Signal Processing: Image Communication 145 (2026) 117551

Contents lists available at ScienceDirect

Signal Processing: Image Communication

journal homepage: www.elsevier.com/locate/image

Du-CIPT: Dual Cross-Modal Interactive Pyramid Transformer for RGB-Thermal salient object detection and segmentation

Jiesheng Wu [a,*], Ji Du [b], Fangwei Hao [b], Jiankang Hong [c]

[a] School of Computer and Information, Anhui Normal University, Wuhu, 241003, China
[b] College of Artificial Intelligence, Nankai University, Tianjin, 300350, China
[c] School of Computer Science and Technology, Tongji University, Shanghai, 201804, China

ARTICLE INFO

Keywords:
Cross-modal pyramid
Transformer
Cross-modal purification
Multi-scale long-range interaction

ABSTRACT

RGB-Thermal (RGB-T) salient object detection and semantic segmentation tasks play a vital role in multiple applications, aiming to enhance pixel-level predictions by fusing information from two modalities. Existing RGB-T cross-modal methods often suffer from unpurified cross-modal features and insufficient multi-scale long-range interactions. To address these challenges, we propose a method that first purifies cross-modal features and then performs multi-scale long-range interaction fusion. Specifically, we propose a Dual Cross-Modal Interactive Pyramid Transformer (Du-CIPT) model, which leverages the proposed Cross-modal Purification Module (CPM) to mutually calibrate and purify the features of the two modalities, enhancing the uni-modal features and improving the feature representative ability of each modality. Subsequently, the proposed Cross-modal Pyramid Pooling Transformer (CP^2T) is utilized to capture multi-scale long-range interactions and generate multi-scale fusion features effectively. Finally, an additional Cross-level Selective Fusion (CSF) module is introduced to filter the noise features of the decoding process. We apply Du-CIPT to two tasks: RGB-T salient object detection and RGB-T semantic segmentation. Across five major datasets for these tasks, compared against 49 state-of-the-art methods based on the same backbones, Du-CIPT achieves state-of-the-art performance.

1. Introduction

Salient Object Detection (SOD) and Semantic Segmentation (SS) are popular pixel-level dense prediction tasks in computer vision. While both tasks involve assigning labels to individual pixels, they serve different purposes. SOD primarily focuses on identifying the most salient objects in an image [1], whereas SS aims to classify each pixel into a predefined category [2]. SOD is a fundamental topic in computer vision and finds applications in various areas such as image understanding [3], video detection [4], defect detection [5,6], and other pre-processing tasks in computer vision [7]. On the other hand, SS plays a crucial role in diverse visual tasks including autonomous driving [8–10], medical image processing [11], and visual surveillance [12].

Existing methods typically rely on RGB images alone for salient object detection and semantic segmentation. However, in adverse conditions such as smoke, fog, or haze, visible light RGB images often struggle to capture effective features and perform poorly, particularly under low-light conditions. In contrast, thermal infrared maps can provide valuable information about the surface temperature of objects, highlighting those with temperatures higher than the surrounding environment. This capability compensates for the limitations of RGB images in adverse conditions [13]. As a result, in recent years, RGB-T SOD and RGB-T SS tasks have emerged. By leveraging the complementary strengths of both modalities, these tasks aim to accurately predict objects even in challenging low-light conditions.

Existing methods for RGB-T dense prediction tasks are commonly categorized into two groups: feature enhancement methods and cross-modal interaction fusion methods. Feature enhancement methods typically employ attention mechanisms to enhance uni-modal features. These enhanced features are then utilized either directly or indirectly as auxiliary cues to aid other modalities in improving their performance. Representative methods in this category include PSTNet [14], FEANet [15], and EGFNet [16]. Cross-modal interaction fusion methods, on the other hand, focus on fusing the two modalities by leveraging their complementary semantics to generate comprehensive fusion features for prediction. This category includes several notable methods such as MFNet [17], RTFNet [18], FuseSeg [19], AFNet [20], MMNet [21], GMNet [22], ABMDRNet [23], APNet [24], MIDD [25], MIA [26], ADF [27], ECFFNet [28], LSNet [29], and others [30–34].

