932	0.887	0.904	0.933	0.889	0.021	0.945	0.927	0.837	0.873	0.807	0.035	0.914	0.883
SwinNet [15]	0.865	0.915	0.846	0.026	0.942	0.912	0.896	0.948	0.894	0.018	0.947	0.938	0.847	0.903	0.818	0.03	0.926	0.904
CMDBIF [39]	0.868	0.892	0.846	0.032	0.933	0.886	0.914	0.931	0.909	0.019	0.952	0.927	0.856	0.887	0.837	0.032	0.923	0.882
Ours	**0.893**	**0.924**	**0.884**	**0.022**	**0.953**	**0.922**	**0.916**	**0.952**	**0.920**	**0.015**	**0.952**	**0.944**	**0.869**	**0.915**	**0.860**	**0.024**	**0.930**	**0.915**

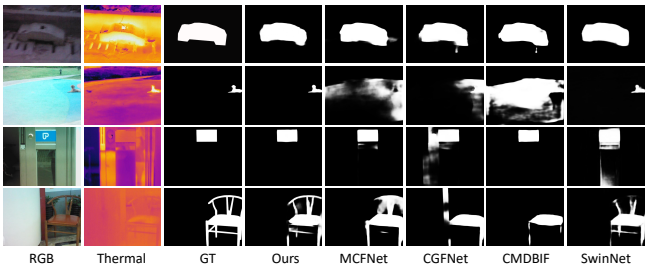

Figure 7: Visual comparison with SOTA RGB-T methods.

Qualitative Evaluation. The visualization results of our method are shown in Figure 7. The upper two rows are examples of poor performance of RGB images (blurry, dim or too small salient area), and the lower two rows are examples of poor performance of thermal images (indistinguishable from the background). It can be seen that our method can achieve better detection results in both cases and is more robust.

5 CONCLUSION

In this paper, we introduced ADNet, an asymmetric network designed for the RGBT-SOD task. An asymmetric dual-stream backbone was utilized to extract both RGB and thermal features. Additionally, the CSI module and SAE module were introduced to enhance cross-modal interactions. Experimental results demonstrated that our ADNet achieved excellent performance on public datasets, validating the effectiveness of our proposed asymmetric network. Moreover, our method used lower model parameters and computational costs.

6 ACKNOWLEDGMENT

This work is supported by the National Science Foundation of China (62072232), Key R&D Project of Jiangsu Province (BE2022138), the Fundamental Research Funds for the Central Universities (021714380026), the Program B for Outstanding Ph.D. candidate of Nanjing University, and the Collaborative Innovation Center of Novel Software Technology and Industrialization.

Table 3: Ablation results of different backbone on VT5000 test set, *S* represents Swin-B, *M* represents Mobilevit.

Backbone	$F_{avg}\uparrow$	$F_{max}\uparrow$	$F^{\omega}\uparrow$	$MAE\downarrow$	$E_m\uparrow$	$S_m\uparrow$
S + S	0.878	0.915	0.873	0.024	0.947	0.914
M + M	0.766	0.836	0.730	0.049	0.883	0.839
S + M	**0.879**	**0.918**	**0.877**	**0.023**	**0.950**	**0.917**

Table 4: Ablation results of different feature fusion. CSI represents feature interaction module, SAE represents multi-head self-attention module. The results are test on VT5000.

Fusion	$F_{avg}\uparrow$	$F_{max}\uparrow$	$F^{\omega}\uparrow$	$MAE\downarrow$	$E_m\uparrow$	$S_m\uparrow$
CSI	0.887	0.918	0.879	0.023	0.949	0.917
SAE	0.889	0.920	0.878	0.023	0.950	0.918
CSI + SAE	**0.893**	**0.924**	**0.884**	**0.022**	**0.953**	**0.922**

Table 5: Ablation of model complexity. *Paramters* denotes the number of parameters of the model while *FLOPs* represents the cost of computation.

Methods	$Paramters(M)\downarrow$	$FLOPs(G)\downarrow$
S+S	174.600	94.795
M+M	11.107	13.396
S+M	92.854	54.095
S+M+CSI	93.302	56.659
S+M+CSI+SAE	93.319	56.683

REFERENCES

[1] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. 2020. nuscenes: A multimodal dataset for autonomous driving. In *IEEE/CVF conference on computer vision and pattern recognition*. 11621–11631.

[2] Runmin Cong, Kepu Zhang, Chen Zhang, Feng Zheng, Yao Zhao, Qingming Huang, and Sam Kwong. 2022. Does thermal really always matter for RGB-T salient object detection? *IEEE Transactions on Multimedia* (2022).

