


## Article

# Real-Time Runway Detection Using Dual-Modal Fusion of Visible and Infrared Data

Lichun Yang <sup>1,2</sup> , Jianghao Wu <sup>1</sup> , Hongguang Li <sup>3,\*</sup>, Chunlei Liu <sup>3</sup> and Shize Wei <sup>4</sup>

<sup>1</sup> School of Transportation Science and Engineering, Beihang University, Beijing 100191, China; yanglc2003@buaa.edu.cn (L.Y.); buaawjh@buaa.edu.cn (J.W.)

<sup>2</sup> Jiangsu Automation Research Institute, Lianyungang 222061, China

<sup>3</sup> Institute of Unmanned System, Beihang University, Beijing 100191, China; chunlei191@gmail.com

<sup>4</sup> School of Electronics and Information Engineering, Beihang University, Beijing 100191, China; weishize@buaa.edu.cn

\* Correspondence: lihongguang@buaa.edu.cn

**Abstract:** Advancements in aviation technology have made intelligent navigation systems essential for improving flight safety and efficiency, particularly in low-visibility conditions. Radar and GPS systems face limitations in bad weather, making visible–infrared sensor fusion a promising alternative. This study proposes a salient object detection (SOD) method that integrates visible and infrared sensors for robust airport runway detection in complex environments. We introduce a large-scale visible–infrared runway dataset (RDD5000) and develop a SOD algorithm capable of detecting salient targets from unaligned visible and infrared images. To enable real-time processing, we design a lightweight dual-modal fusion network (DCFNet) with an independent–shared encoder and a cross-layer attention mechanism to enhance feature extraction and fusion. Experimental results show that the MobileNetV2-based lightweight version achieves 155 FPS on a single GPU, significantly outperforming previous methods such as DCNet (4.878 FPS) and SACNet (27 FPS), making it suitable for real-time deployment on airborne systems. This work offers a novel and efficient solution for intelligent navigation in aviation.

**Keywords:** intelligent navigation; salient object detection; dual-modal fusion; airport runway detection; deep learning; lightweight network



Academic Editor: Weimin Huang

Received: 28 December 2024

Revised: 12 February 2025

Accepted: 13 February 2025

Published: 16 February 2025

**Citation:** Yang, L.; Wu, J.; Li, H.; Liu, C.; Wei, S. Real-Time Runway Detection Using Dual-Modal Fusion of Visible and Infrared Data. *Remote Sens.* **2025**, *17*, 669. <https://doi.org/10.3390/rs17040669>

**Copyright:** © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

The development of advanced intelligent navigation systems [1–4] for aircraft has become a crucial focus in modern aviation, aiming to enhance safety, efficiency, and situational awareness, particularly in low-visibility conditions. One of the critical components of such systems is the ability to detect and identify airport runways accurately and reliably during various flight phases, particularly during takeoff, landing, and approach. Traditional navigation systems, such as radar and GPS-based technologies, often encounter limitations when it comes to precise runway detection under adverse weather conditions like fog, heavy rain, or nighttime operations. These systems struggle with reduced visibility and contrast, making it difficult to accurately identify runways.

To address this issue, this paper presents a novel approach for integrating visible and infrared (IR) sensors into aircraft navigation systems. Specifically, we propose a fusion strategy utilizing visible and IR sensors, coupled with a dual-modal salient object detection (dual-modal SOD) algorithm, to improve airport runway detection and recognition. Salient object detection (SOD) identifies the most prominent regions in an image, mimicking the

human visual system [5–7]. Dual-modal SOD extends this concept by combining data from two different modalities (e.g., visible and infrared) to enhance detection accuracy under challenging conditions. The synergy between visible light and infrared imaging can enhance runway visibility across a broader range of environmental conditions. Visible sensors, providing rich scene information like color and texture, are highly effective under clear daylight conditions. However, their performance deteriorates in low-visibility scenarios, such as fog, haze, or overcast skies, where light scattering and reduced contrast impair scene clarity. In contrast, infrared sensors, especially those operating in the 900–1700 nm wavelength range, perform exceptionally well in low-visibility environments, such as nighttime operations and foggy weather, as they can penetrate atmospheric disturbances better than visible light. Therefore, this paper aims to combine the strengths of both modalities to achieve comprehensive runway detection.

In visual navigation systems for approach and landing, simulating the pilot's visual experience is essential. The system must quickly focus on the airport runway, which becomes a "salient object" for navigation. Therefore, leveraging the SOD method to extract the airport runway region aligns well with cognitive processes.

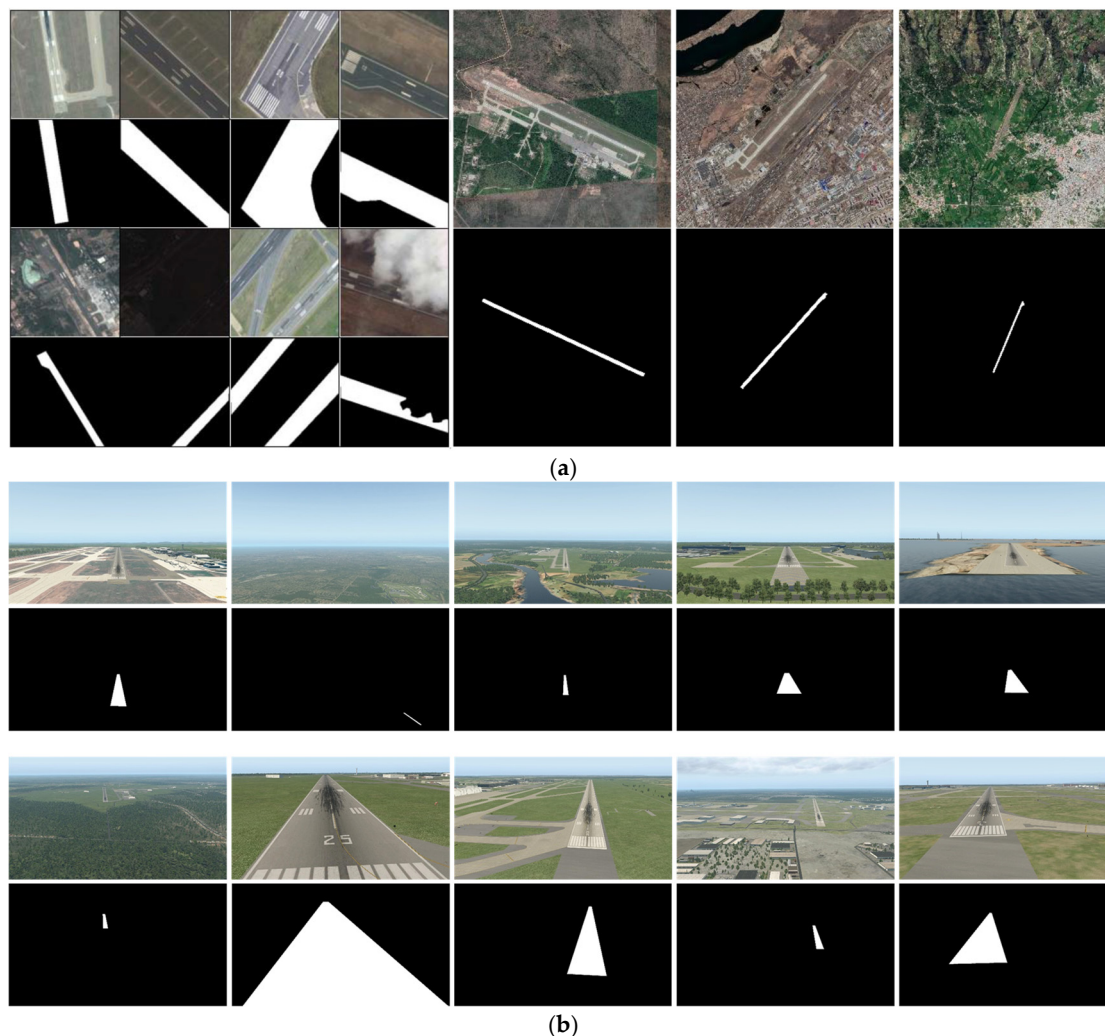
Recent research in deep-learning-based dual-modal SOD has garnered substantial attention, particularly in RGB-T [8–23] and RGB-D [24–28] fusion methods. These methods improve object detection accuracy under challenging conditions by learning features from multiple sources such as visible, infrared, and depth imagery. Their success relies on large-scale, real-world RGB-T and RGB-D datasets [29–35] and the computational power of deep learning models [36–39]. However, these datasets typically require pre-aligned images, which involve additional labor costs in practical applications.

Meanwhile, cross-view matching [40–42] and visible–thermal person re-identification (VI Re-ID) [43] have emerged as significant research areas. Cross-view matching focuses on identifying and matching objects or scenes across different viewpoints, while VI Re-ID leverages multimodal data to recognize individuals under varying conditions. Although our study primarily addresses runway detection from a single perspective, our approach inherently aligns with the broader objectives of cross-view matching and VI Re-ID. Specifically, by processing pairs of visible and infrared images that may be captured at different times or from different angles, our method aims to enhance consistency and accuracy in object detection across diverse viewpoints, thereby contributing to the advancement of these related fields.

However, when dual-modal SOD is applied to airport runway scenarios, it still faces numerous challenges. Firstly, due to the sensitivity and inaccessibility of airport areas, the creation of large-scale, publicly accessible datasets related to airport runways remains a significant challenge. Although several studies [44–46] have constructed airport runway datasets using remote sensing imagery, these are captured from a ground-based viewpoint, making them unsuitable for aircraft landing applications. Moreover, some recent studies provide datasets based on aircraft landing perspectives; however, these datasets are simulations generated using X-Plane [47,48] or Unreal Engine 4 [49], which differ from real-world data (as shown in Figure 1). To date, no visible–infrared dual-modal dataset for airport runway detection is available. To address this research gap, we have created a large-scale, real-world RGB-IR airport runway dataset (RDD5000), captured in actual scenes.

Secondly, most existing dual-modal SOD algorithms rely on strictly aligned datasets for training. However, in the multimodal dataset we use (RDD5000), the visible and infrared images of the same scene are not perfectly aligned due to variations in shooting conditions and camera parameters (as shown in Figure 2). This misalignment is common in real-world applications, where obtaining perfectly aligned images is often impractical. Recently, there has been growing attention on non-aligned dual-modal SOD methods [14,50–52], and

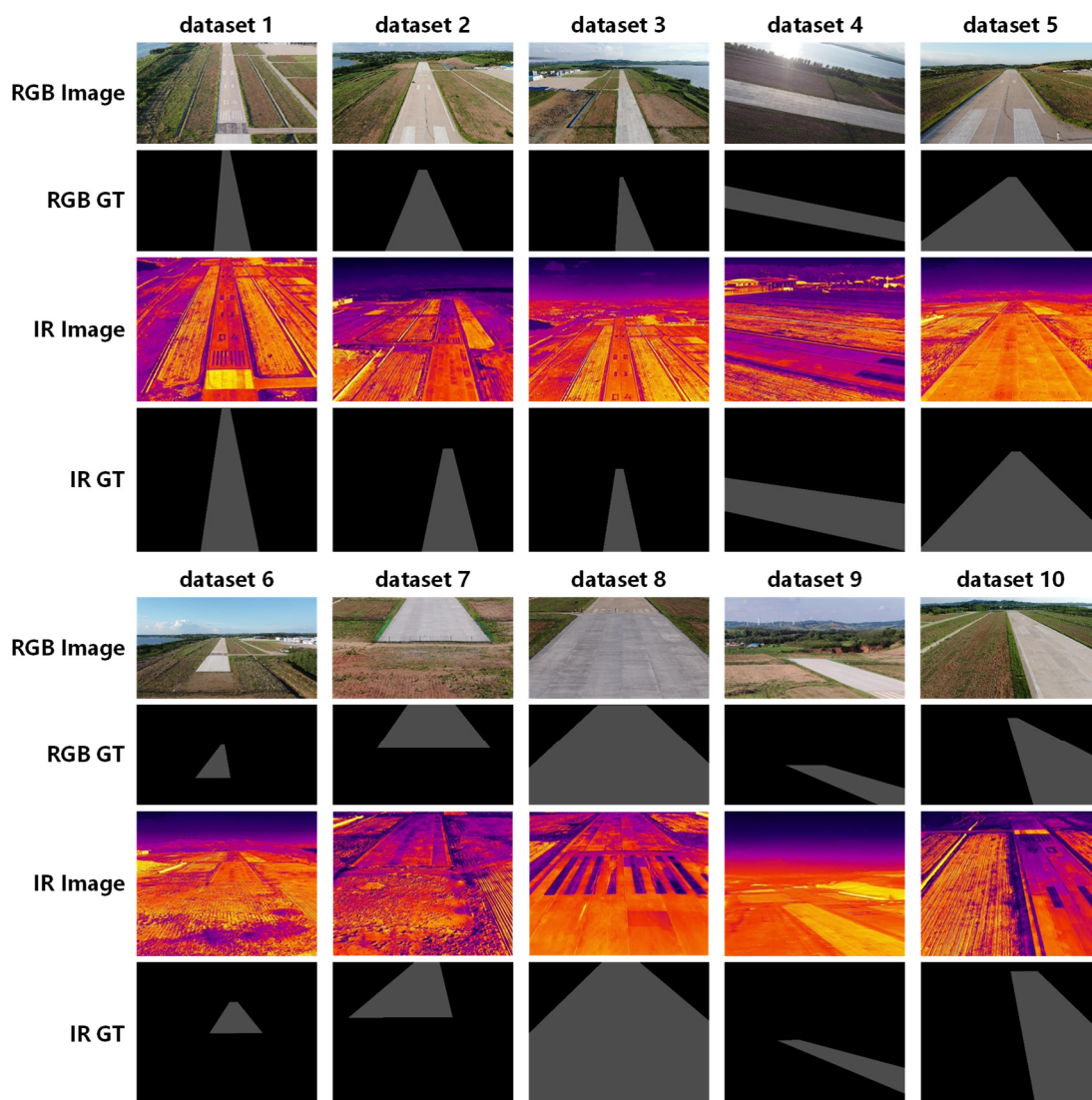
several non-aligned multimodal datasets [50,51] have been introduced. It has been shown that models trained on aligned datasets perform poorly on non-aligned data, such as our RDD5000 dataset. To address this challenge, we propose a practical multimodal SOD algorithm that eliminates the need for manual data alignment, thus avoiding the additional labor costs associated with image alignment.



