西北工業大學

本科毕业设计（论文）

题 目 Multi-Modal Fusion Network for RGB-T Salient Object Detection

专业名称 Computer Science and Technology

学生姓名______________________________

指导教师______________________________

毕业时间______June 2023________________

Multi-Modal Fusion Network for RGB-T Salient Object Detection

A Thesis Submitted

To Faculty of Computer Science of

North-Western Polytechnical University

In Partial Fulfillment of the Requirements

For the Degree of Bachelor of Science

BY

MUSA BAKARR

Advisor:

Abstract

The emergence of salient object detection using deep learning techniques has gathered significant attention due to the possibility of it being able to be applied in computer vision tasks, autonomous driving and human-computer interaction. In this paper, we propose a deep learning network for the multi-modal fusion of RGB-T modalities for salient object detection called cross-attention deep fusion network (CADFNet). In our model, we employ a two-stream multi-modal encoder serving as our backbone network for extracting special multi-modal features. The output is then passed on to a top-down parallel decoder that contains multiple receptive field blocks to learn and predict multi-modal features. After that, a cross-attention complementarity exploration module is proposed to enrich, enhance and refine multi-modal features by exploiting the complementarity between them and finally, a deep supervision progressive fusion module is proposed to fuse these multi-modal features for accurate SOD. In the end, we compare our model to other state of the art methods to see how our method performed. The training and testing were conducted on the 3 benchmark RGB-T datasets; VT5000, VT1000 & VT821. Extensive experiments done on these three benchmark datasets demonstrate that our model is an effective RGB-T SOD framework that outperforms the current state-of-the-art models, both quantitatively and qualitatively.

KEYWORDS: deep learning, convolutional neural networks, salient object detection, multi-modal fusion

Table of Contents

Chapter One Introduction

1.1 RGB-T Salient Object Detection

Salient object detection, plays a crucial role in various applications such as image understanding, object recognition, and scene analysis. Its primary objective is to identify and highlight the objects or regions in an image that draw significant attention from human observers. By pinpointing the visually prominent elements, salient object detection facilitates higher-level understanding of scenes and enables efficient analysis of visual information.

The emergence of RGB-T salient object detection further extends the concept by incorporating thermal information into the detection process. This integration enables the identification of objects that not only stand out based on their visual appearance but also exhibit distinctive thermal signatures. By exploiting both RGB and thermal modalities, RGB-T salient object detection methods gain a more comprehensive understanding of the scene, capturing the most visually and thermally salient regions.

The inclusion of thermal information in RGB-T salient object detection brings several advantages. Firstly, thermal data provides valuable insights into the temperature distribution of objects, offering a complementary perspective beyond the visual appearance. This thermal signature can be particularly useful in scenarios with low lighting conditions or when objects are obscured by visual camouflage, as the thermal characteristics remain distinct and discernible. Secondly, the integration of RGB and thermal modalities facilitates a more robust and accurate detection process by capturing different aspects of saliency. The fusion of visual and thermal cues enhances the overall discriminative power, resulting in more reliable identification and localization of salient objects.

1.2 Multi-modal Fusion Networks for RGB-T Salient Object Detection

Building upon the concept of RGB-T salient object detection, the fusion of RGB and thermal modalities is a crucial step in achieving accurate and reliable results. This fusion process combines the complementary strengths of both modalities, leveraging their unique characteristics to effectively highlight regions that exhibit significant visual and thermal saliency.

To effectively leverage both RGB and thermal modalities, RGB-T salient object detection methods often employ sophisticated fusion strategies and fusion networks. These approaches aim to extract and combine the complementary information from each modality, capitalizing on the strengths of both visual and thermal data.

Deep learning techniques, such as convolutional neural networks (CNNs), have proven to be highly effective in capturing intricate patterns and features from RGB and thermal images. Researchers have proposed various fusion network architectures to integrate RGB and thermal modalities in the context of salient object detection.

1.3 Research Content

In this paper, a deeply supervised fusion module-based model is used to progressively integrate multi-modal features for predicting saliency maps and to obtain an accurate SOD. But before fusion, a cross-attention Complementarity Exploration module is introduced to enrich high-level and low-level features by exploiting the complementarity between them.

The main contribution of the work is as follows:

- We propose a model that makes use of a multi-band decomposition framework which extracts and predicts multi-modal features from RGB-T images from a backbone network.

- A cross-Attention Complementarity Exploration module is then introduced to enhance the high-level and low-level features by exploiting the complementary information between them with a cross-attention mechanism.
- A deeply supervised progressive fusion which includes two main components is then proposed to carry out progressive fusion of the multi-modal feature with deep supervision.
- Our model is evaluated on 3 different RGB-T benchmark datasets and would be evaluated with other state of the art methods.

Chapter Two Related Work

2.1 Salient Object Detection

The first groundbreaking work in the field was presented by Zhang, Ma, and Jiang. Their work, titled "Salient object detection in RGB-D images with saliency evolution," (Zhang, Ma & Jiang, 2014, p. 249) introduced a fusion strategy that combined depth and color cues. By considering both global and local saliency cues, their method achieved accurate detection of salient objects.

The goal of the work was to accurately detect salient objects in RGB-D images by effectively integrating RGB color and depth information. The authors introduce a method called "saliency evolution" to achieve this.

The saliency evolution approach combines global and local saliency cues to detect salient objects. It takes advantage of the depth information to refine and evolve the saliency maps generated from the RGB image. The depth cues help in improving the accuracy of saliency detection by considering the spatial relationships and depth discontinuities of objects in the scene.

Building upon this, Li, Hu, Zhang, and Zhang (Li, Hu & Zhang, 2016, p. 2809) proposed a joint RGB and depth saliency estimation approach. In their work, titled "Joint RGB and depth saliency estimation via a unified multi-scale deep network," they developed a unified multi-scale deep learning network. This network effectively integrated RGB and depth information, capturing salient object features across different scales and enhancing detection performance. The method involves several key steps. First, the RGB and depth images are fed into the network, which consists of multiple layers of convolutional neural networks (CNNs). The network is designed to process the RGB and depth information simultaneously and extract hierarchical features at different scales.

By incorporating multi-scale processing, the network can capture salient object features at various levels of detail. This multi-scale approach allows for the detection of objects of different sizes and scales in the scene.

To further refine the saliency maps, the network employs a joint training strategy that optimizes the model parameters using both RGB and depth data. This joint training process ensures that the network learns to effectively fuse and utilize both modalities for salient object detection.