[3] Pieter-Tjerk De Boer, Dirk P Kroese, Shie Mannor, and Reuven Y Rubinstein. 2005. A tutorial on the cross-entropy method. *Annals of operations research* 134 (2005), 19–67.

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

[5] Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. 2017. Structure-measure: A new way to evaluate foreground maps. In *IEEE international conference on computer vision*. 4548–4557.

[6] Deng-Ping Fan, Cheng Gong, Yang Cao, Bo Ren, Ming-Ming Cheng, and Ali Borji. 2018. Enhanced-alignment measure for binary foreground map evaluation. *arXiv preprint arXiv:1805.10421* (2018).

[7] Jingfan Guo, Tongwei Ren, and Jia Bei. 2016. Salient object detection for RGB-D image via saliency evolution. In *IEEE International Conference on Multimedia and Expo*. 1–6.

[8] Yanming Guo, Yu Liu, Theodoros Georgiou, and Michael S. Lew. 2018. A review of semantic segmentation using deep neural networks. *International journal of Multimedia Information Retrieval* 7 (2018), 87–93.

[9] Xiaowei Hu, Chi-Wing Fu, Lei Zhu, Tianyu Wang, and Pheng-Ann Heng. 2020. SAC-Net: Spatial attenuation context for salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology* 31, 3 (2020), 1079–1090.

[10] Fushuo Huo, Xuegui Zhu, Lei Zhang, Qifeng Liu, and Yu Shu. 2021. Efficient context-guided stacked refinement network for RGB-T salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology* 32, 5 (2021), 3111–3124.

[11] Fushuo Huo, Xuegui Zhu, Qian Zhang, Ziming Liu, and Wenchao Yu. 2022. Real-time one-stream semantic-guided refinement network for RGB-thermal salient object detection. *IEEE Transactions on Instrumentation and Measurement* 71 (2022), 1–12.

[12] Xiurong Jiang, Lin Zhu, Yifan Hou, and Hui Tian. 2022. Mirror complementary transformer network for RGB-thermal salient object detection. *arXiv preprint arXiv:2207.03558* (2022).

[13] Chang Liu, Gang Yang, Shuo Wang, Hangxu Wang, Yunhua Zhang, and Yutao Wang. 2023. TANet: Transformer-based asymmetric network for RGB-D salient object detection. *IET Computer Vision* (2023).

[14] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *IEEE/CVF international conference on computer vision*. 10012–10022.

[15] Zhengyi Liu, Yacheng Tan, Qian He, and Yun Xiao. 2021. SwinNet: Swin transformer drives edge-aware RGB-D and RGB-T salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology* 32, 7 (2021), 4486–4497.

[16] Zhengyi Liu, Yuan Wang, Zhengzheng Tu, Yun Xiao, and Bin Tang. 2021. TriTransNet: RGB-D salient object detection with a triplet transformer embedding network. In *29th ACM international conference on multimedia*. 4481–4490.

[17] Shuai Ma, Kechen Song, Hongwen Dong, Hongkun Tian, and Yunhui Yan. 2023. Modal complementary fusion network for RGB-T salient object detection. *Applied Intelligence* 53, 8 (2023), 9038–9055.

[18] Yunpeng Ma, Dengdi Sun, Qianqian Meng, Zhuanlian Ding, and Chenglong Li. 2017. Learning multiscale deep features and SVM regressors for adaptive RGB-T saliency detection. In *2017 10th International Symposium on Computational Intelligence and Design (ISCID)*, Vol. 1. 389–392.

[19] Ran Margolin, Lihi Zelnik-Manor, and Ayellet Tal. 2014. How to evaluate foreground maps?. In *IEEE conference on computer vision and pattern recognition*. 248–255.

[20] Gellért Máttyus, Wenjie Luo, and Raquel Urtasun. 2017. Deeproadmapper: Extracting road topology from aerial images. In *IEEE international conference on computer vision*. 3438–3446.

[21] Sachin Mehta and Mohammad Rastegari. 2021. Mobilevit: light-weight, general-purpose, and mobile-friendly vision transformer. *arXiv preprint arXiv:2110.02178* (2021).

[22] Youwei Pang, Xiaoqi Zhao, Lihe Zhang, and Huchuan Lu. 2023. CAVER: Cross-modal view-mixed transformer for bi-modal salient object detection. *IEEE Transactions on Image Processing* 32 (2023), 892–904.

[23] Federico Perazzi, Philipp Krähenbühl, Yael Pritch, and Alexander Hornung. 2012. Saliency filters: Contrast based filtering for salient region detection. In *IEEE conference on computer vision and pattern recognition*. 733–740.