**Figure 1.** The examples of several airport runway datasets: (a) airport runway datasets using remote sensing imagery [44–46], (b) airport runway datasets generated by simulation [47,48].

Third, although deep-learning-based dual-modal SOD methods generally offer high accuracy, they are often accompanied by large computational costs and bulky model sizes. This can become a performance bottleneck in practical applications, particularly in intelligent navigation tasks that require real-time processing. To address this challenge, we proposed a dual-modal fusion network with a lightweight MobileNetV2 backbone [53]. The design focuses on algorithm optimization and system integration, aiming to achieve more efficient real-time processing and decision support.

To effectively address the above limitations, this paper proposes a dual-modal cross-level fusion network (DCFNet) based on visible light and infrared RGB-IR images. The network exploits the complementary attributes of visible and infrared image data to facilitate reliable identification of salient targets on airport runways under various conditions.



**Figure 2.** The examples of the proposed RDD5000 dataset consisting of 10 sets of visible images (first row and fifth row), visible ground truth (second row and sixth row), infrared images (third row and seventh row), and infrared ground truth (fourth row and eighth row).

In summary, our main contributions are as follows:

- We created a large-scale visible–infrared dual-modal airport runway dataset (RDD5000) captured in real-world scenarios.
- We designed a lightweight deep learning model, DCFNet, for salient object detection, which is deployable in airborne environments.
- We proposed an independent–shared dual-stream encoder structure to efficiently fuse visible and infrared image information.
- We introduced a multi-scale, multi-level attention fusion decoder module that effectively improves the accuracy of the model.

The remainder of this paper is organized as follows: Section 2 reviews related works on airport runway detection and dual-modal SOD methods. Section 3 describes the proposed DCFNet architecture in detail. Section 4 presents the experimental results and analysis. Finally, Section 5 concludes the paper and discusses future work.

## 2. Related Works

### 2.1. Airport Runway Detection and Related Datasets

Airport runway detection has been an essential focus in computer vision, driving improvements in aviation safety and efficiency. Traditionally, various methods have been employed for runway detection, including edge detection [54–57], template-based approaches [58–60], and color-based segmentation [61,62] techniques. However, these approaches rely on low-level features, require the manual feature design for specific scenarios, and even involve extensive post-processing, which limits their effectiveness in complex environments.

In recent years, with the rapid development of deep learning, an increasing number of researchers have begun to apply deep learning methods to runway detection tasks. These approaches typically involve using deep convolutional neural networks, such as ResNet101 [63] and VGG-19 [64], to extract features from remote sensing images, or employing two-stage networks like R-CNN [65], YOLO series [66], and DeepLabv3 [45] models to identify candidate regions of interest in airport images. Despite the significant potential of deep learning, the lack of large-scale real-world airport runway datasets has become a major challenge limiting the application of deep learning in this field.

Although earlier studies [44–46] have constructed runway datasets, most of these datasets are based on remotely sensed imagery, which is mainly used for detecting runways from the Earth's point of view, and is not suitable for detection tasks in aircraft landing scenarios. Recently, some studies have attempted to construct runway datasets based on the aircraft landing perspective. For example, Cheng et al. [47] presented the BARS dataset, collected using the X-Plane simulation platform, which contains 10,256 images and 30,201 instances; Wang et al. [48] developed the X-Plane-based Runway Landing Dataset (RLD) for runway instance segmentation tasks; and Weng et al. [49] combined Unreal Engine 4 (UE4) with Cesium software (Cesium for Unreal) to create a dataset that closely resembles real airport runway scenes, called RAW. However, despite their efforts, these simulated datasets still lack sufficient real-world relevance, limiting their effectiveness when applied to actual environments.

We have constructed a benchmark dataset named RDD5000, which, unlike existing datasets, consists of 5000 pairs of RGB and infrared images directly captured from real-world scenarios. This dataset aims to advance the application of deep learning in SOD for airport runways.

### 2.2. Deep-Learning-Based Dual-Modal SOD Method

Currently, deep-learning-based dual-modal SOD methods have become mainstream. Recent works, such as [67–69], have explored advancements in fusion techniques, highlighting their effectiveness in improving object detection accuracy under challenging conditions. These methods include CNN-based dual-modal SOD approaches, Transformer-based methods, and the latest SAM (Segment Anything Model)-based approaches, along with various combinations of these methods. CNN-based models [70,71] typically employ specialized network architecture to fuse multimodal information, focusing on fusing features from different modalities for object detection.

The self-attention mechanism of Transformer models [72] can flexibly capture long-range dependencies between data, offering significant advantages over traditional CNNs when handling dual-modal SOD tasks. As a result, recent works [11,14,15,18,22,23,73–75] employed Transformer-based backbones for dual-modal SOD. However, Transformer architectures typically require large model sizes and more computational resources, which leads to challenges such as longer training times and slower inference speeds. Fortunately, the introduction

of variant models like Swin Transformer [11,14,23] has achieved a better balance between efficiency and performance.

Moreover, the emergence of models like the Segment Anything Model (SAM) has further advanced the development of dual-modal SOD technologies. SAM-based models [76,77] achieve higher precision in salient object detection by incorporating advanced image segmentation and object recognition techniques.

Datasets are the foundation of all dual-modal SOD methods. With the increasing demand for dual-modal SOD, several relevant datasets have been released. For example, the VT821 [29], VT1000 [30], and VT5000 [31] datasets provide visible–infrared image pairs, making them suitable for RGB-T dual-modal SOD research. Currently, many RGB-T SOD methods [8–23] have achieved outstanding results on these datasets. Additionally, datasets such as NJU2K [32], NLPR [33], STERE [34], and SIP [35] are also widely used in RGB-D SOD research. However, all of the datasets mentioned above are manually aligned, and methods based on these aligned datasets cannot directly handle misaligned dual-modal image pairs. The additional registration process not only increases manual labor but also imposes a significant computational burden.

In practical applications, misaligned multimodal images are the norm. Tu et al. [14] were among the pioneers in recognizing the misalignment issue in dual-modal SOD. They simulated weakly aligned data using affine transformations, which, while differing from real-world misalignment scenarios, still offered theoretical insights and practical value. Since the work of Tue et al., an increasing number of researchers have focused on the misalignment problem in multimodal data, leading to the development of non-aligned multimodal datasets [50,51] and the proposal of corresponding non-aligned dual-modal SOD models [50–52].

Song et al. [50] constructed the UAV RGB-T 2400 dataset for unaligned RGB-T salient object detection from drone perspectives and proposed the MROS model, which improved bimodal interaction and spatial alignment but had high computational costs, limiting real-time use. Wang et al. [51] developed the UVT2000 dataset with 2000 pairs of unaligned RGB and thermal images and introduced the SACNet model to improve unaligned RGB-T SOD. Lyu et al. [52] proposed AlignSal, an efficient Fourier filter network using contrastive learning to align modalities, reducing model complexity. However, these methods still suffer from high computational cost and large parameter sizes, limiting their use in real-time, low-power applications like airborne edge devices.

To address these challenges, this paper proposes a lightweight dual-modal SOD method for unaligned RGB-IR image pairs, focusing on detection accuracy and computational efficiency. The effectiveness of the approach is demonstrated on the RDD5000 dataset.

### 3. Proposed Method

#### 3.1. Overall Architecture

We adopt an encoder–decoder framework to address the problem of visible–infrared dual-modal SOD. Figure 3 illustrates the overall architecture of the proposed DCFNet, which consists of a combined independent–shared encoder network and a multi-scale, multi-level decoder network. The encoder extracts features from visible and infrared images, while the decoder fuses these features to produce the final saliency map.

**Independent–Shared Encoder Network.** DCFNet’s encoder consists of dual-modal independent encoding layers (IELs) and dual-modal shared encoding layers (SELs), forming a dual-stream convolutional neural network (CNN). In this structure, the IELs are responsible for extracting low-level features from visible light and infrared images, respectively, while the SELs enable feature matching between the two modalities in a unified space, generating high-level features. From the perspective of lightweight design, we selected Mo-

bileNetV2 [53] with a 5-stage structure as the encoder backbone, and optimally configured the number of independent and shared encoding stages in the two-stream network, which is finally configured as 2 for the independent layer and 3 for the shared layer, achieving an optimal balance between performance and efficiency.

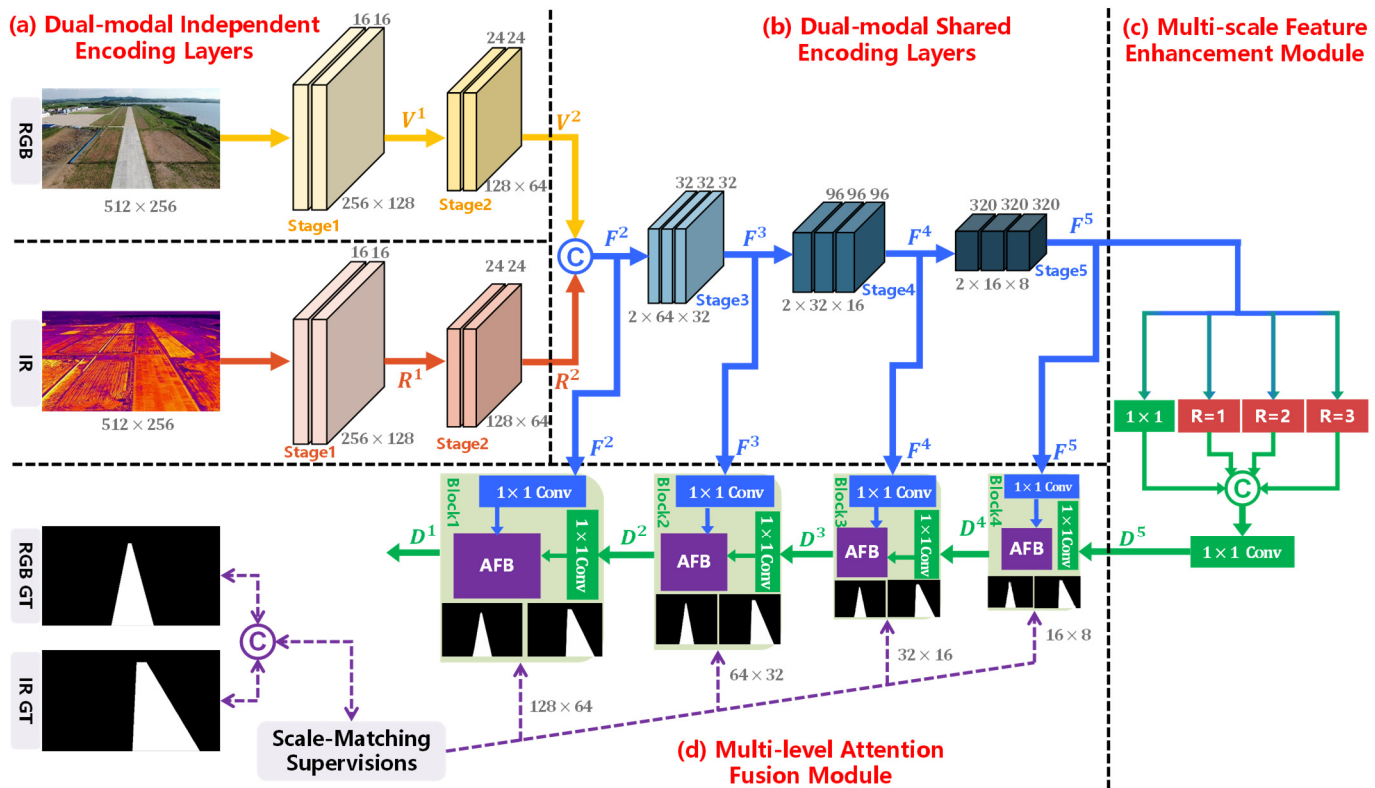


Figure 3. The overall architecture of the proposed DCFNet.

**Multi-Scale and Multi-Level Decoder Network.** The decoder network of DCFNet includes two key modules: the multi-scale feature enhancement module (MFEM) and the multi-level attention fusion module (MAFM). To effectively address potential resolution differences and misalignment between visible and infrared images, we propose the MFEM. This module extracts and enhances features from both modalities at multiple scales, enabling robust fusion even when the input image pairs exhibit resolution differences or misalignment. The MAFM, on the other hand, enhances the model's sensitivity to details and small targets by introducing an adaptive multi-level attention mechanism that fuses features from different levels in a bottom-up manner. Inspired by the modality-cooperative decision-making approach in VI Re-ID, we jointly consider the target annotations from both modalities and weight the fusion of detection results from visible and infrared images.

### 3.2. Independent–Shared Encoder Network

Considering that the visible and infrared image pairs input to our model are raw data directly captured from the onboard platform without manual alignment, significant cross-modal discrepancies exist. These discrepancies not only manifest as distinct differences in color, texture, and appearance between visible and infrared images but also include target variations caused by platform motion and viewpoint changes. From this perspective, the cross-modal challenges resemble those encountered in visible–infrared person re-identification (VI Re-ID) and cross-view matching tasks. Inspired by these tasks, we propose a dual-stream CNN architecture with both independent and shared feature

learning, which is designed to effectively handle the modality differences while allowing for effective cross-modality feature fusion.

A key assumption in our method is that the visible and infrared images must depict the same target, although precise registration is not required. This implies that the two modalities may capture the target at different times or from different perspectives. This assumption is motivated by the practical challenges associated with image registration, which is often computationally expensive and error-prone, especially in real-world scenarios where misalignment between modalities is common. By relaxing the requirement for precise alignment, our method significantly reduces the complexity of the dual-modal SOD pipeline. This not only improves the efficiency of the approach but also enhances its applicability to real-world scenarios where perfect inter-modal alignment is rarely achievable.

Specifically, the different modalities of visible and infrared images differ significantly in the low-level features (e.g., color, texture, edges, etc.), while in the high-level features, although visible and infrared images are different in appearance, the semantic information they express (e.g., shape, structure, category of the target, etc.) is usually similar. Therefore, we employ independent network structures for low-level feature extraction, which better captures the unique information of the respective modalities. A shared network structure is used on the high-level feature extraction to help the model fuse the semantic features of the two modalities and improve the cross-modal representation.

Feature extraction follows a 'first independent, then shared' approach. However, the choice of encoder backbone and the specific division between independent and shared layers has a direct impact on the performance and efficiency of the dual-stream encoder network. The following sections will describe each of these aspects in detail.