To address the challenges of salient object detection in RGB-D scenes, Wang, Liu, Huang, and Wang introduced a method based on depth-weighted information entropy. Their work, titled "Salient object detection in RGB-D scenes using depth-weighted information entropy,"(Wang, Liu, Huang & Wang, 2016, p. 2082) utilized information entropy to measure the saliency of RGB and depth information. By assigning different weights to depth cues based on their information, their method achieved robust detection of salient objects.

Subsequently, motivated by and building upon these groundbreaking researches, the first work on RGB-T Salient Object detection (Wang et al., 2018, pp. 359-369) was introduced. It proposes a multi-task manifold ranking algorithm for RGB-T image saliency detection which made use of a unified RGB-T image dataset that includes 821 spatially aligned RGB-T image pairs. The research in this direction was limited by the lack of a comprehensive benchmark.

The second important work on RGB-T salient object detection (Tu et al., 2019, pp. 160–173) proposes an approach that relies on a collaborative graph learning algorithm. In this work, super pixels are taken as graph nodes, and it uses hierarchical deep features to simultaneously learn graph affinity and node saliency in a unified optimization framework. This work was done on a more challenging dataset, which contains 1000 spatially aligned RGB-T image pairs and their ground truth annotations.

A more recent work of RGB-T salient object detection, titled “RGBT Salient Object Detection: A Large-scale Dataset and Benchmark” (Tu et al., 2022, pp. 1–1) made use of a large-scale dataset containing 5000 pairs of RGB-T images with ground truth annotations with great improving in the complexity and diversity of the scenes.

2.2 Multi-Modal Salient Object Detection Datasets

With the emergence of multi-modal data, RGB-T salient object detection has been proposed, and the related RGB-T datasets have been constructed. Wang et al. (Wang et al., 2018, pp. 359-369) constructed the first RGB-T dataset VT821 with 821 pairs of RGB-T images. Next, Tu et al. (Tu et al., 2019, pp. 160–173) contribute a more challenging dataset VT1000 for RGB-T image saliency detection. The most recent RGB-T dataset VT5000, (Tu et al., 2022, pp. 1–1), introduced a much more challenging set consisting of 5000 pairs of RGB-T images. The image pairs in this dataset were recorded in different places, locations and environments. VT5000 records different categories, illuminations, sizes, positions and quantities. The aspects that were considered when creating VT5000 were size of object, illumination conditions, center bias, amounts of salient object, and background factor.

For our dataset, we use half of the VT5000 dataset as the training set and the rest of the VT5000 dataset combined with the VT1000 and VT821 dataset would be used as the test set.

2.3 Attention Mechanism

Attention mechanism was initially proposed by Bahdanau et al. (Bahdanau, Cho, & Bengio, 2014, arXiv: 1409.0473) for neural machine translation, the attention mechanisms in deep neural networks have been studied widely recently. Attention mechanisms are proven to be useful in many tasks, such as scene recognition (Cao et al., 2015, pp. 2956-2964), (Hong, You, Kwak, & Han. 2015. pp. 597–606), question answering (Yang, He, Gao, Deng, & Smola, 2016, pp. 21-29), caption generation (Xu et al., 2015, pp. 2048–2057.) and pose estimation (Chu et al., 2017, pp. 1831-1840).

In the domain of salient object detection, both cross-attention mechanisms and self-attention mechanisms have been widely explored and used to capture dependencies and attend to informative regions. These mechanisms, including channel attention and spatial attention, play crucial roles in enhancing the accuracy and performance of salient object detection models.

Cross-attention mechanisms in salient object detection allows interactions between different regions or elements within an image, facilitating the exchange of information between the encoder and decoder modules. By making use of the channel attention, which captures interdependencies between channels of feature maps. These attention maps are typically generated through a series of operations involving convolutions, pooling, and non-linear transformations. The decoder then utilizes these attention maps to guide its saliency prediction process, giving higher weights to regions that are deemed more salient based on the attended features.

The use of cross-attention mechanisms in salient object detection has shown promising results in numerous studies (Lv et al., 2023, p. 953) (Li & Yu, 2017, pp. 478-487). These mechanisms facilitate the integration of multi-scale and multi-level information, enabling the model to effectively capture the diverse characteristics of salient objects. By attending to both local details and global context, the model becomes more robust and accurate in distinguishing salient objects from the background clutter.

The distinction between cross-attention mechanisms and self-attention mechanisms lies in the scope of attention computation. Cross-attention mechanisms enable the model to attend to relevant regions that may be spatially distant but contextually significant, thanks to the incorporation of channel attention. Self-attention mechanisms, on the other hand, emphasize attending to informative regions within a single input sequence or feature map, leveraging spatial attention.

Based off this, a cross-attention complementarity module is proposed in our model to exploit the complementary information between the multi-modal features before fusing them for predicting the saliency map.

2.4 Fusion Strategies and Techniques

In the field of computer vision and image processing, various fusion techniques are utilized to combine information from multiple sources or modalities, aiming to enhance system understanding, representation, or performance. These techniques play a crucial role in addressing the challenges associated with integrating heterogeneous data. Some commonly used fusion techniques include:

(i) Early fusion (Peng, Li, Xiong, Hu, & Ji, 2014, pp. 92–109), (Song et al., 2017, pp. 4204–4216): Also known as data-level fusion, early fusion combines the input data from different sources or modalities at the beginning of the processing pipeline. The thermal infrared map is concatenated with the RGB image as the network input. Methods in this category are relatively simple, but are unable to take advantage of the difference and complementarity between RGB-T images.

(ii) Late fusion (Han, Chen, Liu, Yan & Li. 2018. pp. 3171–3183), (Wang and Gong, 2019, pp. 55277–55284): also referred to as decision-level fusion, involves processing the data from each modality separately and combining the final decisions or results at a later stage. The prediction maps provided by multi-modal branches are integrated with a fusion module. This can be achieved through techniques such as averaging, voting, or weighted combination of the individual modality outputs. This strategy is better than the early fusion, but the complementary information between the two modalities is still not fully explored.

(iii) Middle fusion (Zhang et al. 2020a, pp. 8582–8591) (Qu et al., 2016, pp. 2274-2285): RGB-T/RGB-D features are fused in different layers before being fed into a decoder for the final saliency prediction. In middle fusion, the individual modalities are fused at intermediate stages of the processing pipeline. This means that the modalities are processed separately up to a certain point, and then their representations or features are combined. The fused representations are then used for subsequent processing and decision making. Middle fusion aims to benefit from the independent processing of each modality while also allowing for the integration of modality-specific information at a later stage in the pipeline.