While existing methods have demonstrated impressive performance, they often fall short of fully exploring and exploiting cross-modal

* Corresponding author.
E-mail address: jasonwu@mail.nankai.edu.cn (J. Wu).

https://doi.org/10.1016/j.image.2026.117551
Received 15 May 2025; Received in revised form 11 March 2026; Accepted 23 March 2026
Available online 4 April 2026

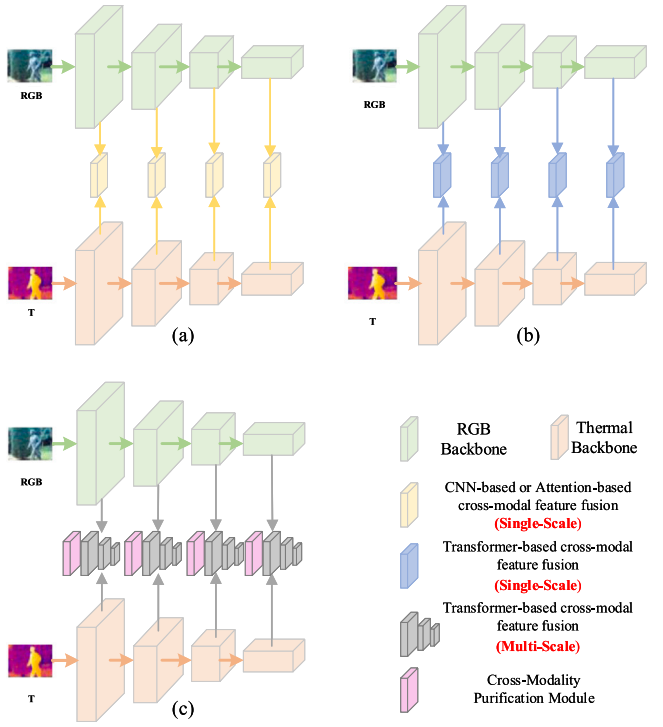

Fig. 1. Three different cross-modal feature fusion methods: (a) Cross-modal fusion based on CNN or attention mechanism (single-scale interactive fusion); (b) Cross-modal fusion based on Transformer (single-scale long-range interactive fusion); (c) Cross-modal fusion based on Transformer (multi-scale long-range interactive fusion).

semantics. Moreover, most fusion strategies rely on simplistic operations such as element-wise addition, multiplication, or channel-level concatenation, along with Convolutional Neural Network (CNN) blocks for fusion. These approaches may lead to insufficient cross-modal fusion and consequently result in poor performance.

Furthermore, although existing works leverage Transformers for long-range cross-modal fusion, interaction fusion is typically limited to single-scale interaction. This design may lack sensitivity to multi-scale objects and may fail to provide sufficient semantics in scenarios where RGB images are affected by low light or contain significant noise. Consequently, directly fusing features from the two modalities may not yield satisfactory results.

Specifically, as depicted in Fig. 1, (a) illustrates a commonly used cross-modal fusion strategy that achieves interaction fusion of cross-modal features. However, its limitation lies in its inability to achieve long-range dependent interaction and being restricted to single-scale interaction. On the other hand, (b) shows the current Transformer-based cross-modal fusion method, which achieves long-range cross-modal interaction fusion and generates efficient fusion features. However, it may fail to capture multi-scale fusion features, potentially hindering further performance improvement.

Moreover, both of these methods directly fuse features from the two modalities, which can lead to performance degradation when RGB images are in noisy conditions. To address these limitations, we propose a method that first performs cross-modal feature purification followed by multi-scale long-range interaction fusion. An overview of the entire method is illustrated in Fig. 1(c). Initially, features extracted by the backbones are fed into the cross-modal purification module to calibrate the uni-modal features. Subsequently, they are fed into the Transformer-based multi-scale feature interaction fusion module to achieve long-range cross-modal interaction fusion and generate multi-scale long-range fusion features for optimal performance. We argue that purifying features before fusion can mitigate the aforementioned limitations, and simultaneously incorporating multi-scale and long-range interaction semantics can enhance the performance of dense prediction tasks.

Based on the above analysis, we develop the Dual Cross-Modal Interactive Pyramid Transformer (Du-CIPT) model for RGB-T cross-modal fusion, achieving state-of-the-art performance on RGB-T SOD and RGB-T SS tasks. Specifically, three modules are proposed: Cross-modal Purification Module (CPM), Cross-modal Pyramid Pooling Transformer (CP^2T), and Cross-level Selective Fusion (CSF). The CPM module employs an attention mechanism to mutually calibrate and purify uni-modal features extracted by the backbones, enhancing the original features. CP^2T utilizes a pyramid pooling structure to extract multi-scale features and then employs the proposed cross-modal pyramid Multi-Dconv transposed self-attention mechanism to achieve multi-scale long-range cross-modal feature interaction. Finally, the CSF module filters out noise features from the decoding outputs during the decoding process, leading to the final prediction.