### 3.2.1. Lightweight Backbone

Some lightweight backbone networks, such as MobileNets [53,78], ShuffleNets [79,80], and EfficientNet [81], are specifically designed for environments with limited computational resources. MobileNetV2, with its innovative inverted residuals structure and bottleneck layer, is able to effectively preserve feature representation capabilities while maintaining high efficiency, making MobileNetV2 the lightweight backbone of choice for many CV tasks. We have chosen MobileNetV2 as the backbone network in our proposed model, given the real-time requirements of visual navigation tasks. The ablation study in Section 4.3 also compares the performance of DCFNet with different lightweight backbone networks, which confirms the rationale behind our choice of MobileNetV2.

### 3.2.2. Independent and Shared Layers Architecture

Another key factor of the dual-stream network is the design of the independent encoding layers (IELs) and shared encoding layers (SELs), which directly affects the performance, complexity, and efficiency of our model.

The IELs we designed are primarily responsible for extracting low-level features from the visible light and infrared modalities. Of course, the different modalities can adopt different network structures. The ablation experiments in Section 4.3.1 also compare the performance of different backbones in the IELs for the visible and infrared modalities. To simplify the description, we use the same network architecture for the IELs of both streams. The SELs designed by us are mainly responsible for the extraction and fusion of high-level features.

The division between low-level and high-level features is typically determined by the depth of the convolutional layers in the network. In the case of the ResNets, for example, they can be divided into five main stages, each consisting of a number of convolutional layers, with the lower features coming from the early stages of the network and the higher



features coming from the later stages of the network. The layer configuration of the IELs and SELs needs to be balanced according to the performance and efficiency of the task.

We also divide MobileNetV2 into five stages (to adapt it to the SOD task, we remove the global average pooling layer and the final fully connected layer from the backbone) and compare its structure with that of ResNet18, as shown in Table 1. Based on insights from Liu et al. [43] and the experimental results in Section 4.3.2, we have determined the optimal configuration: IELs for the first two stages of the backbone (stages 1–2) and SELs for the last three stages (stages 3–5). This configuration provides the best balance between performance and efficiency.

**Table 1.** Configuration of dual-modal independent and shared encoding layers. #Params. (M) denotes the number of parameters in millions.

Layers	Output Size	ResNet18			MobileNetV2		
		Channels	RGB Image	IR Image	Channels	RGB Image	IR Image
Input	$512 \times 256$	3	/	/	3	/	/
Stage 1	$256 \times 128$	64	$7 \times 7, 64,$ stride = 2	$7 \times 7, 64,$ stride = 2	16	$3 \times 3, 32,$ stride = 2 Dwise $3 \times 3, 16$	$3 \times 3, 32,$ stride = 2 Dwise $3 \times 3, 16$
Stage 2	$128 \times 64$	64	$3 \times 3, \max$ pool, stride = 2 $[3 \times 3, 64] \times 4$	$3 \times 3, \max$ pool, stride = 2 $[3 \times 3, 64] \times 4$	24	Dwise $3 \times 3,$ 24, stride = 2 Dwise $3 \times 3, 24$	Dwise $3 \times 3,$ 24, stride = 2 Dwise $3 \times 3, 24$
Stage 3	$64 \times 32$	128	$[3 \times 3, 128] \times 4$		32	Dwise $3 \times 3, 32,$ stride = 2 [Dwise $3 \times 3, 32] \times 2$	
Stage 4	$32 \times 16$	256	$[3 \times 3, 256] \times 4$		96	Dwise $3 \times 3, 64,$ stride = 2 [Dwise $3 \times 3, 64] \times 3$ [Dwise $3 \times 3, 96] \times 3$	
Stage 5	$16 \times 8$	512	$[3 \times 3, 512] \times 4$		320	Dwise $3 \times 3, 160,$ stride = 2 [Dwise $3 \times 3, 160] \times 2$ [Dwise $3 \times 3, 320]$	
#Params. (M)		11.7			3.4		

In our proposed model, the IELs for the visible light and infrared branches, as well as the SELs, can utilize either the same or different lightweight backbone networks. For clarity, we describe the processing steps assuming MobileNetV2 as the backbone network.

Visible Branch in the IELs:

- Input: RGB image of size  $512 \times 256$ .
- Feature Extraction: The first two layers of MobileNetV2 are used to extract low-level features at two different scales:

Stage 1: Output feature of size  $256 \times 128$  with 16 channels, denoted as  $V^1$ .

Stage 2: Output feature of size  $128 \times 64$  with 24 channels, denoted as  $V^2$ .

Infrared Branch in the IELs:

- Input: IR image of size  $512 \times 256$ .
- Feature Extraction: Similar to the visible branch, the first two layers of MobileNetV2 extract low-level features:

Stage 1: Output feature of size  $256 \times 128$  with 16 channels, denoted as  $R^1$ .

Stage 2: Output feature of size  $128 \times 64$  with 24 channels, denoted as  $R^2$ .

Feature Fusion:

The feature maps  $V^2$  and  $R^2$  are concatenated to form the joint feature  $F^2$ :

$$F^2 = \text{Concat}(V^2, R^2), \quad (1)$$

Shared Encoding Layers (SELS):

The concatenated feature  $F^2$  is then passed through the shared encoding layers to extract high-level features:

Stage 3: Output feature of size  $64 \times 32$  with 32 channels, denoted as  $F^3$ .

Stage 4: Output feature of size  $32 \times 16$  with 96 channels, denoted as  $F^4$ .

Stage 5: Output feature of size  $16 \times 8$  with 320 channels, denoted as  $F^5$ .

Handling Different Backbones:

If the IELs for visible light and infrared use different backbones (e.g., MobileNetV2 for visible light and ResNet-18 for infrared), the infrared feature  $R^2$  is first compressed to 32 channels using a  $1 \times 1$  convolution before concatenating with  $V^2$ :

$$F^2 = \text{Concat}(V^2, \text{Conv}_1BR(R^2)), \quad (2)$$

The resulting feature  $F^2$  is then passed through the SELs for high-level feature extraction.

### 3.3. Multi-Scale and Multi-Layer Decoder Network

Although there has been a substantial amount of research on decoders [26,38,82–84], most approaches suffer from complex designs and low efficiency. Therefore, this paper proposes a lightweight multi-scale fusion decoder that not only effectively integrates multi-level features but also emphasizes improving efficiency. The decoder consists of two key modules: the MFEM and MAFM.

#### 3.3.1. Multi-Scale Feature Enhancement Module

Compared to large CNNs, lightweight backbone networks have inherent disadvantages, especially in complex scenarios. For example, the MobileNetV2 model only performs a  $3 \times 3$  convolution operation on a 320-dimensional channel. To address this challenge, this paper proposes a multi-scale feature enhancement module (MFEM), as shown in Figure 3c, which enhances the semantic feature  $F^5$  output by the last layer of the encoder, resulting in an enhanced feature  $D^5$  of the same size as  $F^5$ . This design not only improves the feature representation, but also provides an additional input  $D^5$  (in addition to  $F^5$ ) for the subsequent MAFM Block 4. This design ensures structural consistency across the four attention fusion modules, improving model stability.

The MFEM employs depthwise separable convolutions [85] to process the input feature maps in parallel using three different expansion rate strategies [85] to enhance the capture of contextual information. Additionally, a  $1 \times 1$  convolution layer is used to perform a linear transformation of the feature map, further improving the feature representation. Finally, the module sums the four output feature maps at different scales pixel-wise and integrates them further through convolution, batch normalization (BN) [86], and ReLU activation functions [87], enhancing the non-linear expression. The overall process can be formally expressed as:

$$D^5 = \text{Conv}_1BR \left( \sum_{i=1}^3 \text{DWConv}_3BR^{(i)}(F^5) + \text{Conv}_1BR(F^5) \right), \quad (3)$$

where  $F^5$  and  $D^5$  represent the input and output feature maps, respectively.

$\text{Conv}_1BR(\cdot)$  represents a  $1 \times 1$  convolution followed by batch normalization and ReLU activation.

$\text{DWConv}_3BR^{(i)}(\times)$  refers to the  $i$ -th depthwise separable convolution with a  $3 \times 3$  kernel and a dilation rate  $i$ , followed by batch normalization and ReLU activation.

The lightweight nature of the MFEM is primarily attributed to the combination of depthwise separable convolutions and dilation convolution strategies. Depthwise separable

convolutions decompose traditional convolutions into depthwise convolutions and point-wise convolutions, significantly reducing the number of parameters and computational complexity, thus improving the network's efficiency. At the same time, dilation convolutions introduce holes (dilations) into the convolutional kernels, allowing the network to capture a larger receptive field without adding extra computational overhead.

### 3.3.2. Multi-Level Attention Fusion Module

Many SOD models carefully design the encoder to extract dual-modal features, while the decoder typically employs simple upsampling strategies to recover feature maps. Differentiating from the existing methods, we design a multi-level attention fusion module (MAFM) in the decoder, positioned after the MFEM. The MAFM not only considers the complementary nature of features from different network layers but also integrates features from various scales and receptive fields. More importantly, we introduce the contribution weights of features from different levels, combined with multiple attention mechanisms (attention modules, AM). By guiding low-level detail features with high-level semantic features, the MAFM progressively refines the salient objects, optimizing them from coarse localization to precise boundaries.

The proposed framework of the MAFM is illustrated in Figure 4. The input to the MAFM consists of two components: (1) high-level feature maps  $F^i$  (where  $i = 2, 3, 4, 5$ ) output laterally by the encoding module; and (2) low-level feature maps  $D^i$  (where  $i = 2, 3, 4$ ) produced by the previous MAFM stage, along with the enhanced feature map  $D^5$  (where  $i = 5$ ) specifically generated by the MFEM. As mentioned in the previous section, the MFEM is designed not only to enhance the features of the lightweight encoding network, but also to provide inputs to MFEM Block 4 that lack support from the upper level, thus allowing MAFM Block 4 to also fuse features from different levels (i.e.,  $F^5$  and  $D^5$ ). The joint design of the MFEM and MAFM ensures the consistency of the structure of MAFM Blocks 1–4, simplifies the process of multi-level feature fusion and improves the maintainability of the system, and allows the flexibility to stack, expand, or replace MAFMs as required.

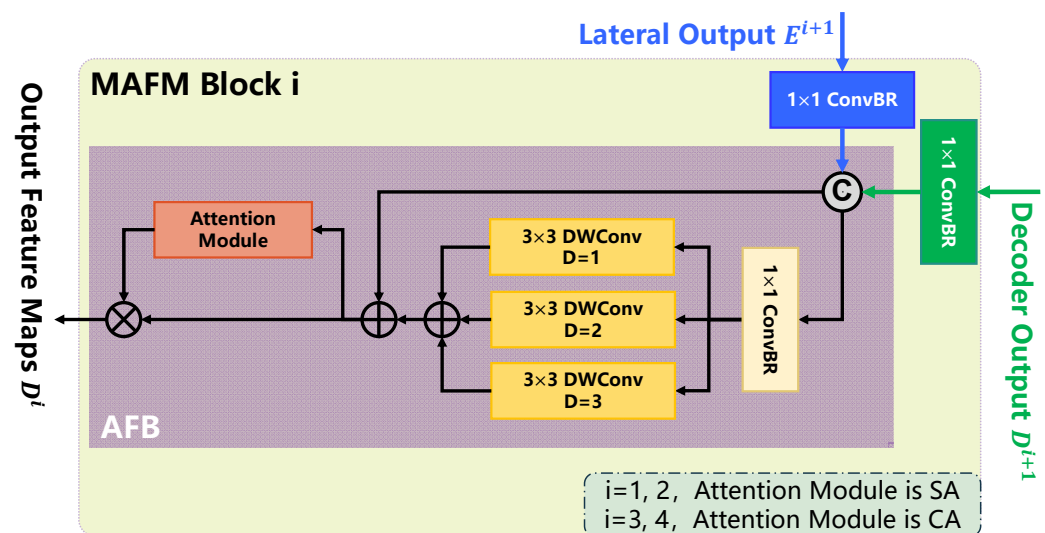


Figure 4. The proposed MAFM framework.

Figure 4 clearly depicts the details of feature fusion and attention allocation in the MAFM. Our feature fusion process is implemented through the attention fusion block (AFB). First, we concatenate the feature maps  $D^i$  and  $F^i$  from different layers along the channel dimension, rather than simply adding them pixel-wise. This approach aims to preserve the complete information from both different-level features to the greatest extent. To maintain

dimensional consistency and improve model efficiency, we apply a  $1 \times 1$  convolution layer and batch normalization before concatenation, reducing the number of channels by half. The concatenated feature map  $DF^i$  is then passed into the AFB for fusion processing.

$$DF^i = \text{Concat}\left(\text{Conv}_1\text{BR}\left(D^i\right), \text{Conv}_1\text{BR}\left(F^i\right)\right) \quad (4)$$

AFB first preprocesses the input feature map  $DF^i$  using a  $1 \times 1$  convolutional layer, followed by batch normalization to enhance the non-linear representation of the features. Then, the feature map  $DF^i$  is passed through multiple depthwise separable convolution layers ( $DWConvBR$ ) with different dilation rates ( $D$ ) in parallel to capture contextual information at different scales. The depthwise separable convolutions use  $3 \times 3$  kernels, and the dilation rates  $D$  are set to 1, 2, 3, or 1, 3, 5 to accommodate feature maps at different depths. To facilitate network convergence and alleviate the vanishing gradient problem, the AFB introduces residual connections by adding the input feature  $DF^i$  to the features processed in parallel, resulting in the fused feature map  $DF^m$ . This process can be formally described as:

$$DF^m = \sum_{i=1}^3 DWConv_3BR^{(i)}\left(\text{Conv}_1\text{BR}\left(DF^i\right)\right) \oplus DF^i \quad (5)$$

Finally, we apply an attention mechanism to the preliminary fused feature  $DF^m$  to further optimize the representation of the saliency prediction map. In our model, we employ two different attention mechanisms based on the different levels of feature maps. Specifically, for high-level feature maps ( $i = 3, 4$ ), where the number of channels is relatively large, we use a channel attention (CA) mechanism to emphasize important channels and suppress noisy or redundant ones. For low-level feature maps ( $i = 1, 2$ ), where the spatial dimensions are gradually recovering, we apply a spatial attention (SA) mechanism to better focus on regions containing effective details, resulting in clearer saliency boundaries. To improve efficiency, we incorporate a CBAM (convolutional block attention module) [88], an easy-to-plug-and-play module that integrates both spatial and channel attention. It is worth noting that the model framework is highly flexible, allowing for the use of alternative spatial or channel attention mechanisms based on specific requirements.