(iv) Model-level Fusion: Model-level fusion involves training separate models on each modality and then combining them at the model level. This can be done through techniques like ensemble methods, where predictions from different models are aggregated to make a final decision. Model-level fusion can leverage the strengths of individual models and improve overall performance.

(v) Attention-based Fusion (Tu et al., 2022, pp. 1–1): Attention mechanisms are used to selectively weigh the importance of different modalities or features at various spatial or temporal locations. Attention-based fusion allows the model to dynamically focus on the most relevant information from each modality, improving the fusion process.

(vi) Hybrid Fusion: Hybrid fusion techniques combine multiple fusion strategies to achieve a more comprehensive integration of information. This can involve a combination of early fusion, late fusion, feature-level fusion, or model-level fusion methods tailored to the specific requirements of the task.

(vii) Multi-scale fusion: A technique used in various computer vision applications, including salient object detection, to integrate information from multiple scales or resolutions. It aims to capture both fine-grained details and global context by combining features extracted at different scales. The process of multi-scale fusion involves extracting features or representations at different scales and then integrating them to form a comprehensive representation. This can be achieved through various methods like feature concatenation or multiple pooling or striding layers.

Chapter 3 Method

3.1 Proposed Network Architecture

Our proposed network architecture is named Cross-Attention Deep Fusion Network (CADFNet) and Figure 1 shows an overview. CADFNet comprises of several key components, including a two-stream multi-modal encoder, parallel high-level feature and low-level feature decoders, a cross-attention complementarity exploration (CACE) module, and a deeply supervised progressive fusion (DSPF) module. The final prediction is obtained from the last fusion unit within the DSPF module, which encapsulates the collective knowledge acquired throughout the network.

Delving into the intricacies of CADFNet, let's begin by examining the two-stream multi-modal encoder. This encoder encompasses two separate streams dedicated to different modalities, enabling the simultaneous processing of RGB and thermal data. By employing distinct pathways, the encoder captures and extracts modality-specific features, ensuring that both RBG and thermal information is adequately represented and utilized.

The high-level feature and low-level feature decoders operate in parallel, each responsible for decoding features from the respective modality-specific pathways. The high-level feature decoder primarily focuses on capturing and refining high-level semantic information, facilitating a more comprehensive understanding of the scene. Meanwhile, the low-level feature decoder emphasizes the precise delineation of object boundaries and fine details, enhancing the model's segmentation accuracy.

To facilitate effective integration and exploration of cross-modal complementarity, our CADFNet incorporates the CACE module. This module leverages cross-attention mechanisms to enable the network to dynamically adapt and align features between RGB and thermal modalities. By attending to complementary features, the network can effectively leverage the distinctive information present in each modality, enhancing the overall discriminative power and fusion capability.

Lastly, we introduce the DSPF module, which plays a critical role in facilitating progressive fusion and supervision within CADFNet. This module deeply supervises the fusion process at multiple stages, allowing for the gradual integration of information from different modalities and levels of abstraction. By iteratively refining and consolidating feature representations, the DSPF module ensures the coherent and accurate fusion of RGB and thermal information, leading to improved prediction quality.

3.2 Two-stream Multi-modal Encoder

Following (Zhai et al., 2020, pp. 8727-8742.), we employ a two-stream multi-modal encoder to extract the multi-modal features from RGB-T images. The encoder, as depicted in Figure 1, comprises two parallel backbones dedicated to processing thermal infrared images and RGB images separately. To further enhance the quality of thermal image features, we introduce multiple channel attention (CA) and spatial attention (SA) components, (Woo, Park, Lee, & So Kweon, 2018, pp. 3–19), strategically positioned between the thermal image and RGB backbones.

The channel attention helps the model to pay more attention to the important aspects or features by assigning different levels of importance to each aspect. It does this by analyzing the relationship between different aspects and determining their importance. By emphasizing the important aspects and de-emphasizing the less important ones, channel attention helps the model focus on the most relevant information for the task it is trying to solve and the spatial attention.

Spatial attention helps the model to selectively focus on the important regions or parts of the image. It does this by analyzing the spatial relationships or dependencies between different regions. By identifying the important regions and giving them more attention, spatial attention helps the model make more precise and accurate predictions.

The addition of CA and SA components to our model aims to improve the spatial awareness of the extracted thermal image features. The encoder makes use of the channel attention component to make the network to be able selectively amplify the important channel-wise information within the thermal image features thereby effectively enhancing their representational quality. Correspondingly, the spatial attention component focuses on capturing relevant spatial context and this enables the model to better understand the spatial relationships and correlations within the thermal image features.

To facilitate effective fusion of the thermal image and RGB features, we combine the enhanced thermal image features with the corresponding RGB features. The side-output RGB and thermal image features extracted from the backbones are denoted as:

$$\{X_{\mathrm{RGB}}^{(\mathrm{i})} | i = 1, 2 \ldots, 5\} \tag{3-1}$$

and

$$\{\mathrm{X}_{\mathrm{T}}^{(\mathrm{i})} | \mathrm{i}{=}1, 2\ldots, 5\}, \tag{3-2}$$

the fused features $\mathrm{X}^{(\mathrm{i})}$ at level *i* can be defined as:

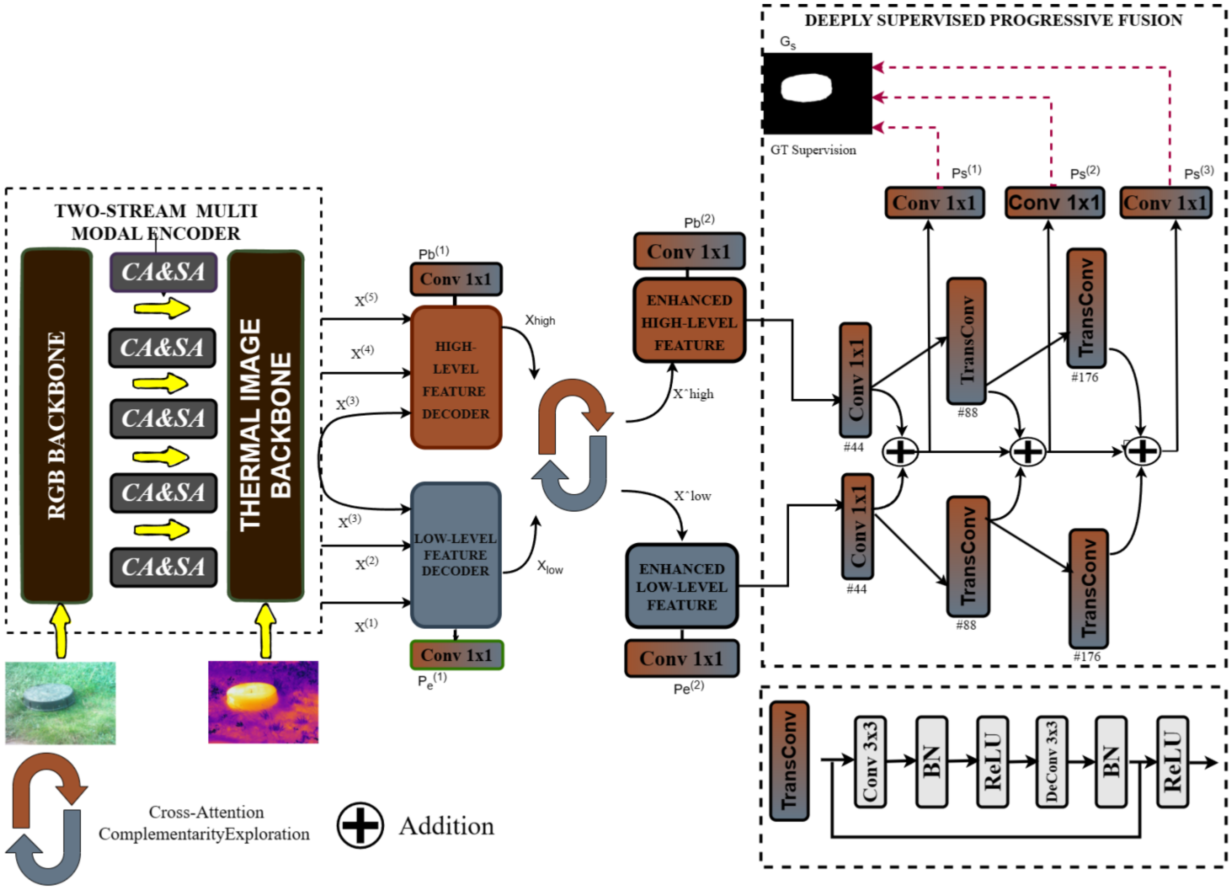
$$\mathrm{X}^{(\mathrm{i})} = \mathrm{SA}\left(\mathrm{CA}\left(\mathrm{X}_{\mathrm{T}}^{(\mathrm{i})}\right)\right) + \mathrm{X}_{\mathrm{RGB}}^{(\mathrm{i})} \tag{3-3}$$

where CA(·) and SA(·) denote the CA and SA components, respectively. The resulting features are then fed to the subsequent high-level feature and low-level feature decoders. The right of Figure 2 for the detailed architectures of the CA and SA components.

Figure 1

Architecture of CADFNet.

The multi-modal features are extracted from in the two-stream encoder and the high-level features {$X^{(3)}$*,* $X^{(4)}$*,* $X^{(5)}$*} and low-level features {*$X^{(1)}$*,* $X^{(2)}$*,* $X^{(3)}$*} from the encoder are fed to two parallel high-level feature and low-level feature decoders to learn high-level and low-level features (*X_{high}*,* X_{low}*). The CACE model exploits the complementary information to enrich the features and resulting features are fused with the DSPF module to predict the final saliency map. At each fusion bock, a GT supervision is done.*



3.3 Parallel Decoders

The parallel decoders are designed based on (Wu, Su, & Huang, 2019a, pp. 3907-3916), where a top-down architecture is proposed to propagate high-level features to the lower level in a layer-wise manner. Top-down architecture decoders are a type of decoding module commonly used in neural network models, particularly in computer vision tasks such as image segmentation. The flow of information follows a "top-down" direction, where higher-level features are progressively downsampled and passed to the lower layers. This allows the model to combine the rich semantic information from the higher-level features with the detailed spatial information from the lower-level features. The objective is to refine the predictions and generate more accurate and detailed outputs. We also include multiple receptive field block (RFB) modules (Liu, Huang, & Wang, 2018, pp. 404-419) in the decoders to enhance the receptive field of the network, allowing it to capture more context and global information from the input data and enlarging the receptive field for improved performance.

Therefore, we propose a parallel architecture with two decoders to refine the high-level features X_{high} and low-level features X_{low} obtained from the encoders. Denoting the two decoders as D_{high} and D_{low}, we have

$$X_{\text{high}} = D_{\text{high}}(X^{(3)}, X^{(4)}, X^{(5)}) \tag{3-4}$$

and

$$X_{\text{low}} = D_{\text{low}}(X^{(1)}, X^{(2)}, X^{(3)}). \tag{3-5}$$

High-level features capture more abstract and rich semantic information about the input data. These features are typically obtained from deeper layers of the network or from layers that have undergone several levels of feature transformation. High-level features are particularly useful for predicting features pertaining to the body as they encode complex patterns, global context, and higher-level semantics of the input. Low-level features contain a substantial amount of primary information about image boundary information, thus making them suitable for predicting features that have to do with the image borders. Low-level features are sensitive to fine-grained details, contour, and texture information. The proposal of this module in our model was motivated by the involvement and observation of high-level and low-level features.

The decoding process of our parallel decoders involves the following steps:

(i) Upsampling: The high-level features are upsampled to match the spatial resolution of the corresponding lower-level features. This is done to ensure that the features from different levels can be fused or concatenated effectively.

(ii) Feature Fusion: The upsampled high-level features are combined with the lower-level features at each corresponding spatial resolution. This fusion process allows the model to integrate the semantic information from the high-level features with the spatial details from the lower-level features.

(iii) Local Refinement: The fused features are further processed through convolutional layers or other operations to refine the predictions. This step helps to improve the spatial accuracy and enhance the finer details in the output.