Our main contributions can be summarized as follows:

- A cross-modal purification followed by a multi-scale long-range interaction method is proposed for RGB-T fusion. Building upon this method, the Du-CIPT model is introduced, which conducts comprehensive multi-scale long-range multi-modal interaction on purified uni-modal features to boost detection performance.
- A CPM module is designed to purify the features of the two modalities, enhancing the uni-modal feature representative ability.
- The CP^2T module is proposed to achieve multi-scale long-range interactions and generate multi-scale fusion features. To the best of our knowledge, few existing RGB-T fusion models explicitly address multi-scale long-range interactions.
- An additional Cross-level Selective Fusion (CSF) module is incorporated into the decoding stage to suppress redundant information and enhance the fidelity of the final prediction.
- The proposed Du-CIPT model is evaluated on five datasets for both RGB-T SOD and RGB-T SS tasks. Extensive experimental results demonstrate that the proposed Du-CIPT achieves the best performance compared to a total of 49 state-of-the-art methods for both tasks.

2. Related work

Generally, RGB-Depth, RGB-Thermal, and RGB-Optical Flow fusion tasks each target different aspects of multimodal perception. RGB-Depth fusion primarily captures scene geometry through distance information, while RGB-Thermal fusion leverages surface temperature distributions to provide complementary semantic cues, especially under low-illumination conditions. In contrast, RGB-Optical Flow focuses on encoding temporal motion dynamics rather than intrinsic object properties. Distinct from the geometry or motion modeling pursued in these modalities, RGB-T fusion aims to enhance visual robustness and semantic consistency under adverse environmental conditions, making it particularly valuable for perception tasks in complex real-world scenarios. Our research focuses on the RGB-Thermal (RGB-T) task.

RGB-T tasks encompass two prominent tasks: RGB-T Salient Object Detection (RGB-T SOD) and RGB-T Semantic Segmentation (RGB-T SS). The former focuses on identifying the most salient objects or regions in a scene, serving various applications such as object segmentation and recognition. In contrast, the latter specializes in segmenting target objects under challenging lighting conditions, including nighttime or haze scenarios.

2.1. RGB-T salient object detection

In recent years, significant progress has been made in RGB-T SOD. Wang et al. introduced the first RGB-T SOD dataset named VT821 [35], alongside a multi-task learning algorithm tailored for this task. Tu et al. proposed a collaborative graph learning algorithm for RGB-T SOD [36].

With the advent of deep learning, CNN-based methods have been at the forefront of advancing RGB-T SOD [37,38]. These methods typically involve extracting uni-modal features followed by designing fusion components to integrate these features. For instance, Zhang et al. introduced a dual-stream interactive decoder to facilitate cross-modal feature fusion [39], while Wang et al. proposed a cross-guided fusion network for comprehensive cross-modal fusion and utilization of high-level semantic information [40]. Zhou et al. employed efficient bilateral fusion and multi-level coherent fusion modules for cross-modal fusion [28], whereas Liang et al. developed a multi-modal interactive attention unit along with decoding modules for multi-source and multi-level feature fusion [26]. Furthermore, Cong et al. recently explored the significance of thermal infrared information for salient object detection and introduced a global illumination estimation module to control the involvement of thermal infrared information [41], offering valuable insights for future developments in RGB-T SOD. More recently, methods such as UDNet [42] by Wang et al. CAVER [43] by Pang et al. and MCNet [44] by Jiang et al. have leveraged Transformer-based approaches to enhance the performance of RGB-T SOD.

2.2. RGB-T semantic segmentation

In the realm of RGB-T SS, Ha et al. made pioneering efforts by combining visible light and thermal infrared maps for scene segmentation and introducing the first RGB-T SS dataset named MFNet [17]. Building upon this, Shivalumar et al. proposed the PSTNet model, which leverages the confidence map of semantic segmentation as auxiliary clues, synthesizing it with RGB and thermal infrared maps to enhance cross-modal fusion representation. They also introduced the PST900 dataset for RGB-T SS [14]. In recent years, Deng et al. developed the FEANet, a two-stage feature enhancement attention model equipped with feature enhancement attention modules to extract and enhance multi-level RGB and thermal infrared features from both channel and spatial perspectives, thus preserving spatial information [15]. Guo et al. introduced MLFNet, a novel model that employs multi-level fusion of RGB and thermal images, which utilizes two identical ResNets as backbone encoders to extract RGB and thermal features separately, followed by hierarchical fusion of thermal features into RGB features during the encoding stage [45]. Zhou et al. proposed EGFNet, which adopts a unique strategy of utilizing edge information to enhance multi-modal features for facilitating cross-modal fusion. This method incorporates an MFM module to fuse multi-modal features, introduces a prior edge map to capture detail and texture information, and finally embeds edge semantics into the multi-modal fusion features by multiplying them with the prior edge, thereby enhancing the overall semantic segmentation performance [16].