In the equation,  $AM(\cdot)$  represents the attention mechanism.

$$F_o^{i-1} = AM(DF^m) \otimes DF^m \quad (6)$$

In summary, the MAFM enhances the expressive power of the saliency prediction map through carefully designed non-linear transformations, multi-scale receptive field capture, residual connections, and flexible application of attention mechanisms at different layers, all while maintaining computational efficiency.

### 3.4. Loss Function

The outputs of MAFM Blocks 1–4 are denoted as  $D^i$  (for  $i = 1, \dots, 4$ ). These outputs are first passed through a  $1 \times 1$  convolutional layer to reduce the number of channels to 1. Then, bilinear interpolation is applied to resize the feature maps to match the original input image dimensions. Finally, a Sigmoid function is applied to normalize the pixel values to the range  $[0, 1]$ , resulting in four saliency prediction maps  $S^i$  (for  $i = 1, \dots, 4$ ), as shown in Figure 3. For both visible and infrared modalities, the ground truth is derived by concatenating the visible light and infrared ground truth, and is denoted as  $G$ . This cascade supervision mechanism allows the model to better integrate information from both visible and infrared images, thereby improving the accuracy of saliency object detection.

$$G = \text{Concat}(G_{RGB}, G_{IR}) \quad (7)$$

In this section, we define the objective function used to train the proposed model. The objective function combines binary cross-entropy (BCE) loss and Dice loss to effectively optimize the saliency object detection task. Therefore, the loss for each saliency prediction map  $S^i$  can be expressed as:

$$L(S^i, G) = L_{BCE}(S^i, G) + L_{Dice}(S^i, G), \quad (8)$$

where  $S^i$  and  $G$  represent the saliency prediction map and the ground truth, respectively.  $L_{BCE}(S^i, G)$  represents the binary cross-entropy loss, and  $L_{Dice}(S^i, G)$  represents the Dice loss.

BCE loss is widely used in SOD tasks, as it computes the classification error for each pixel in the image but does not account for the global structure of the image. Therefore, we incorporate the Dice loss in the objective function to measure the spatial overlap between the predicted result and the true target. By using both loss functions simultaneously, the model is more comprehensively constrained during training. The specific loss functions are defined as follows:

$$L_{BCE}(S^i, G) = -\sum G \log(S^i) + (1 - G) \log(1 - S^i), \quad (9)$$

$$L_{Dice}(S^i, G) = 1 - \frac{2 \times \sum(S^i \times G) + \epsilon}{\sum(S^i) + \sum(G) + \epsilon}, \quad (10)$$

where  $\epsilon$  (set to  $1 \times 10^{-5}$ ) is a small constant used to avoid division by zero.

To optimize the model, we employ multi-scale supervision to accelerate convergence. This approach improves predictive accuracy by delivering gradient signals at various levels of abstraction. During implementation, we compute the loss  $L(S^i, G)$  for each channel in the model's output and aggregate the losses across all channels. The overall loss is the sum of these individual losses:

$$L(S, G) = \sum_{i=1}^4 L(S^i, G) \quad (11)$$

During training, to better balance the ground truth information from both visible and infrared modalities, we compute separate losses  $L(S, G)$  for each modality and introduce a visible light weight parameter  $\alpha$  for weighted summation. The final objective function is:

$$L = \alpha \times L_{RGB}(S, G) + (1 - \alpha) \times L_{IR}(S, G) \quad (12)$$

We performed ablation experiments under different weight settings (see Section 4.3.5) to determine the optimal weight value.

## 4. Experimental Results and Analysis

### 4.1. Experiment Setup

#### 4.1.1. Datasets

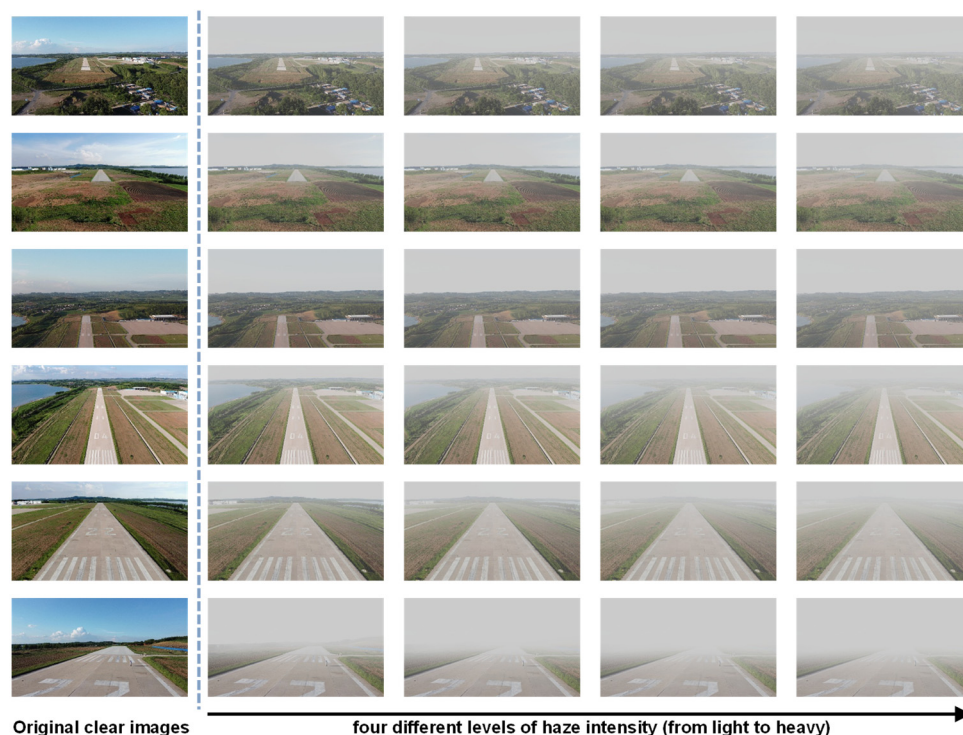
We have constructed a specialized dataset for airport runway detection, named the Runway Detection Dataset (RDD5000), which is a significant contribution of our research. The RDD5000 dataset was collected using a DJI drone equipped with both visible light ( $1920 \times 1080$  resolution) and infrared ( $640 \times 512$  resolution) cameras. The dataset consists of 5000 pairs of visible and infrared images, which were carefully selected from real-world airport runway scenes captured at Shenyang Faku Caihu Airport. A total of 16 GB of video data was recorded during the collection process.

The dataset includes 10 sequences of clear visible light images and 10 sequences of infrared images, each containing 500 images. Due to differences in shooting conditions and

camera parameters, the visible light and infrared images in the RDD5000 dataset are not aligned (as shown in Figure 2). This phenomenon is consistent with real-world conditions, as perfectly aligned dual-modal images are difficult to obtain directly. Currently, existing dual-modal datasets are typically aligned manually, which is a labor-intensive process.

All images were manually annotated with ground truth labels for runway detection. Although the dataset was collected using a UAV, our proposed method is designed to be applicable to real-world navigation systems. We acknowledge potential differences between drone-collected data and data from actual navigation systems, particularly in terms of environmental variations. For example, our drone dataset was collected under clear weather conditions, whereas real navigation systems must operate in diverse and complex environments, including adverse conditions such as fog, rain, or haze.

To address this gap and enhance the dataset's realism, we employed a haze generation method based on the MiDaS algorithm [89] to synthesize hazy conditions from the visible light data collected by the drone. This approach simulates challenging weather scenarios, thereby improving the robustness and generalizability of our method for real-world navigation tasks. Specifically, we simulated four different levels of haze intensity (from light to heavy) for each image in the 10 sequences, resulting in four additional sets of 5000 images each. Figure 5 illustrates examples of haze image generation, showing the original visible light images and their corresponding synthetic hazy versions at four intensity levels (light to heavy). These variations allow for comprehensive training and evaluation of the model under various low-visibility scenarios.



**Figure 5.** Examples of haze image generation.

To evaluate the effectiveness of our method, we conducted experiments on the RDD5000 dataset and compared the results with the latest non-aligned dual-modal SOD algorithms.

#### 4.1.2. Evaluation Metrics

We adopted four popular evaluation metrics to comprehensively evaluate the performance of different methods, including the maximum F-measure ( $F_\beta$ ) [90], mean absolute error (MAE), E-measure ( $E_\zeta$ ) [91], and S-measure ( $S_m$ ) [92].

##### 1. F-Measure ( $F_\beta$ ).

$F_\beta$  is a combined metric that comprehensively considers both precision and recall, and is commonly used to evaluate the accuracy of SOD results. It is formulated as:

$$F_\beta = \frac{(1 + \beta^2) \times Precision \times Recall}{\beta^2 \times Precision + Recall}, \quad (13)$$

where  $\beta^2$  is typically set to 0.3, as suggested in [6], to emphasize precision.  $F_\beta$  is a ‘larger is better’ type of metric. In our paper, we report the maximum  $F_\beta$  under different binarization thresholds.

##### 2. Mean Absolute Error (MAE).

The MAE is defined as the average pixel-wise absolute difference between the predicted saliency map and the corresponding ground truth. The calculation formula is:

$$MAE = \frac{1}{W \times H} \sum_{i=1}^W \sum_{j=1}^H |S_{ij} - G_{ij}|, \quad (14)$$

where  $W$  and  $H$  represent the width and height of the image, respectively, with  $W \times H$  being the total number of pixels in the image.  $S_{ij}$  and  $G_{ij}$  are predicted saliency map and ground truth, respectively. MAE is a ‘smaller is better’ type of metric.

##### 3. E-measure ( $E_\zeta$ ).

$E_\zeta$  [91] is a recently proposed enhanced alignment metric. This metric, based on cognitive vision, jointly captures image-level statistics and local pixel matching information. It is computed by:

$$E_\zeta = \frac{1}{W \times H} \sum_{i=1}^W \sum_{j=1}^H f\left(\frac{2\varphi_G \circ \varphi_F}{\varphi_G \circ \varphi_G + \varphi_F \circ \varphi_F}\right), \quad (15)$$

where  $\varphi$  is the bias matrix, representing the distance between each pixel-wise value of the ground truth and its image-level mean.  $f(\cdot)$  denotes the enhanced-alignment matrix [91].  $E_\zeta$  is a ‘larger is better’ type of metric.

##### 4. Structure-Measure ( $S_m$ ) [92].

The  $S_m$  simultaneously evaluates region-aware and object-aware structural similarity between a saliency map and the ground truth.  $S_m$  is calculated as:

$$S_m = \alpha \times S_0 + (1 - \alpha) \times S_r, \quad (16)$$

where  $\alpha \in [0, 1]$ , and we set  $\alpha = 0.5$  as recommended in [92].  $S_0$  and  $S_r$  denote the object-aware and region-aware structure similarities, respectively.  $S_m$  is a ‘larger is better’ type of metric.

In addition, we evaluate all methods by PR curves via binarizing the saliency map with a threshold sliding from 0 to 255 and then comparing the binary maps with the ground truth.

#### 4.1.3. Implementation Details

##### 1. Training Set and Data Augmentation.

To address potential resolution differences between visible and infrared images, both modalities are resized to a uniform resolution of  $512 \times 256$  pixels before being fed into the network. For data augmentation, we use horizontal flips, random cropping, and multi-scale operations to process input image pairs. These steps enhance the diversity and robustness of the training data, ensuring that the extracted features from both modalities are aligned and suitable for fusion.

## 2. Parameter Settings.

We utilize the popular PyTorch framework to implement the proposed network. Unless otherwise specified, we use MobileNetV2 as our backbone. During the training phase, we apply the Adam algorithm with a momentum of 0.9 and weight decay of  $1 \times 10^{-4}$  to optimize our network. The batch size is set to 4, and the number of epochs is set to 150. Training was performed on a desktop equipped with an Intel Core i7-13700 CPU, an NVIDIA RTX 4060 GPU, and 16 GB of RAM, with a total training time of approximately 14 h. For image pairs with an input size of  $512 \times 256$ , the average inference time is 0.0065 s (running at about 155 FPS).

### 4.2. Comparison with SOTA Methods on the RDD5000

We compare our method with two state-of-the-art non-aligned dual-modal SOD methods, DCNet [14] and SACNet [51], on the provided RDD5000 dataset.

#### 4.2.1. Quantitative Comparisons

As shown in Table 2 and Figure 6, we provide a comprehensive quantitative and visual comparison of our method (DCFNet) against two other unaligned-based methods (DCNet and SACNet) on the RDD5000 dataset.

**Table 2.** Quantitative comparison with two other unaligned-based methods on RDD5000 datasets.  $\uparrow$  and  $\downarrow$  indicate ‘larger is better’ and ‘smaller is better’, respectively. #Params. (M) denotes the number of parameters in millions. The best results are highlighted in **bold**.

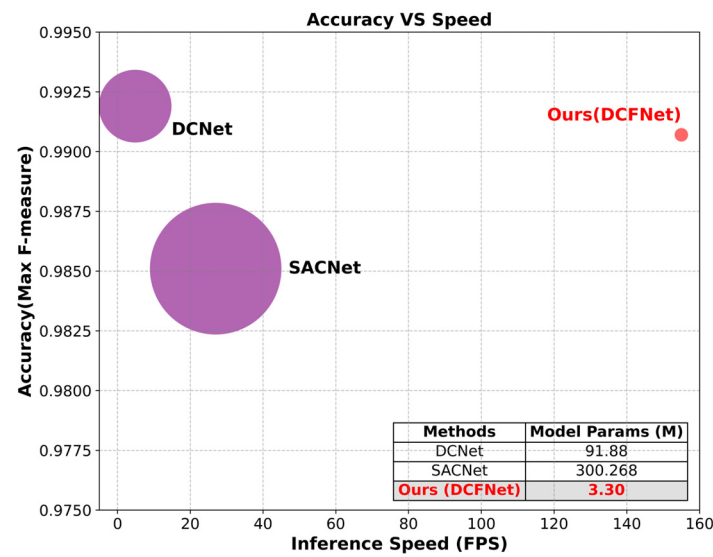
Metric		DCNet <sub>23</sub> [14]	SACNet <sub>24</sub> [51]	Ours (DCFNet)
Backbone		Res2Net-50	SwinB	MobileNetV2
Input size		$352 \times 352$	$384 \times 384$	$512 \times 256$
#Params. (M) $\downarrow$		91.88	300.268	<b>3.30</b>
FPS $\uparrow$		4.878	27	<b>155</b>
FLOPs (G) $\downarrow$		207.31	143.787	<b>2.548</b>
RDD5000 Visible Light Test Dataset	$F_\beta \uparrow$	<b>0.9919</b>	0.9851	0.9907
	MAE $\downarrow$	<b>0.0034</b>	0.0092	0.0053
	$E_\zeta \uparrow$	<b>0.9964</b>	0.9879	0.9939
	$S_m \uparrow$	0.9662	0.9663	<b>0.9780</b>
RDD5000 Infrared Test Dataset	$F_\beta \uparrow$	0.6773	0.8646	<b>0.9881</b>
	MAE $\downarrow$	0.3470	0.1856	<b>0.0108</b>
	$E_\zeta \uparrow$	0.4233	0.7544	<b>0.9885</b>
	$S_m \uparrow$	0.3933	0.7156	<b>0.9001</b>

In the RDD5000 visible light dataset, DCFNet achieves competitive results, with an  $F_\beta$  of 0.9907, MAE of 0.0053, and  $E_\zeta$  of 0.9939. Although slightly behind DCNet in terms of  $F_\beta$  (0.9919) and  $E_\zeta$  (0.9964), DCFNet outperforms SACNet (0.9851, 0.0092, 0.9879) across all three metrics. Additionally, DCFNet excels in the S-measure, scoring 0.9780, higher than both DCNet (0.9662) and SACNet (0.9663), suggesting its superior overall performance in the visible dataset.