3.4 Cross- Attention Complementarity Exploration

We propose a novel cross-attention complementarity exploration (CACE) module to exploit the complementary information between the high-level and low-level features and refine them before fusing them for predicting the saliency map. As shown in Figure 2, as the first step, we begin by computing the cross-attention coefficients from the high-level and low-level features. These computed coefficients are then used weights to extract complementary features from one branch to enrich the features in the other branch via a concatenation operation. Mathematically, our CACE module takes the feature outputs given by the two decoders, i.e., (X_{high}, X_{low}), as input and provides the refined ones ($\widehat{\mathrm{X}}_{\text{high}}$, $\widehat{\mathrm{X}}_{\text{low}}$) as the final product. Given the high-level feature X_{high}, the channel and spatial attention coefficient maps $\mathrm{A}_{\text{high}}^{(\text{CA})}$and $\mathrm{A}_{\text{high}}^{(\text{SA})}$are defined as:

$$\mathrm{A}_{\text{high}}^{(\text{CA})}) = \mathrm{CA}(X_{\text{high}}) \tag{3-6}$$

And

$$\mathrm{A}_{\text{high}}^{(\text{SA})} = \mathrm{SA}(X_{\text{high}}) \tag{3-7}$$

The detailed architectures of the CA and SA components are shown on the right of Figure 2. The channel and spatial attention coefficient maps $\mathrm{A}_{\text{low}}^{(\text{CA})}$and $\mathrm{A}_{\text{low}}^{(\text{SA})}$of the low-level feature branch as:

$$\mathrm{A}_{\text{low}}^{(\text{CA})}) = \mathrm{CA}(X_{\text{low}}) \tag{3-8}$$

And

$$\mathrm{A}_{\text{low}}^{(\text{SA})} = \mathrm{SA}(X_{\text{low}}) \tag{3-9}$$

Based on the resulting attention coefficient maps, we exploit the complementarity between high-level and low-level features and obtain the refined ones as

$$\hat{X}_{\text{high}} = \text{Cat}((\text{A}^{(\text{CA})}_{\text{high}})\ \text{A}^{(\text{SA})}_{\text{high}} = \ \odot X_{\text{low}}), X_{\text{high}}) \tag{3-10}$$

$$\hat{X}_{\text{low}} = \text{Cat}((\text{A}^{(\text{CA})}_{\text{low}})\ \text{A}^{(\text{SA})}_{\text{low}} = \ \odot X_{\text{high}}), X_{\text{low}}) \tag{3-11}$$

Where Cat(·) denotes a concatenation operation and $\odot$ represents the Hadamard product.

Figure 2

*Architecture of the CACE module***.** *High-level and low-level features are passed through two parallel branches, each of which consists of channel and spatial attention components for exploiting the complementary information. Finally, the complementary feature is concatenated with the original one as the refined feature.*

3.5 Deeply Supervised Progressive Fusion

In our proposed method, we introduce the Deeply Supervised Progressive Fusion (DSPF) module as a final step to fuse the multi-modal features in order to obtain accurate salient object detection (SOD). The DSPF module, shown on the right side of Figure 1, is designed with three layers, each consisting of a fusion unit that is supervised by the ground truth.

The overall architecture of the DSPF module is organized in a progressive framework, where the fusion process gradually refines and enhances the saliency prediction to get a final output. This progressive approach allows for a more comprehensive and detailed fusion of the multi-modal features.

Our DSPF module generally comprises of two main components which are progressive feature fusion and deep supervision.

(i) Progressive Feature Fusion: Being that the outputs of the CACE model are the paired high-level and low-level features ($\hat{X}_{\text{high}}$, $\hat{X}_{\text{low}}$) $\in \mathbb{R}\,\frac{\mathrm{H}}{\mathrm{W}} \times \frac{\mathrm{W}}{4}\ \times C$ with H and W denoting the height and width of the input images, and C denoting the number of channels, we aim to first fuse the two types of features and secondly fill the dimension gap between ($\hat{\mathrm{X}}_{\text{high}}$, $\hat{\mathrm{X}}_{\text{low}}$) and the ground truth saliency map $G \in \mathbb{R}^{H\times W}$. Therefore, we progressively enlarge the dimension of feature maps in a layer-by-layer manner and instead of using a combination of convolutional and pooling layers, we employ the transpose of convolutional layers (Jiang, Zheng, Luo & Zhang, 2018, *arXiv preprint arXiv:1806.01054*), otherwise called TransConv (shown on the bottom right of Figure 1), to simultaneously learn feature representations and enlarge the dimensions of feature maps. Finally, after two transpose of convolutional layers (TransConv) followed by a 1 × 1 convolutional layer and an upsampling layer, we obtain full-resolution feature maps, which match the dimensions of the ground truth saliency maps. At each layer, we fuse the multi-modal features using an efficient fusion unit, which combines the high-level feature, low-level feature, and upsampled output of the fusion unit from the previous layer. The final output, $P_{\mathrm{s}}^{\,(3)}$, is provided by the last fusion unit.

(ii) Deep Supervision: An integral component incorporated into our DSPF model, we design a deep supervision (Lee, Xie, Gallagher, Zhang & Tu, 2015, pp. 562-570) component to mitigate the challenge of vanishing gradients, facilitate better gradient flow and to enhance and improve the overall performance of our method. To implement deep supervision, we introduce a convolutional layer with a kernel size of 1×1 for each side output of the DSPF module. This convolutional layer serves the purpose of reducing the number of channels to one, thereby simplifying the saliency map representation. Subsequently, the output of this convolutional layer, denoted as $P_{\mathrm{s}}^{(3),}$ undergoes upsampling to match the resolution of the ground truth saliency map. To enhance the learning process and to enable effective gradient propagation, we apply the same supervision unit to each layer of the DSPF module. By doing so, we ensure that each layer receives direct feedback from the ground truth saliency map thus facilitating the optimization of the model at multiple levels. This deep supervision strategy enables the model to capture meaningful information from different layers and promote accurate saliency predictions.

Chapter Four Experiment Results

4.1 Experiment Setup

Our CADFNet is implemented using the popular deep learning framework PyTorch, and in building our architecture, we adopt the widely acclaimed ResNet-50 (He, Zhang, Ren, & Sun, 2016, pp. 770-778) as our backbone network. Similar to previous works such as (Zhai et al., 2020, pp. 8727-8742.), (Wu, Su & Huang, 2019, pp. 3907-3916), we modify the network by removing the last pooling and fully connected layers, and we extract features from the 5 intermediate layers, which serve as the side outputs of our model.

To train our model, we utilize the convenient VSCode IDE, which offers a user-friendly coding experience and facilitates efficient model development. Furthermore, we execute our experiments on a remote SSH server, allowing us to leverage its computational resources for faster and more efficient training. Throughout the training process, we set the number of epochs to 201, ensuring sufficient iterations to capture the complex patterns and relationships in the data. For each iteration, we utilize a mini-batch size of 8, striking a balance between computational efficiency and model optimization.

We employ the Adam optimizer (Kingma & Ba, 2014, abs/1412.6980.) to update the network weights during training. We initialize the learning rate to 1e-4 for achieving stable and effective convergence. To further enhance training dynamics, we schedule the learning rate to decrease by a factor of 10 every 60 epochs, allowing the model to gradually refine its predictions and converge towards the optimal solution.