[24] Sucheng Ren, Qiang Wen, Nanxuan Zhao, Guoqiang Han, and Shengfeng He. 2021. Unifying global-local representations in salient object detection with transformer. *arXiv preprint arXiv:2108.02759* (2021).

[25] Tongwei Ren and Ao Zhang. 2019. *RGB-D Salient Object Detection: A Review*. Springer International Publishing, Cham, 203–220.

[26] Zhengzheng Tu, Zhun Li, Chenglong Li, Yang Lang, and Jin Tang. 2021. Multi-interactive dual-decoder for RGB-thermal salient object detection. *IEEE Transactions on Image Processing* 30 (2021), 5678–5691.

[27] Zhengzheng Tu, Yan Ma, Chenglong Li, Jin Tang, and Bin Luo. 2020. Edge-guided non-local fully convolutional network for salient object detection. *IEEE transactions on circuits and systems for video technology* 31, 2 (2020), 582–593.

[28] Zhengzheng Tu, Yan Ma, Zhun Li, Chenglong Li, Jieming Xu, and Yongtao Liu. 2022. RGBT salient object detection: A large-scale dataset and benchmark. *IEEE Transactions on Multimedia* (2022).

[29] Zhengzheng Tu, Tian Xia, Chenglong Li, Yijuan Lu, and Jin Tang. 2019. M3S-NIR: Multi-modal multi-scale noise-insensitive ranking for RGB-T saliency detection. In *2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*. 141–146.

[30] Zhengzheng Tu, Tian Xia, Chenglong Li, Xiaoxiao Wang, Yan Ma, and Jin Tang. 2019. RGB-T image saliency detection via collaborative graph learning. *IEEE Transactions on Multimedia* 22, 1 (2019), 160–173.

[31] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[32] Guizhao Wang, Chenglong Li, Yunpeng Ma, Aihua Zheng, Jin Tang, and Bin Luo. 2018. RGB-T saliency detection benchmark: Dataset, baselines, analysis and a novel approach. In *Image and Graphics Technologies and Applications: 13th Conference on Image and Graphics Technologies and Applications, IGTA 2018, Beijing, China, April 8–10, 2018, Revised Selected Papers 13*. 359–369.

[33] Jie Wang, Kechen Song, Yanqi Bao, Liming Huang, and Yunhui Yan. 2021. CGFNet: Cross-guided fusion network for RGB-T salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology* 32, 5 (2021), 2949–2961.

[34] Wenguan Wang, Jianbing Shen, and Haibin Ling. 2018. A deep network solution for attention and aesthetics aware photo cropping. *IEEE transactions on pattern analysis and machine intelligence* 41, 7 (2018), 1531–1544.

[35] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. 2021. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *IEEE/CVF international conference on computer vision*. 568–578.

[36] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. 2003. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, Vol. 2. 1398–1402.

[37] Jun Wei, Shuhui Wang, Zhe Wu, Chi Su, Qingming Huang, and Qi Tian. 2020. Label decoupling framework for salient object detection. In *IEEE/CVF conference on computer vision and pattern recognition*. 13025–13034.

[38] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. 2018. Cbam: Convolutional block attention module. In *European conference on computer vision*. 3–19.

[39] Zhengxuan Xie, Feng Shao, Gang Chen, Hangwei Chen, Qiuping Jiang, Xiangchao Meng, and Yo-Sung Ho. 2023. Cross-modality double bidirectional interaction and fusion network for RGB-T salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology* (2023).

[40] Chang Xu, Qingwu Li, Qingkai Zhou, Xiongbiao Jiang, Dabing Yu, and Yaqin Zhou. 2022. Asymmetric cross-modal activation network for RGB-T salient object detection. *Knowledge-Based Systems* 258 (2022), 110047.

[41] Jin Zhang, Yanjiao Shi, Qing Zhang, Liu Cui, Ying Chen, and Yugen Yi. 2022. Attention guided contextual feature fusion network for salient object detection. *Image and Vision Computing* 117 (2022), 104337.

[42] Qiang Zhang, Nianchang Huang, Lin Yao, Dingwen Zhang, Caifeng Shan, and Jungong Han. 2019. RGB-T salient object detection via fusing multi-level CNN features. *IEEE Transactions on Image Processing* 29 (2019), 3321–3335.

[43] Qiang Zhang, Tonglin Xiao, Nianchang Huang, Dingwen Zhang, and Jungong Han. 2020. Revisiting feature fusion for RGB-T salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology* 31, 5 (2020), 1804–1818.

[44] Heqin Zhu, Xu Sun, Yuexiang Li, Kai Ma, S Kevin Zhou, and Yefeng Zheng. 2022. DFTR: Depth-supervised fusion transformer for salient object detection. *arXiv preprint arXiv:2203.06429* (2022).