In the RDD5000 infrared dataset, DCFNet demonstrates a clear advantage over both DCNet and SACNet. It achieves a remarkable  $F_\beta$  of 0.9881, MAE of 0.0108, and  $E_\zeta$  of 0.9885, far surpassing the scores of DCNet ( $F_\beta = 0.6773$ , MAE = 0.3470,  $E_\zeta = 0.4233$ ) and SACNet



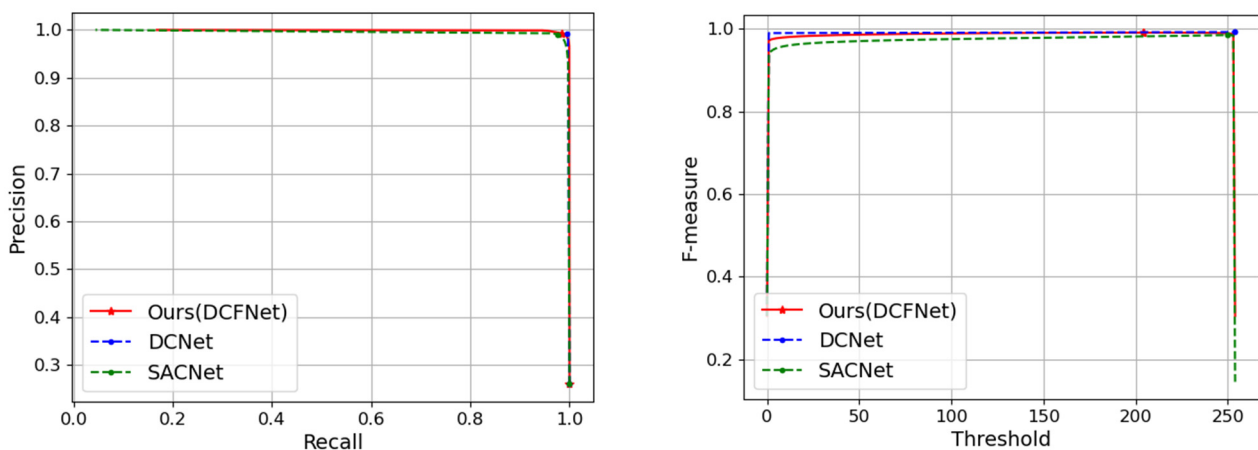
( $F_{\beta} = 0.8646$ ,  $MAE = 0.1856$ ,  $E_{\xi} = 0.7544$ ). This highlights DCFNet's robust performance in handling unaligned bimodal datasets, particularly in the infrared modality where both DCNet and SACNet struggle.



**Figure 6.** Visual comparison of speed and accuracy on the RDD5000 datasets.

In terms of real-time performance, DCFNet achieves an impressive 155 FPS, significantly outperforming DCNet (4.878 FPS) and SACNet (27 FPS), demonstrating its high computational efficiency and suitability for real-world applications.

In addition, Figure 7 shows the overall evaluation results of the PR curve and F-measure curve for our method and comparative methods on the RDD5000 visible dataset. As seen from the figure, the performance curve of the proposed method demonstrates superior overall distribution and trend, further validating the significant advantage of DCFNet in terms of performance.

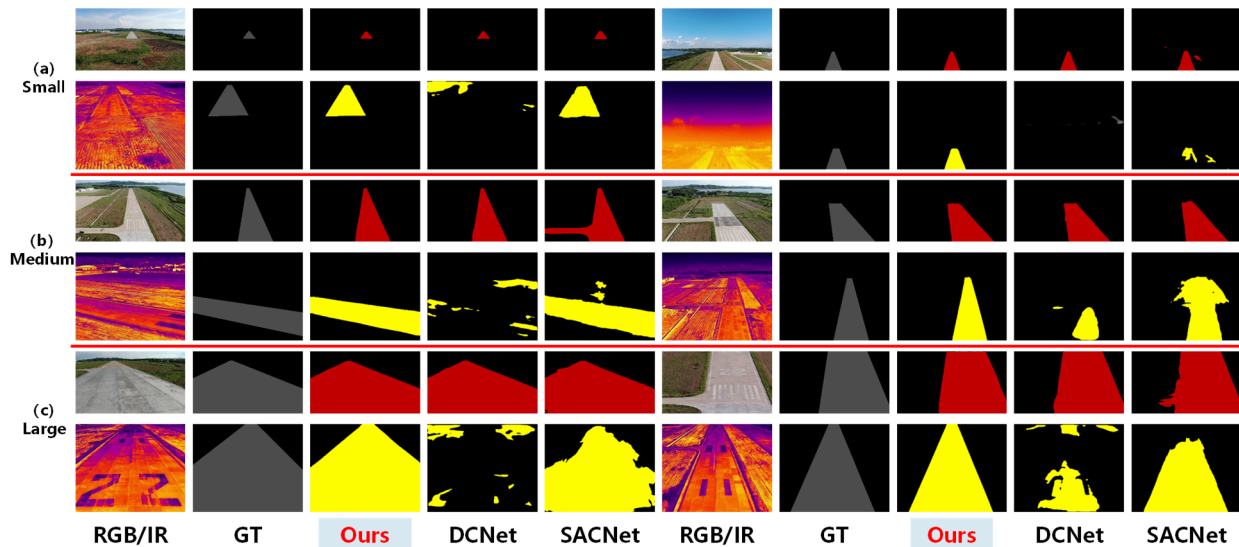


**Figure 7.** PR curves and threshold F-measure curves (from left to right) of different models on the RDD5000 visible dataset.

#### 4.2.2. Qualitative Comparison

Figure 8 provides a visual comparison of the saliency maps predicted by our method and other competing methods. The examples in Figure 8 are taken from the RDD5000 dataset and represent several challenging scenarios, including small objects (Rows 1 and 2), medium objects (Rows 3 and 4), and large objects (Rows 5 and 6). From these intuitive views, it is evident that our method performs best in terms of completeness and clarity.

Furthermore, the object boundaries predicted by our method are clearer and sharper than those predicted by other methods. These results demonstrate the effectiveness of our approach. These visual comparisons further validate the effectiveness and robustness of DCFNet in dealing with various challenging scenarios.



**Figure 8.** Visual comparisons with other SOTA methods under different challenging scenarios, including small objects (Rows 1 and 2), medium objects (Rows 3 and 4), and large objects (Rows 5 and 6).

#### 4.3. Ablation Studies

Section 3 provides a detailed description of the DCFNet network architecture and its components. We perform a series of ablation studies based on the RDD5000 dataset to evaluate the contribution of each key component in our method. In each ablation experiment, only one component is modified, and the same experimental settings as those in Section 4.1.3 are used. The specific ablation experiments are as follows:

##### 4.3.1. Different Backbone Designs in the Encoder

We analyzed the impact of different backbone network designs in the DCFNet encoder component. Specifically, we modified the backbone networks for the visible and infrared independent encoding layers, as well as the shared encoding layer. In the “Default” configuration, MobileNetV2 was used as the backbone network for both the visible and infrared independent and shared encoding layers, and this was used as the baseline. We also explored several alternative backbone configurations, as detailed in Table 3, which involve combinations of independent and shared encoding layers using different backbones. For example, the “Mobile + Res + Mobile” configuration indicates that the visible independent encoding layer uses MobileNetV2, the infrared independent encoding layer uses ResNet18, and the shared encoding layer uses MobileNetV2. Other similar backbone configurations follow the same pattern.

Table 3 evaluates the performance of different backbone designs in the encoder. The default configuration, which uses MobileNetV2 as the backbone, achieves the best performance on both the visible and infrared datasets, with  $F_\beta$  scores of 0.9907 and 0.9866, respectively. This confirms the effectiveness of MobileNetV2 for dual-modal feature extraction and fusion.

**Table 3.** Evaluation of different backbone network designs in the encoder.  $\uparrow$  and  $\downarrow$  indicate ‘larger is better’ and ‘smaller is better’, respectively. The best results are highlighted in **bold**.

Method	RDD5000 Visible Light Test Dataset				RDD5000 Infrared Test Dataset			
	MAE $\downarrow$	$F_\beta$ $\uparrow$	$E_\zeta$ $\uparrow$	$S_m$ $\uparrow$	MAE $\downarrow$	$F_\beta$ $\uparrow$	$E_\zeta$ $\uparrow$	$S_m$ $\uparrow$
Default	<b>0.0053</b>	<b>0.9907</b>	0.9939	<b>0.9780</b>	<b>0.0117</b>	<b>0.9866</b>	<b>0.9868</b>	<b>0.9072</b>
Res + Res + Res	0.0092	0.9848	0.9907	0.9545	0.0280	0.9666	0.9701	0.8529
Shuffle + Shuffle + Shuffle	0.0056	0.9901	0.9939	0.9625	0.0256	0.9711	0.9713	0.8617
Shuffle + Shuffle + Mobile	<b>0.0053</b>	0.9902	<b>0.9942</b>	0.9625	0.0232	0.9746	0.9745	0.8710
Mobile + Res + Mobile	<b>0.0053</b>	0.9904	0.9940	0.9634	0.0140	0.9834	0.9848	0.8922
Res + Mobile + Mobile	0.0056	0.9905	0.9939	0.9627	0.0159	0.9818	0.9817	0.8892
Mobile + Mobile + Res	0.0065	0.9893	0.9928	0.9608	0.0346	0.9549	0.9598	0.8432

#### 4.3.2. Independent and Shared Encoding Layers Configuration

The performance of the DCFNet model is highly influenced by the balance between independent and shared encoding layers. Specifically, the total number of independent and shared layers must be equal to the total number of layers in the backbone network (e.g., MobileNetV2 has five layers). Consequently, increasing the number of independent layers reduces the number of shared layers, and vice versa. This complementary relationship highlights the importance of selecting the optimal balance to achieve the best model performance.

For MobileNetV2, we established the default configuration with two independent layers and three shared layers as the baseline. To assess the impact of different layer configurations, we varied the number of independent and shared layers (while keeping the total at five) and evaluated their performance, as shown in Table 4.

**Table 4.** Evaluation of independent versus shared coding layer settings.  $\uparrow$  and  $\downarrow$  indicate ‘larger is better’ and ‘smaller is better’, respectively. #Params. (M) denotes the number of parameters in millions. The best results are highlighted in **bold**.

Method	#Params. (M) $\downarrow$	RDD5000 Visible Light Test Dataset				RDD5000 Infrared Test Dataset			
		MAE $\downarrow$	$F_\beta$ $\uparrow$	$E_\zeta$ $\uparrow$	$S_m$ $\uparrow$	MAE $\downarrow$	$F_\beta$ $\uparrow$	$E_\zeta$ $\uparrow$	$S_m$ $\uparrow$
IL 1, SL 4	3.29 M	<b>0.0047</b>	<b>0.9913</b>	<b>0.9945</b>	0.9652	0.0150	0.9833	0.9827	0.8892
IL 2, SL 3	3.30 M	0.0053	0.9907	0.9939	<b>0.9780</b>	<b>0.0108</b>	<b>0.9881</b>	<b>0.9885</b>	<b>0.9001</b>
IL 3, SL 2	3.33 M	0.0055	0.9903	0.9939	0.9626	0.0150	0.9828	0.9835	0.8933
IL 4, SL 1	3.82 M	0.0061	0.9897	0.9932	0.9617	0.0146	0.9830	0.9840	0.8947

The results in Table 4 show that the configuration of two independent layers and three shared layers (IL 2, SL 3) achieves the best performance for both visible and infrared datasets. Specifically, for the RDD5000 visible light test dataset, this configuration yields an MAE of 0.0053,  $F_\beta$  of 0.9907,  $E_\zeta$  of 0.9939, and an  $S_m$  of 0.9780. Similarly, for the infrared test dataset, it achieves an MAE of 0.0108,  $F_\beta$  of 0.9881,  $E_\zeta$  of 0.9885, and an  $S_m$  of 0.9001, outperforming other layer configurations.

Our analysis reveals that the interplay between independent and shared layers is critical to model performance. With their sum fixed at five, it is essential to find a proper balance. A higher number of independent layers enhances feature extraction but reduces the number of shared layers, weakening cross-modal fusion. Conversely, increasing shared layers improves fusion at the expense of reducing the capacity for feature extraction in each modality. Ultimately, the IL 2, SL 3 configuration strikes the best balance between performance and computational efficiency, ensuring optimal model output.

#### 4.3.3. Effectiveness of the MFEM

The MFEM is primarily designed to enhance the semantic features output by the lightweight encoder. We conducted several ablation experiments, including removing the MFEM (denoted as “w/o MFEM”), where the decoder receives the encoder output directly via the MAFM; and comparing the effects of different dilation strategies in the MFEM, where “MFEM<sub>T=123</sub>” and “MFEM<sub>T=135</sub>” represent parallel three  $3 \times 3$  depthwise separable convolution units with dilation rates T set to 1, 2, 3 and 1, 3, 5, respectively. Additionally, another design of the MFEM module was tested, where parallel  $3 \times 3$ ,  $5 \times 5$ , and  $7 \times 7$  depthwise separable convolution kernels are used to process the input feature map, labeled as “MFEM<sub>K=357</sub>”.