4.2 Experimental Dataset

For our experiment, we used the approach taken by Tu et al. (Tu et al., 2022, pp. 1–1) to train and test our model. We make use of the VT5000 (Tu et al., 2019, pp. 160–173), VT1000 (Tu et al., 2019, pp. 160–173) and VT821 (Wang et al., 2018, pp. 359-369) datasets.

For the VT821 dataset, 821 RGB-T image pairs were collected by recording system which consisted of an online thermal imager (FLIR A310) and a CCD camera (SONY TD-2073). For alignment, a number of point correspondences in each image pairs were uniformly selected, and the homography matrix was computed by the least-square algorithm.

The imaging hardware used for the VT1000 dataset was the FLIR SC620, which consist of a thermal infrared camera and a CCD camera. This meant that the two cameras have the same imaging parameters but not the same focus. The optical axes was aligned as parallel and the image alignment was manually done to totally overlap the RGB image with the thermal infrared image. This was achieved by enlarging and cropping the visible image.

And finally for the VT5000, the equipment used to collect the RGB and the thermal infrared images were the FLIR T640 and T610. They are equipped with a thermal infrared camera and a CCD camera. Both cameras have same imaging parameters, and thus there was no need to manually align the RGB and thermal infrared images singly. This was an advantage as it reduces errors gotten from manual alignment.

To train the model, we use 2500 pairs of RGB-T in the VT5000 dataset and the other half of the VT5000 dataset combined with the VT1000 and VT821 altogether is used to test the model.

4.3 Data pre-processing and Data augmentation

The input images are resized to 352 × 352 to ensure a consistent size for training. To enhance the variety of training data, all the images undergo several augmentation processes, including random flipping, rotation, and boundary cropping. These augmentations used are as follows:

During the random flipping process, each image is randomly mirrored either horizontally or vertically. This augmentation helps the model learn to recognize objects from different perspectives and improves its generalization capability.

The rotation augmentation randomly rotates the images within a range of -15 to +15 degrees. By introducing variations in orientation, the model becomes more resilient to object rotations in real-world scenarios.

To further increase the variability of the training data, random boundary cropping is applied. This process randomly selects a region of interest within each image, effectively clipping the boundaries. By focusing on specific regions, the model learns to extract relevant features and becomes more attentive to important image details.

Additionally, the color enhancement technique is employed to introduce variations in brightness, contrast, color intensity, and sharpness. By randomly adjusting these parameters within certain ranges, the model becomes more adaptable to variations in lighting conditions and color distributions.

Two noise-based augmentations, namely Gaussian noise and pepper noise, are also utilized. Gaussian noise is added by perturbing the pixel values of the image using a Gaussian distribution. This helps the model become resilient to random noise present in real-world images. Pepper noise randomly modifies individual pixels, setting them to either black or white. This augmentation simulates noisy image conditions and improves the model's ability to handle image artifacts.

4.4 Evaluation Metrics

To prove the efficiency of our method, we compare it to other state of the art methods. To carry out this comparison, we use the widely used evaluation metrics F-measure (F_{β}) (Achanta, Hemami, Estrada & Susstrunk, 2009, pp. 1597-1604), S-measure (S_{α}) (Fan, Cheng, Liu, Li & Borji, 2017, pp. 4548-4557), E-measure (E_{ξ}) (Fan et al., 2018, pp. 689-704), Precision Recall (PR) curves and mean absolute error (MAE), (Perazzi et al., 2012, 733–740). They are detailed on below:

- F-measure ($F\beta$) is used to set the evaluation criteria and is used to evaluate the saliency map quality by computing the weighted harmonic mean of the precision and recall.

$$F_{\beta}=\frac{(1+\beta^2)\cdot \text{Precision}\cdot \text{Recall}}{\beta^2\cdot \text{Precision}+\text{Recall}} \tag{4-1}$$

- The S-measure(S_{α}) is combines region-aware structural similarity (S_r) and object-aware structural similarity for evaluation (S_o):

$$S_{\alpha} = (\alpha)\ (\text{So}) + (1 - \alpha)\ (\text{S}_\text{r}) \tag{4-2}$$

Where $\alpha = 0.5$.

- E-measure (E_{ξ}) carries out evaluation by utilizing image-level and local-pixel statistics. It is based on cognitive visual studies.
- The Precision-Recall (PR) curve is used for evaluating the performance of saliency detection. It is generated by converting the saliency map into a binary map using

thresholds ranging from 0 to 255, and then comparing these binary maps with the ground truth.

- Mean Absolute Error (MAE) is a complement to the PR curve and it does a quantitative measurement of the difference in average between the predicted saliency map S and the ground truth G at the pixel level.

$$\text{MAE}=\frac{1}{\text{W x H}}\sum_{x=1}^{W}\sum_{y=1}^{H}|\,\text{S(x,y)-G(x,y)}| \tag{4-3}$$

4.5 Comparison with Existing Methods

We compare our methods with other state of the art methods to prove that our model is effective in detecting salient objects ion RGB-T images. To achieve this, we compare our method with 9 other state of the art methods. The methods are: ADFNet (Tu et al., 2022, pp. 1–1), BASNet (Qin et al., 2019. pp. 7479-7489), CPD (Wu, Su & Huang, 2019, pp. 3907-3916), EGNet (Zhao et al., 2019, pp. 8779-8788), MTMR (Wang et al., 2018, pp. 359-369)), PFA (Zhao & Wu, 2019, pp. 3084-3094), PoolNet (Liu, Hou, Cheng, Feng & Jiang, 2019, pp. 3917-3926), R3Net (Deng et al., 2018, pp.684 – 690), and SCGL (Tu et al., 2019, pp. 160–173). All these methods are deep learning based and utilize deep features except MTMR, which is a traditional model. Thus, we compare the combined deep features from our method and compare it to these mentioned methods.

4.6 Quantitative Comparison

We compare our method with other methods using the F-measure scores, E-measure scores, S-measure scores and the MAE.

In the VT821 dataset, our model CADFNet achieves a max F-measure of 82.56%, outperforming all other methods in the F-measure scores, E-measure scores, S-measure scores and the MAE.

It outperforms the top methods such as ADFNet by 10.07%, BASNet by 8.52%, and EGNet by 8.86%. In terms of max E-measure, CADFNet achieves 90.77%, surpassing well-known state of the art methods like CPD by 5.81% and EGNet by 10.13%. CADFNet also achieves a high S-measure of 85.86%, indicating its ability to measure the overall similarity of saliency maps and an MAE of 4.22%.

For the VT1000 dataset, our method is outperformed in the F-measure scores by ADFNet for approximately 2.81%, BASNet for approximately 2.26% and slightly by CPD and EGNet both which are below 1% but our method outperforms the rest of the methods.

BASNet and CPD also slightly outperform our method by less than 1% with our method outperforming the rest of the methods. Our method also performs better than the majority of the other methods in the S-measure scores and MAE scores.

CADFNet excels on the challenging VT5000 dataset, achieving outstanding results. It achieves a remarkable Max F-measure of 93.48%, Max E-measure of 98.62%, S-measure of 95.11%, and an MAE as low as 1.52%. It outperforms every method in every score. The score of each baseline methods in each benchmark dataset is shown in Table 4-1. We also commutated the PR curves across the 3 datasets (Figure 3). This is to compare the truth-positive rate of our saliency map to other methods.

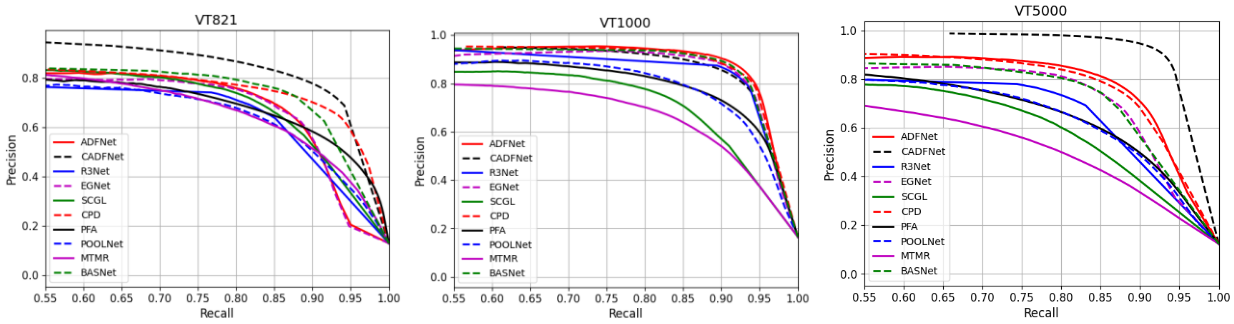
Table 4-1

Quantitative comparison of our model with other state of the art models using max F-measure (F_β), max E-measure (E_ξ), S-measure (S_α), and MAE (MAE) scores, on the 3 benchmark datasets VT5000, VT1000 & VT821 with ↑ (↓) denotes that the higher (lower) the better. Our model, CADFNet, is in ***bold****.*

From this quantitative analysis, we can see that our model performs exceptionally well in the VT821 dataset and in the VT5000 dataset but falters diminutively in its performance in the VT1000 dataset.

Figure 3

Precision-Recall curves of our model compared to other methods to analyze the performance on the three datasets.



4.7 Visual Comparison

For qualitative comparison between our proposed method and other methods, we make use of a visual comparison. The images used in this comparison all contain a different type of challenging scenario and the output given by every method is displayed. (Figure 4)

Figure 4

A visual comparison between our model and other methods

Chapter Five Ablation Studies

5.1 Effectiveness of CADFNet model

For the ablation analysis, we investigate the effectiveness of the CACE and the DSPF modules of the proposed method. We perform this action by removing both the CACE module and the DSPF modules from the network. For the CACE, removing the model creates the first instance of the ablated version A1. In this model, the features from the decoders are directly fed to our DSPF module for the prediction of saliency maps. Table 5-1 shows the decrease in performance after the CACE module has been removed.

Table 5-1

Quantitative comparison results of the ablated models using the S-measure (S_α) and MAE (MAE) scores on the three benchmark datasets. ↑ (↓) denotes that the higher (lower) the better. Best score in each row is highlighted in ***bold****.*

	VT821		VT1000		VT5000	
MODEL	S_α↑	MAE↓	S_α↑	MAE↓	S_α↑	MAE↓
A1	0.8498	0.0501	0.8726	0.0412	0.9490	0.0161
B1	0.8499	0.0502	0.8726	0.0413	0.9501	0.0159
B2	0.8554	0.0474	0.8891	0.0395	0.9504	0.0158
CADFNet	**0.8586**	**0.0422**	**0.8969**	**0.0377**	**0.9511**	**0.0152**

We then perform two experiments on the DSPF module to create ablated versions B1 and B2. The first ablated version, B1, is created by removing the DSPF module from CADFNet and using a simple addition operation, followed by a 1 × 1 convolutional layer and ground truth supervision. The results shown in Table 5-1 demonstrate that our model CADFNet outperforms “B1” by a large margin on all datasets. This indicates that our DSPF module is an effective fusion module capable of integrating the high-level and low-level features for accurate SOD. In addition, we create the second ablated version “B2” by removing the deep supervision. As shown in Table 5-1, “B2” outperforms “B1”, but cannot compete with CADFNet. This demonstrates the importance of deep supervision for improving the performance.

Chapter Six Conclusion

In conclusion, this thesis dissertation highlights the advancements and advances made in salient object detection achieved through RGB-T images and deep learning approaches. Throughout this report, we demonstrated the effectiveness of various architectural components and evaluation metrics. These findings contribute to the development of better salient object detection algorithms and to gain more accurate results.

In this work, we propose a novel cross-attention and deep progressive fusion based network to refine, enhance and fuse multi-modal features to obtain accurate saliency detection results. We build our network on a two-stream multi-modal encoder to extract multi-modal features from RGB-T images and then these features are passed on to a parallel decoder to enhance multi-modal features by leveraging the high level and low level features. The output is then passed through a cross-attention complementary exploration (CACE) module to refine and enhance these features and finally, a deeply supervised progressive fusion (DSPF) module is proposed to carry out progressive fusion of these multi-modal features.

We train and test our model using the combination of the VT5000, VT1000 and VT821 datasets and then compare our model to other existing state of the art methods. Based on the results obtained, we drew conclusions on how saliency detection in RGB-T images can be improved in the future. For RGB-T salient object detection to be improved, deep learning methods have to be further explored. There should be more exploration on how to design deep networks that extract special features of RGB and thermal modalities and take them into consideration and there should be further studies on how to make the best use of attention mechanisms and semantic information.

References

Achanta, R.; Hemami, S.; Estrada, F.; and Susstrunk, S, “Frequency-tuned salient region detection,” in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 1597–1604.

C. Cao, X. Liu, Y. Yang, Y. Yu, J. Wang, Z. Wang, Y. Huang, L. Wang, C. Huang, W. Xu, D. Ramanan, and T. S. Huang, “Look and think twice: Capturing top-down visual attention with feedback convolutional neural networks,” in Proceedings of IEEE International Conference on Computer Vision", 2015. pp. 2956-2964

C. Lv, B. Wan, X. Zhou, Y. Sun, J. Hu, J. Zhang, and C. Yan, “CAE-Net: Cross-Modal Attention Enhancement Network for RGB-T Salient Object Detection,” Electronics, 2023, p. 953.