The experimental results, presented in Table 5, indicate a notable performance degradation when the MFEM is removed, highlighting the importance of this module for the overall model performance. The removal of MFEM (w/o MFEM) led to higher MAE values and lower performance metrics across both visible and infrared test datasets. In contrast, the configurations incorporating the MFEM consistently outperformed the baseline without the MFEM, particularly the “MFEM<sub>T=123</sub>” configuration, which achieved the best results for both visible and infrared datasets.

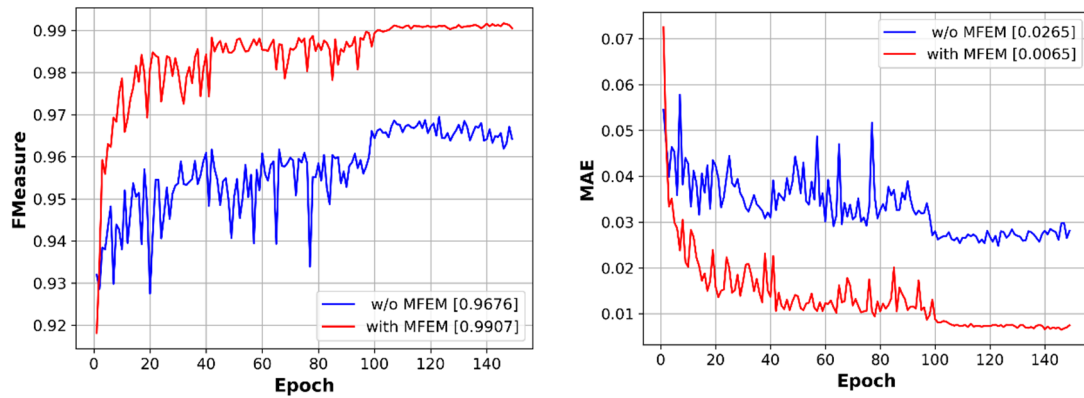
**Table 5.** Evaluation of the MFEM design.  $\uparrow$  and  $\downarrow$  indicate ‘larger is better’ and ‘smaller is better’, respectively. #Params. (M) denotes the number of parameters in millions. The best results are highlighted in bold.

Method	#Params. (M) $\downarrow$	RDD5000 Visible Light Test Dataset				RDD5000 Infrared Test Dataset			
		MAE $\downarrow$	$F_\beta$ $\uparrow$	$E_\xi$ $\uparrow$	$S_m$ $\uparrow$	MAE $\downarrow$	$F_\beta$ $\uparrow$	$E_\xi$ $\uparrow$	$S_m$ $\uparrow$
w/o MFEM	3.07 M	0.0103	0.9789	0.9807	0.9641	0.0455	0.9485	0.9528	<b>0.9357</b>
MFEM <sub>T=123</sub>	3.30 M	0.0053	0.9907	0.9939	<b>0.9780</b>	<b>0.0108</b>	<b>0.9881</b>	<b>0.9885</b>	0.9001
MFEM <sub>T=135</sub>	3.30 M	<b>0.0049</b>	<b>0.9909</b>	<b>0.9946</b>	0.9641	0.0144	0.9832	0.9836	0.8897
MFEM <sub>K=357</sub>	3.41 M	0.0054	0.9900	0.9941	0.9625	0.0167	0.9814	0.9792	0.8850

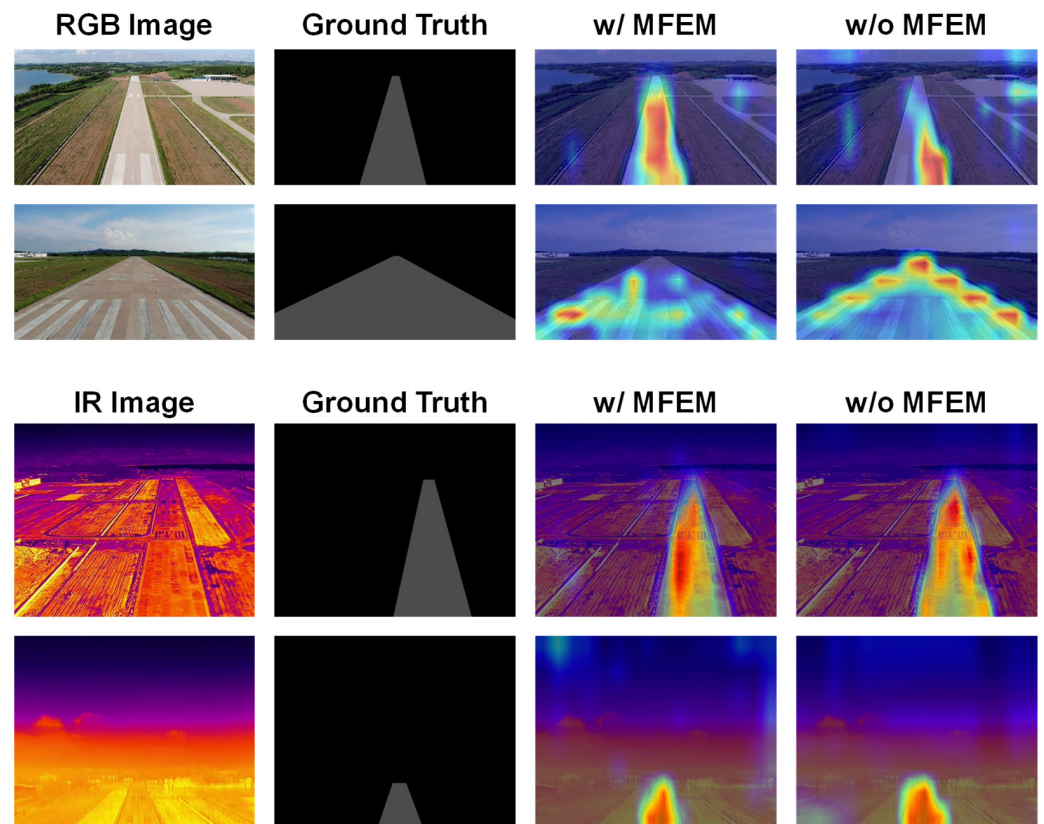
Regarding the MFEM design choices, different dilation strategies (MFEM<sub>T=123</sub> and MFEM<sub>T=135</sub>) and convolution kernel size combinations (MFEM<sub>K=357</sub>) showed only marginal differences in model performance. For both visible and infrared datasets, the changes in dilation rates and kernel sizes had minimal impact on key metrics such as MAE,  $F_\beta$ ,  $E_\xi$ , and  $S_m$ . This suggests that the MFEM, in its current form, is already highly effective at improving the model’s performance, with the dilation rates and kernel sizes playing a secondary role.

Figure 9 depicts the training dynamics of the DCFNet model with and without the MFEM on the RDD5000 dataset. The red curves represent the model with the MFEM, while the blue curves correspond to the model without the MFEM. The figure consists of two subplots: one illustrating the progression of the maximum F-measure score, and the other showing the MAE throughout the training process. The results clearly demonstrate that the model with the MFEM not only achieves higher F-measure scores at an accelerated pace but also shows a substantial reduction in MAE during training. These findings underscore the crucial role of the MFEM in enhancing the model’s performance, providing faster convergence and more accurate predictions.

Figure 10 provides a visual comparison of the feature heatmaps generated by the DCFNet model with and without the MFEM. The first column displays the original visible light or infrared images of airport runways, the second column shows the corresponding ground truth (GT), the third column presents the feature heatmaps when the MFEM is used, and the fourth column shows the feature heatmaps without the MFEM.



**Figure 9.** Max F-measure scores and MAE on RDD5000 in the whole training procedure to verify the effectiveness of the MFEM.



**Figure 10.** Visual comparison of the feature heatmaps with/without the MFEM.

As shown in Figure 10, the model with the MFEM (third column) captures the airport runway regions more accurately and confidently than the model without the MFEM (fourth column). The feature heatmaps generated with the MFEM exhibit stronger activation in the runway areas, indicating that the MFEM significantly enhances the model's ability to extract and focus on salient features. In contrast, the heatmaps without the MFEM show weaker and less precise activations, particularly in challenging regions such as low-visibility or complex backgrounds.

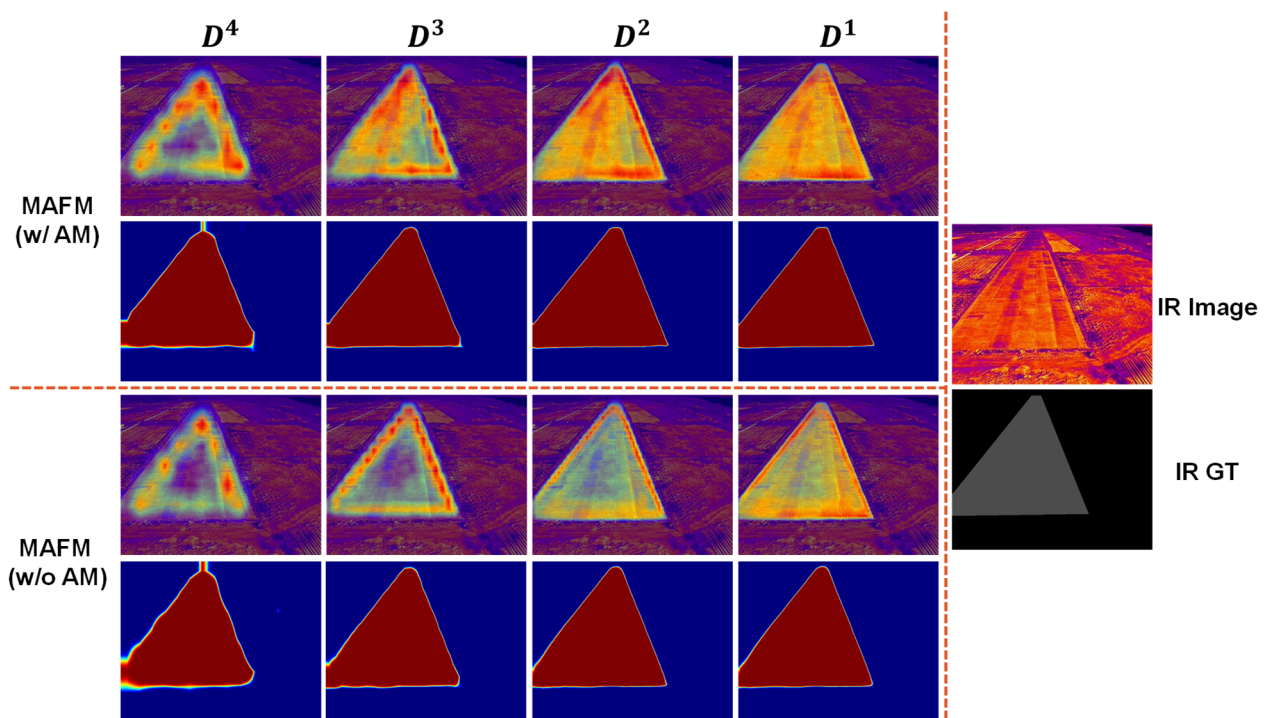
#### 4.3.4. Attention Mechanisms in the MAFM

The MAFM optimizes saliency object detection performance by introducing attention modules (AM) at multiple feature levels. To evaluate the contribution of these attention mechanisms, we conducted ablation experiments to assess the performance of the model with and without attention mechanisms, as well as with different combinations of attention

mechanisms (e.g., channel attention (CA) and spatial attention (SA)). The experimental results are presented in Table 6, and a visual comparison of the feature heatmaps with and without AM is shown in Figure 11.

**Table 6.** Evaluation of AM design in the MAFM.  $\uparrow$  and  $\downarrow$  indicate ‘larger is better’ and ‘smaller is better’, respectively. The best results are highlighted in **bold**.

Method	RDD5000 Visible Light Test Dataset				RDD5000 Infrared Test Dataset			
	MAE $\downarrow$	$F_\beta$ $\uparrow$	$E_\xi$ $\uparrow$	$S_m$ $\uparrow$	MAE $\downarrow$	$F_\beta$ $\uparrow$	$E_\xi$ $\uparrow$	$S_m$ $\uparrow$
MAFM (CA <sub>1-4</sub> )	0.0054	0.9905	0.9940	0.9627	0.1140	0.9857	0.9840	0.8904
MAFM (SA <sub>1-4</sub> )	<b>0.0051</b>	0.9906	<b>0.9944</b>	0.9636	0.0242	0.9742	0.9740	0.8639
MAFM (SA <sub>1</sub> CA <sub>2-4</sub> )	0.0053	0.9903	0.9942	0.9632	0.0157	0.9836	0.9818	0.8877
MAFM (SA <sub>1-2</sub> CA <sub>3-4</sub> )	0.0053	<b>0.9907</b>	0.9939	<b>0.9780</b>	<b>0.0108</b>	<b>0.9881</b>	<b>0.9885</b>	<b>0.9001</b>
MAFM (SA <sub>1-3</sub> CA <sub>4</sub> )	0.0052	0.9905	0.9942	0.9638	0.0119	0.9857	0.9867	0.8961
MAFM (w/o AM)	0.0053	0.9901	0.9942	0.9625	0.0287	0.9644	0.9687	0.8525



**Figure 11.** Visual comparison of the features with/without AM in the MAFM.

Table 6 evaluates the performance of different attention mechanism configurations in the MAFM. Several configurations are compared, including models with only channel attention (CA), only spatial attention (SA), and combinations of both. For example, the configuration “MAFM (CA<sub>1-4</sub>)” refers to applying CA across all blocks (Block 1 to Block 4) in the MAFM module, while “MAFM (SA<sub>1</sub>CA<sub>2-4</sub>)” uses SA in Block 1 and CA in Blocks 2 to 4.

The results in Table 6 demonstrate that removing the AM (“MAFM (w/o AM)”) leads to a significant performance drop, demonstrating the critical role of attention mechanisms in the MAFM. The “MAFM (SA<sub>1-2</sub>CA<sub>3-4</sub>)” configuration performed well on both the visible and infrared test sets, particularly on the infrared dataset, where the  $F_\beta$  and MAE scores were 0.9881 and 0.0108, respectively. These results validate the effectiveness of the proposed attention mechanism design in improving saliency detection accuracy.

Figure 11 provides a visual comparison of the output feature maps generated by the MAFM with and without AM at different scales. The left side of the figure shows the output feature maps at multiple scales, from coarse to fine, corresponding to Block 4, Block 3,

Block 2, and Block 1, denoted as  $D^4$ ,  $D^3$ ,  $D^2$  and  $D^1$ , respectively. The first and second rows show the output feature heatmaps and saliency maps when AM is used, while the third and fourth rows show the output feature heatmaps and saliency maps without AM. The right side of the figure displays the original images and their corresponding ground truth (GT) for intuitive comparison between the model outputs and the true salient regions.

The comparative results in Figure 11 clearly illustrate the positive impact of integrating attention mechanisms into the MAFM module, improving both feature extraction quality and saliency detection precision. By applying AM at different scales, the model is better able to capture and emphasize key features within the image, thus improving the overall performance of saliency object detection.

#### 4.3.5. Loss Function Design

The proposed model uses a hybrid loss function that combines binary cross-entropy (BCE) loss and Dice loss, along with a visible light weighting parameter  $\alpha$  to balance the contributions of visible and infrared modalities, as shown in Equation (16). We conducted ablation experiments with different values of  $\alpha$ , and the results are presented in Table 7. In addition, we also tested the performance of the model using only BCE loss and only Dice loss.

**Table 7.** Comparison of the performance of different loss functions.  $\uparrow$  and  $\downarrow$  indicate ‘larger is better’ and ‘smaller is better’, respectively. The best results are highlighted in **bold**.

Loss Setting	RDD5000 Visible Light Test Dataset				RDD5000 Infrared Test Dataset			
	MAE $\downarrow$	$F_\beta$ $\uparrow$	$E_\xi$ $\uparrow$	$S_m$ $\uparrow$	MAE $\downarrow$	$F_\beta$ $\uparrow$	$E_\xi$ $\uparrow$	$S_m$ $\uparrow$
$\alpha = 0.5$	0.0053	<b>0.9907</b>	0.9939	<b>0.9780</b>	<b>0.0108</b>	<b>0.9881</b>	<b>0.9885</b>	0.9001
$\alpha = 0.6$	<b>0.0052</b>	<b>0.9907</b>	<b>0.9940</b>	0.9634	0.0155	0.9822	0.9835	0.8889
$\alpha = 0.4$	0.0054	0.9904	0.9939	0.9632	0.0130	0.9859	0.9851	0.8947
BCE only	0.0059	0.9905	0.9934	0.9642	0.0112	0.9876	0.9873	<b>0.9020</b>
Dice only	0.0056	0.9893	0.9939	0.9608	0.0169	0.9736	0.9808	0.8929

The experimental results indicate that the model performs optimally when  $\alpha = 0.5$ , i.e., when the contributions of the visible and infrared data are balanced. Therefore, we chose  $\alpha = 0.5$  as the default setting for network training. Further analysis revealed that introducing Dice loss supervision improved the MAE by approximately 0.1% to 0.2%, but had minimal impact on the  $F_\beta$  score. Given that  $F_\beta$  is the primary evaluation metric for SOD and is mainly enhanced by BCE loss, we conclude that the hybrid BCE and Dice loss function is a reasonable choice for the default setting.

#### 4.4. Experimental Example of Our Method

In addition to runway detection, we have extended the application of the proposed method to other aviation tasks, such as airport runway line instance segmentation. As shown in Table 8, our method achieves competitive performance on both visible and infrared datasets for runway area segmentation and runway line segmentation tasks. For instance, on the RDD5000 visible light dataset, DCFNet attains an F1 score of 0.9940 for runway area segmentation and an IoU of 0.6301 for runway line segmentation. These results demonstrate the potential of our method for broader applications in aviation and remote sensing. In future work, we plan to further explore the generalization of our approach to other object detection tasks, such as aircraft detection and obstacle detection.

**Table 8.** Performance analysis of DCFNet applied to other aviation tasks. ↑ and ↓ indicate ‘larger is better’ and ‘smaller is better’, respectively.

Method	RDD5000 Visible Light Test Dataset				RDD5000 Infrared Test Dataset			
	Runway Area Segmentation		Runway Line Segmentation		Runway Area Segmentation		Runway Line Segmentation	
	F1 ↑	IoU ↑	Accuracy ↑	IoU ↑	F1 ↑	IoU ↑	Accuracy ↑	IoU ↑
Ours (DCFNet)	0.9940	0.9882	0.9818	0.6301	0.9676	0.9415	0.8855	0.3353

## 5. Conclusions

This study presents a novel approach for real-time runway detection by integrating visible and infrared data through dual-modal saliency object detection (SOD) techniques, addressing the critical need for improved detection accuracy in complex environments. We propose an innovative dual-modal SOD algorithm that effectively extracts salient objects from misaligned visible and infrared images, successfully overcoming the alignment challenges inherent in traditional methods.

A key contribution of this work is the creation of the RDD5000 dataset, a large-scale visible–infrared runway dataset captured from real-world scenarios. This dataset fills a significant gap in publicly available datasets for airport runway detection from an aircraft landing perspective and is expected to serve as a valuable resource for future research in aviation safety.

Experimental results demonstrate the superior performance of our proposed dual-modal cross-layer lightweight fusion network (DCFNet), achieving a maximum F-measure of 99.07% and a mean absolute error (MAE) of 0.0053 on the RDD5000 dataset. The lightweight version of DCFNet, based on MobileNetV2, operates at an impressive speed of 155 FPS on a single GPU, making it highly suitable for real-time use in airborne systems. These results highlight the effectiveness of our method in improving both detection accuracy and computational efficiency, which are critical for intelligent navigation systems.

Despite the promising results, the proposed method has certain limitations, such as performance degradation under extreme weather conditions, reliance on GPU deployment, and the need for further dataset scalability. Future work will focus on enhancing robustness through multi-modal fusion, exploring deployment on low-power embedded systems, expanding the dataset to include more data from real navigation systems, and extending the method to other target detection tasks in aviation and remote sensing, enabling real-time applications in more complex environments.

In conclusion, this research provides an effective solution for airport runway detection, and we look forward to its potential to make a greater impact within intelligent transportation systems.

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

**Funding:** This research was funded by the National Key Research and Development Program of China (No. 2022YFB3904303) and the National Natural Science Foundation of China (No. 62076019).

**Data Availability Statement:** The datasets presented in this article are not readily available because the data are part of an ongoing study. Requests to access the datasets should be directed to yanglc2003@buaa.edu.cn.



**Acknowledgments:** We express our sincere gratitude to the researchers behind DCNet and SAC-Net for generously sharing their algorithm codes, which greatly facilitated the execution of our comparative experiments.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

1. Zuo, Z.; Yang, B.; Li, Z.; Zhang, T. A GNSS/IMU/vision ultra-tightly integrated navigation system for low altitude aircraft. *IEEE Sens. J.* **2022**, *22*, 11857–11864.
2. Watanabe, Y.; Manecy, A.; Hiba, A.; Nagai, S.; Aoki, S. Vision-integrated navigation system for aircraft final approach in case of GNSS/SBAS or ILS failures. In Proceedings of the AIAA Scitech 2019 Forum, San Diego, CA, USA, 7–11 January 2019; p. 0113.
3. Krammer, C.; Mishra, C.; Holzapfel, F. Testing and evaluation of a vision-augmented navigation system for automatic landings of general aviation aircraft. In Proceedings of the AIAA Scitech 2020 Forum, Orlando, FL, USA, 6–10 January 2020; p. 1083.
4. Gao, Y.; Wang, Y.; Tian, L.; Li, D.; Wang, F. Visual Navigation Algorithms for Aircraft Fusing Neural Networks in Denial Environments. *Sensors* **2024**, *24*, 4797. [[CrossRef](#)] [[PubMed](#)]
5. Liu, T.; Yuan, Z.; Sun, J.; Wang, J.; Zheng, N.; Tang, X.; Shum, H.Y. Learning to detect a salient object. *IEEE Trans. Pattern Anal. Mach. Intell.* **2010**, *33*, 353–367.
6. Borji, A.; Cheng, M.M.; Jiang, H.; Li, J. Salient object detection: A benchmark. *IEEE Trans. Image Process.* **2015**, *24*, 5706–5722. [[CrossRef](#)] [[PubMed](#)]
7. Borji, A.; Cheng, M.M.; Hou, Q.; Jiang, H.; Li, J. Salient object detection: A survey. *Comput. Vis. Media* **2019**, *5*, 117–150. [[CrossRef](#)]
8. Tu, Z.; Li, Z.; Li, C.; Lang, Y.; Tang, J. Multi-interactive dual-decoder for RGB-thermal salient object detection. *IEEE Trans. Image Process.* **2021**, *30*, 5678–5691. [[CrossRef](#)]
9. Huo, F.; Zhu, X.; Zhang, L.; Liu, Q.; Shu, Y. Efficient context-guided stacked refinement network for RGB-T salient object detection. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *32*, 3111–3124. [[CrossRef](#)]
10. Wang, J.; Song, K.; Bao, Y.; Huang, L.; Yan, Y. CGFNet: Cross-guided fusion network for RGB-T salient object detection. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *32*, 2949–2961. [[CrossRef](#)]
11. Liu, Z.; Tan, Y.; He, Q.; Xiao, Y. SwinNet: Swin transformer drives edge-aware RGB-D and RGB-T salient object detection. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *32*, 4486–4497. [[CrossRef](#)]
12. Huo, F.; Zhu, X.; Zhang, Q.; Liu, Z.; Yu, W. Real-time one-stream semantic-guided refinement network for RGB-thermal salient object detection. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 1–12. [[CrossRef](#)]
13. Cong, R.; Zhang, K.; Zhang, C.; Zheng, F.; Zhao, Y.; Huang, Q.; Kwong, S. Does thermal really always matter for RGB-T salient object detection? *IEEE Trans. Multimed.* **2022**, *25*, 6971–6982. [[CrossRef](#)]
14. Tu, Z.; Li, Z.; Li, C.; Tang, J. Weakly alignment-free RGBT salient object detection with deep correlation network. *IEEE Trans. Image Process.* **2022**, *31*, 3752–3764. [[CrossRef](#)]
15. Tang, B.; Liu, Z.; Tan, Y.; He, Q. HRTransNet: HRFormer-driven two-modality salient object detection. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *33*, 728–742. [[CrossRef](#)]
16. Ma, S.; Song, K.; Dong, H.; Tian, H.; Yan, Y. Modal complementary fusion network for RGB-T salient object detection. *Appl. Intell.* **2023**, *53*, 9038–9055. [[CrossRef](#)]
17. Zhou, W.; Zhu, Y.; Lei, J.; Yang, R.; Yu, L. LSNet: Lightweight spatial boosting network for detecting salient objects in RGB-thermal images. *IEEE Trans. Image Process.* **2023**, *32*, 1329–1340. [[CrossRef](#)] [[PubMed](#)]
18. Pang, Y.; Zhao, X.; Zhang, L.; Lu, H. CAVER: Cross-modal view-mixed transformer for bi-modal salient object detection. *IEEE Trans. Image Process.* **2023**, *32*, 892–904. [[CrossRef](#)] [[PubMed](#)]
19. Zhou, W.; Sun, F.; Jiang, Q.; Cong, R.; Hwang, J.N. WaveNet: Wavelet network with knowledge distillation for RGB-T salient object detection. *IEEE Trans. Image Process.* **2023**, *32*, 3027–3039. [[CrossRef](#)] [[PubMed](#)]
20. Zhang, Z.; Wang, J.; Han, Y. Saliency prototype for RGB-D and RGB-T salient object detection. In Proceedings of the 31st ACM International Conference on Multimedia, Ottawa, ON, Canada, 29 October–3 November 2023; pp. 3696–3705.
21. Wang, K.; Tu, Z.; Li, C.; Zhang, C.; Luo, B. Learning Adaptive Fusion Bank for Multi-modal Salient Object Detection. *IEEE Trans. Circuits Syst. Video Technol.* **2024**, *34*, 7344–7358. [[CrossRef](#)]
22. Guo, R.; Ying, X.; Qi, Y.; Qu, L. UniTR: A Unified TRansformer-based Framework for Co-object and Multi-modal Saliency Detection. *IEEE Trans. Multimed.* **2024**, *26*, 7622–7635. [[CrossRef](#)]
23. Luo, Z.; Liu, N.; Zhao, W.; Yang, X.; Zhang, D.; Fan, D.P.; Han, J. VSCoDe: General Visual Salient and Camouflaged Object Detection with 2D Prompt Learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 17169–17180.