D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” 2014, arXiv: 1409.0473. [Online]. Available: https://arxiv.org/abs/1409.0473

D.P. Kingma, & J. Ba, “Adam: A Method for Stochastic Optimization,” 2014, arXiv: 1412.6980. [Online]. Available: https://arxiv.org/abs/1412.6980

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

Fan, D.-P.; Gong, C.; Cao, Y.; Ren, B.; Cheng, M.-M.; and Borji, A., “Enhanced-alignment Measure for Binary Foreground Map Evaluation,” in Proceedings of the International Joint Conference on Artificial Intelligence, 2018, pp. 698–704.

F. Perazzi, P. Krahenbuhl, Y. Pritch, and A. Hornung, “Saliency filters: Contrast based filtering for salient region detection,” In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 733–740.

G. Li, and Y. Yu, “Deep Contrast Learning for Salient Object Detection,” in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 478-487.

G. Wang, C. Li, Y. Ma, A. Zheng, J. Tang and B. Luo, “Rgb-t saliency detection benchmark: Dataset, baselines, analysis and a novel approach,” in Proc. Chin. Conf. Image Graph. Technol., 2018, pp. 359-369

H. Peng, B. Li, W. Xiong, W. Hu, and R. Ji, “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.

H. Peng, B. Li, W. Xiong, W. Hu, and R. Ji, “RGBD salient object detection: a benchmark and algorithms,” In Proceedings of the European Conference on Computer Vision, 2014, pp. 92–109. Springer.

H. Song, Z. Liu, H. Du, G. Sun, O. Le Meur, and T. Ren, “Depth-aware salient object detection and segmentation via multiscale discriminative saliency fusion and bootstrap learning,” IEEE Transactions on Image Processing, 2017, pp. 4204–4216.

He, K.; Zhang, X.; Ren, S.; and Sun, J, “Deep Residual Learning for Image Recognition,” In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

J. Han, H. Chen, N. Liu, C. Yan, and X. Li, "CNNs Based RGB-D Saliency Detection via Cross-View Transfer and Multiview Fusion," IEEE transactions on cybernetics, 2018, pp. 3171–3183.

J. Li, W. Hu, Z. Zhang, and L. Zhang, "Joint RGB and depth saliency estimation via a unified multi-scale deep network," In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2809-2817.

J. Zhang, D.-P. Fan, Y. Dai, S. Anwar, F. S. Saleh, T. Zhang, and N. Barnes, "UC-Net: uncertainty inspired RGB-D saliency detection via conditional variational autoencoders," In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2020a, pp. 8582–8591.

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

J.-X. Zhao, J.-J. Liu, D.-P. Fan, Y. Cao, J. Yang, and M.-M. Cheng, "Egnet: Edge guidance network for salient object detection," in Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 8779-8788

Jiang, J.; Zheng, L.; Luo, F.; and Zhang, Z, "Rednet: Residual encoder-decoder network for indoor RGB-D semantic segmentation," 2018, arXiv:1806.01054, [Online]. Available: https://arxiv.org/abs/1806.01054

Kelvin Xu, J. Ba, R. Kiros, K. Cho, A. C. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio, "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention," In: Proceedings of the 32nd International Conference on Machine Learning, July 2015, pp. 2048–2057.

Lee, C.-Y.; Xie, S.; Gallagher, P.; Zhang, Z.; and Tu, Z, “Deeply-supervised nets,” In Artificial intelligence and statistics, 2015, pp. 562–570. PMLR.

N. Wang, and X. Gong, “Adaptive fusion for RGB-D salient object detection,” IEEE Access, 2019, pp. 55277–55284.

Q. Wang, H. Liu, X. Huang, and L. Wang, “Salient object detection in RGB-D scenes using depth-weighted information entropy,” Signal Processing: Image Communication, 47, 2016, pp. 207-217.

Qu, L., He, S., Zhang, J., Tian, J., Tang, Y., and Yang, Q, “RGBD Salient Object Detection via Deep Fusion,” IEEE Transactions on Image Processing, 2016, 26, pp. 2274-2285.

S. Hong, T. You, S. Kwak, and B. Han, “Online tracking by learning discriminative saliency map with convolutional neural network,” in International conference on machine learning (PMLR) (IEEE), 2015, pp. 597–606

S. Liu, D. Huang, and Y. Wang, “Receptive Field Block Net for Accurate and Fast Object Detection,” In Proceedings of the European Conference on Computer Vision, 2018, pp. 404–419.

S. Woo, J. Park, J.-Y. Lee, and I. So Kweon,” CBAM: Convolutional Block Attention Module,” In Proceedings of the European Conference on Computer Vision, 2018, pp. 3–19.

T. Zhao and X. Wu, “Pyramid feature attention network for saliency detection,” in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 3084-3094

X. Chu, W. Yang, W. Ouyang, C. Ma, A. L. Yuille, and X. Wang, “Multi context attention for human pose estimation,” in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1831-1840

X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan, and M. Jagersand, “Basnet: Boundary-aware salient object detection,” in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 7479-7489

Y. Zhai, D. Fan, J. Yang, A. Borji, L. Shao, J. Han, and L. Wang, “Bifurcated Backbone Strategy for RGB-D Salient Object Detection,” IEEE Transactions on Image Processing, 30, 2020, pp. 8727-8742.

Y. Zhang, J. Ma, and S. Jiang, “Salient object detection in RGB-D images with saliency evolution,” In Proceedings of the 22nd ACM international conference on Multimedia, 2014, pp. 249-258.

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

Z. Tu, T. Xia, C. Li, X. Wang, Y. Ma, and J. Tang, “Rgb-t image saliency detection via collaborative graph learning,” IEEE Transactions on Multimedia, vol. 22, no. 1, pp. 160–173, 2019.

Z. Tu, Y. Ma, Z. Li, C. Li, J. Xu, and Y. Liu, “RGBT salient object detection: A large-scale dataset and benchmark,” IEEE Trans. Multimedia, 2022, pp. 1–1,

Z. Wu, L. Su, and Q. Huang, “Cascaded Partial Decoder for Fast and Accurate Salient Object Detection,” In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019a, pp. 3907–3916.

Z. Yang, X. He, J. Gao, L. Deng, and A. J. Smola, “Stacked attention networks for image question answering,” in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 21-29