24. Zhang, W.; Ji, G.P.; Wang, Z.; Fu, K.; Zhao, Q. Depth quality-inspired feature manipulation for efficient RGB-D salient object detection. In Proceedings of the 29th ACM International Conference on Multimedia, Chengdu, China, 20–24 October 2021; pp. 731–740.
25. Wu, Y.H.; Liu, Y.; Xu, J.; Bian, J.W.; Gu, Y.C.; Cheng, M.M. MobileSal: Extremely efficient RGB-D salient object detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 10261–10269. [[CrossRef](#)] [[PubMed](#)]
26. Jin, X.; Yi, K.; Xu, J. MoADNet: Mobile asymmetric dual-stream networks for real-time and lightweight RGB-D salient object detection. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 7632–7645. [[CrossRef](#)]
27. Ling, L.; Wang, Y.; Wang, C.; Xu, S.; Huang, Y. Depth-aware lightweight network for RGB-D salient object detection. *IET Image Process.* **2023**, *17*, 2350–2361. [[CrossRef](#)]
28. Dong, S.; Feng, Y.; Yang, Q.; Huang, Y.; Liu, D.; Fan, H. Efficient multimodal semantic segmentation via dual-prompt learning. *arXiv* **2023**, arXiv:2312.00360.
29. Wang, G.; Li, C.; Ma, Y.; Zheng, A.; Tang, J.; Luo, B. RGB-T saliency detection benchmark: Dataset, baselines, analysis and a novel approach. In *Image and Graphics Technologies and Applications, Proceedings of the 13th Conference on Image and Graphics Technologies and Applications, IGTA 2018, Beijing, China, 8–10 April 2018; Revised Selected Papers 13*; Springer Nature: Cham, Switzerland, 2018; pp. 359–369.
30. Tu, Z.; Xia, T.; Li, C.; Wang, X.; Ma, Y.; Tang, J. RGB-T image saliency detection via collaborative graph learning. *IEEE Trans. Multimed.* **2019**, *22*, 160–173. [[CrossRef](#)]
31. Tu, Z.; Ma, Y.; Li, Z.; Li, C.; Xu, J.; Liu, Y. RGBT salient object detection: A large-scale dataset and benchmark. *IEEE Trans. Multimed.* **2022**, *25*, 4163–4176. [[CrossRef](#)]
32. Ju, R.; Ge, L.; Geng, W.; Ren, T.; Wu, G. Depth saliency based on anisotropic center-surround difference. In Proceedings of the 2014 IEEE International Conference on Image Processing (ICIP), Paris, France, 27–30 October 2014; pp. 1115–1119.
33. Peng, H.; Li, B.; Xiong, W.; Hu, W.; Ji, R. RGBD salient object detection: A benchmark and algorithms. In *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, 6–12 September 2014, Proceedings, Part III 13*; Springer International Publishing: Berlin/Heidelberg, Germany, 2014; pp. 92–109.
34. Niu, Y.; Geng, Y.; Li, X.; Liu, F. Leveraging Stereopsis for Saliency Analysis. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 454–461.
35. Fan, D.P.; Lin, Z.; Zhang, Z.; Zhu, M.; Cheng, M.M. Rethinking RGB-D salient object detection: Models, data sets, and large-scale benchmarks. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *32*, 2075–2089. [[CrossRef](#)]
36. Yin, B.; Zhang, X.; Li, Z.; Liu, L.; Cheng, M.M.; Hou, Q. Dformer: Rethinking RGBD representation learning for semantic segmentation. *arXiv* **2023**, arXiv:2309.09668.
37. Hao, S.; Zhong, C.; Tang, H. Cola: Conditional dropout and language-driven robust dual-modal salient object detection. In *European Conference on Computer Vision*; Springer Nature: Cham, Switzerland, 2025; pp. 354–371.
38. Zhang, W.; Jiang, Y.; Fu, K.; Zhao, Q. BTS-Net: Bi-directional transfer-and-selection network for RGB-D salient object detection. In Proceedings of the 2021 IEEE International Conference on Multimedia and Expo (ICME), Shenzhen, China, 5–9 July 2021; pp. 1–6.
39. Lee, M.; Park, C.; Cho, S.; Lee, S. SPSN: Superpixel prototype sampling network for RGB-D salient object detection. In Proceedings of the European Conference on Computer Vision 2022, Tel Aviv, Israel, 23–27 October 2022; pp. 630–647.
40. Yan, C.; Gong, B.; Wei, Y.; Gao, Y. Deep Multi-View Enhancement Hashing for Image Retrieval. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *43*, 1445–1451. [[CrossRef](#)] [[PubMed](#)]
41. Li, G.; Qian, M.; Xia, G.S. Unleashing Unlabeled Data: A Paradigm for Cross-View Geo-Localization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2024; pp. 16719–16729.
42. Guan, F.; Zhao, N.; Fang, Z.; Jiang, L.; Zhang, J.; Yu, Y.; Huang, H. Multi-level Representation Learning via ConvNeXt-Based Network for Unaligned Cross-View Matching. *Geo-Spat. Inf. Sci.* **2025**, 1–14. [[CrossRef](#)]
43. Liu, H.; Tan, X.; Zhou, X. Parameter sharing exploration and hetero-center triplet loss for visible-thermal person re-identification. *IEEE Trans. Multimed.* **2020**, *23*, 4414–4425.
44. Akbar, J.; Shahzad, M.; Malik, M.I.; Ul-Hasan, A.; Shafait, F. Runway detection and localization in aerial images using deep learning. In Proceedings of the 2019 Digital Image Computing: Techniques and Applications (DICTA), Perth, Australia, 2–4 December 2019; pp. 1–8.
45. Men, Z.C.; Jiang, J.; Guo, X.; Chen, L.J.; Liu, D.S. Airport runway semantic segmentation based on DCNN in high spatial resolution remote sensing images. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2020**, *42*, 361–366. [[CrossRef](#)]
46. Wang, Y.; Jiang, H.; Liu, C.; Pei, X.; Qiu, H. An airport runway detection algorithm based on semantic segmentation. *Navig. Position. Timing CSTPCD* **2021**, *8*, 97–106.
47. Chen, W.; Zhang, Z.; Yu, L.; Tai, Y. BARS: A benchmark for airport runway segmentation. *Appl. Intell.* **2023**, *53*, 20485–20498. [[CrossRef](#)]
48. Wang, Q.; Feng, W.; Zhao, H.; Liu, B.; Lyu, S. VALNet: Vision-Based Autonomous Landing with Airport Runway Instance Segmentation. *Remote Sens.* **2024**, *16*, 2161. [[CrossRef](#)]

49. Weng, X.; Guo, B. A method of airport runway dataset construction for the visual detection algorithm. *J. Phys. Conf. Ser.* **2023**, *2435*, 012016. [[CrossRef](#)]
50. Song, K.; Wen, H.; Xue, X.; Huang, L.; Ji, Y.; Yan, Y. Modality Registration and Object Search Framework for UAV-Based Unregistered RGB-T Image Salient Object Detection. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–15.
51. Wang, K.; Lin, D.; Li, C.; Tu, Z.; Luo, B. Alignment-Free RGBT Salient Object Detection: Semantics-guided Asymmetric Correlation Network and A Unified Benchmark. *IEEE Trans. Multimed.* **2024**, *26*, 10692–10707. [[CrossRef](#)]
52. Lyu, P.; Yeung, P.H.; Cheng, X.; Yu, X.; Wu, C.; Rajapakse, J.C. Efficient Fourier Filtering Network with Contrastive Learning for UAV-based Unaligned Bi-modal Salient Object Detection. *arXiv* **2024**, arXiv:2411.03728.
53. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.C. Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2018, Salt Lake City, UT, USA, 18–22 June 2018; pp. 4510–4520.
54. Kou, Z.; Shi, Z.; Liu, L. Airport detection based on line segment detector. In Proceedings of the 2012 International Conference on Computer Vision in Remote Sensing, Xiamen, China, 16–18 December 2012; pp. 72–77.
55. Wu, W.; Xia, R.; Xiang, W.; Hui, B.; Chang, Z.; Liu, Y.; Zhang, Y. Recognition of airport runways in FLIR images based on knowledge. *IEEE Geosci. Remote Sens. Lett.* **2014**, *11*, 1534–1538. [[CrossRef](#)]
56. Budak, M.; Alcin, O.F.; Sengur, A. Automatic Airport Detection with Line Segment Detector and Histogram of Oriented Gradients from Satellite Images. In Proceedings of the 2018 International Conference on Artificial Intelligence and Data Processing (IDAP), Malatya, Turkey, 28–30 September 2018; pp. 1–5.
57. Wang, J.; Cheng, Y.; Xie, J.; Niu, W. A real-time sensor guided runway detection method for forward-looking aerial images. In Proceedings of the 2015 11th International Conference on Computational Intelligence and Security (CIS), Shenzhen, China, 19–20 December 2015; pp. 150–153.
58. Qu, Y.; Li, C.; Zheng, N. Airport detection base on support vector machine from a single image. In Proceedings of the 2005 5th International Conference on Information Communications & Signal Processing, Bangkok, Thailand, 6–9 December 2005; pp. 546–549.
59. Meng, D.; Yun-feng, C.; Lin, G. A method to recognize and track runway in the image sequences based on template matching. In Proceedings of the 2006 1st International Symposium on Systems and Control in Aerospace and Astronautics, Harbin, China, 19–21 January 2006; p. 4.
60. Jackson, P.T.; Nelson, C.J.; Schiefele, J.; Obara, B. Runway detection in High Resolution remote sensing data. In Proceedings of the 2015 9th International Symposium on Image and Signal Processing and Analysis (ISPA), Zagreb, Croatia, 7–9 September 2015; pp. 170–175.
61. Aytekin, Ö.; Zöngür, U.; Halici, U. Texture-based airport runway detection. *IEEE Geosci. Remote Sens. Lett.* **2012**, *10*, 471–475. [[CrossRef](#)]
62. Ajith, B.; Adlinge, S.D.; Dinesh, S.; Rajeev, U.P.; Padmakumar, E.S. Robust method to detect and track the runway during aircraft landing using colour segmentation and runway features. In Proceedings of the 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, 23–25 April 2019; pp. 751–757.
63. Chen, L.; Tan, S.; Pan, Z.; Xing, J.; Yuan, Z.; Xing, X.; Zhang, P. A new framework for automatic airports extraction from SAR images using multi-level dual attention mechanism. *Remote Sens.* **2020**, *12*, 560. [[CrossRef](#)]
64. Hua, Z.; Bian, Z.; Li, J. Airport Detection-Based Cosaliency on Remote Sensing Images. *Math. Probl. Eng.* **2021**, *2021*, 8956396. [[CrossRef](#)]
65. Amit, R.A.; Mohan, C.K. A robust airport runway detection network based on R-CNN using remote sensing images. *IEEE Aerosp. Electron. Syst. Mag.* **2021**, *36*, 4–20. [[CrossRef](#)]
66. Ji, C.; Cheng, L.; Li, N.; Zeng, F.; Li, M. Validation of global airport spatial locations from open databases using deep learning for runway detection. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *14*, 1120–1131. [[CrossRef](#)]
67. Luo, Y.; Shao, F.; Mu, B.; Chen, H.; Li, Z.; Jiang, Q. Dynamic Weighted Fusion and Progressive Refinement Network for Visible-Depth-Thermal Salient Object Detection. *IEEE Trans. Circuits Syst. Video Technol.* **2024**, *34*, 10662–10677. [[CrossRef](#)]
68. Wang, J.; Li, G.; Shi, J.; Xi, J. Weighted Guided Optional Fusion Network for RGB-T Salient Object Detection. *ACM Trans. Multimedia Comput. Commun. Appl.* **2024**, *20*, 1–20. [[CrossRef](#)]
69. Wang, J.; Zhang, Z.; Yu, N.; Han, Y. Progressive Expansion for Semi-Supervised Bi-Modal Salient Object Detection. *Pattern Recognit.* **2025**, *157*, 110868. [[CrossRef](#)]
70. Lin, J.; Zhu, L.; Shen, J.; Fu, H.; Zhang, Q.; Wang, L. ViDSOD-100: A New Dataset and a Baseline Model for RGB-D Video Salient Object Detection. *Int. J. Comput. Vis.* **2024**, *132*, 1–19. [[CrossRef](#)]
71. Wang, K.X.; Liu, C.H.; Zhang, R.F. CMA-SOD: Cross-modal Attention Fusion Network for RGB-D Salient Object Detection. *Visual Comput.* **2024**, 1–17. [[CrossRef](#)]
72. Vaswani, A. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**.

73. Liu, N.; Zhang, N.; Wan, K.; Shao, L.; Han, J. Visual saliency transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 4722–4732.
74. Wu, J.; Hao, F.; Liang, W.; Xu, J. Transformer fusion and pixel-level contrastive learning for RGB-D salient object detection. *IEEE Trans. Multimed.* **2023**, *26*, 1011–1026. [[CrossRef](#)]
75. Hu, X.; Sun, F.; Sun, J.; Wang, F.; Li, H. Cross-modal fusion and progressive decoding network for RGB-D salient object detection. *Int. J. Comput. Vis.* **2024**, *132*, 3067–3085. [[CrossRef](#)]
76. Gao, S.; Zhang, P.; Yan, T.; Lu, H. Multi-scale and detail-enhanced segment anything model for salient object detection. In Proceedings of the 32nd ACM International Conference on Multimedia, Melbourne, Australia, 28 October–1 November 2024; pp. 9894–9903.
77. Liu, Z.; Deng, S.; Wang, X.; Wang, L.; Fang, X.; Tang, B. SSFam: Scribble Supervised Salient Object Detection Family. *arXiv* **2024**, arXiv:2409.04817.
78. Howard, A.G. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv* **2017**, arXiv:1704.04861.
79. Zhang, X.; Zhou, X.; Lin, M.; Sun, J. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6848–6856.
80. Ma, N.; Zhang, X.; Zheng, H.T.; Sun, J. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 116–131.
81. Tan, M.; Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In Proceedings of the 36th International Conference on Machine Learning, Long Beach, CA, USA, 10–15 June 2019; pp. 6105–6114.
82. Cong, R.; Lin, Q.; Zhang, C.; Li, C.; Cao, X.; Huang, Q.; Zhao, Y. CIR-Net: Cross-modality interaction and refinement for RGB-D salient object detection. *IEEE Trans. Image Process.* **2022**, *31*, 6800–6815. [[CrossRef](#)]
83. Luo, Y.; Shao, F.; Xie, Z.; Wang, H.; Chen, H.; Mu, B.; Jiang, Q. HFMDNet: Hierarchical fusion and multi-level decoder network for RGB-D salient object detection. *IEEE Trans. Instrum. Meas.* **2024**, *73*, 1–15. [[CrossRef](#)]
84. Dong, P.; Wang, B.; Cong, R.; Sun, H.-H.; Li, C. Transformer with large convolution kernel decoder network for salient object detection in optical remote sensing images. *Comput. Vis. Image Underst.* **2024**, *240*, 103917. [[CrossRef](#)]
85. Yu, F. Multi-Scale Context Aggregation by Dilated Convolutions. *arxiv* **2015**, arXiv:1511.07122.
86. Ioffe, S. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv* **2015**, arXiv:1502.03167.
87. Nair, V.; Hinton, G.E. Rectified linear units improve restricted boltzmann machines. In Proceedings of the 27th International Conference on Machine Learning, Haifa, Israel, 21–24 June 2010; ICML-10. pp. 807–814.
88. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
89. Ranftl, R.; Lasinger, K.; Hafner, D.; Schindler, K.; Koltun, V. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *44*, 1623–1637. [[CrossRef](#)] [[PubMed](#)]
90. Achanta, R.; Hemami, S.; Estrada, F.; Susstrunk, S. Frequency-tuned salient region detection. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition 2009, Miami, FL, USA, 20–25 June 2009; pp. 1597–1604.
91. Fan, D.P.; Gong, C.; Cao, Y.; Ren, B.; Cheng, M.M.; Borji, A. Enhanced-alignment measure for binary foreground map evaluation. *arXiv* **2018**, arXiv:1805.10421.
92. Cheng, M.M.; Fan, D.P. Structure-measure: A new way to evaluate foreground maps. *Int. J. Comput. Vis.* **2021**, *129*, 2622–2638. [[CrossRef](#)]

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.