Despite the impressive performance achieved by existing methods [17–29], they often fall short of fully leveraging semantic information from cross-modal fusion. Current approaches for cross-modal fusion typically rely on convolutional blocks or attention mechanisms, which may not sufficiently explore the information within each modality and adequately fuse cross-modal information, thus directly impacting final performance. Furthermore, while previous works have employed Transformer architectures for cross-modal fusion [46,47], such interaction fusion tends to be single-scale, potentially limiting sensitivity to multi-scale objects. Moreover, directly fusing features from both modalities may struggle in scenarios where RGB images are affected by low-light conditions or contain significant noise, thereby failing to provide adequate semantic features. To address these challenges, we propose a novel approach that first purifies cross-modal features and subsequently performs multi-scale long-range interaction fusion for RGB-T dense prediction.

3. Proposed method

3.1. Overall architecture

The overall architecture of Du-CIPT is depicted in Fig. 2. Du-CIPT comprises dual-stream encoders, a cross-modal fusion stage, and a decoder. The dual-stream encoder uses either CNN-based backbones to extract multi-stage and multi-scale features from two modalities. In our study, we leverage ResNet [48] as the encoder. The fusion stage facilitates multi-scale interactive fusion learning, consisting of the Cross-Modal Purification Module (CPM) and the Cross-Modal Pyramid Pooling Transformer (CP^2T). Subsequently, the decoder interprets the obtained fusion features for dense prediction tasks, employing the Cross-Level Selective Fusion (CSF) module.

Specifically, let $\boldsymbol{I} \in \mathbb{R}^{H\times W\times 3}$ and $\boldsymbol{T} \in \mathbb{R}^{H\times W\times 1}$ represent the RGB image and the thermal infrared map fed into the model, respectively. Initially, $\boldsymbol{I}$ and $\boldsymbol{T}$ undergo separate processing through ResNet-50, resulting in downsampled representations, denoted as $\boldsymbol{R}_i$ and $\boldsymbol{T}_i$ for different scales, where $i \in \{1,2,3,4\}$. These representations are downscaled by factors of 1/4, 1/8, 1/16, and 1/32, respectively.

Next, $\boldsymbol{R}_i$ and $\boldsymbol{T}_i$ are merged along the channel dimension using a 1×1 convolutional layer equipped with batch normalization and ReLU activation (Conv + Batch Normalization + ReLU, CBR). Subsequently, the fused features are fed into the fusion stage to generate multi-level fusion features $\boldsymbol{P}_i$. Finally, $\boldsymbol{P}_i$ is forwarded into the decoder to generate the final predictions denoted as $\boldsymbol{O}_i$.

3.2. Cross-modal purification module

RGB images offer rich color, texture, detail, and semantic information, but they often falter in adverse conditions such as low light, darkness, or heavy fog, leading to performance degradation. Conversely, thermal infrared maps capture temperature information of object surfaces, accentuating objects with temperatures higher than their surroundings, thereby compensating for the deficiencies of RGB images under adverse conditions [13]. Nonetheless, inherent differences in spatial and channel features between the two modalities pose challenges. Directly fusing features from both modalities may introduce noise and degrade performance. Therefore, an effective modality purification strategy becomes imperative for cross-modal feature fusion. To tackle this issue, we propose a CPM designed for cross-modal feature calibration and alignment. The specific details of this module are elaborated in Fig. 3. CA (channel attention) models global modality-level importance, while SA (spatial attention) captures localized modality cues. Their cross-modal coupling ensures complementary purification—CA refines global semantics, SA enhances local spatial consistency.

Specifically, let $\boldsymbol{R}_i \in \mathbb{R}^{h\times w\times c}$ and $\boldsymbol{T}_i \in \mathbb{R}^{h\times w\times c}$ represent the input feature maps into the CPM, with $\boldsymbol{R}_p^i \in \mathbb{R}^{h\times w\times c}$ and $\boldsymbol{T}_p^i \in \mathbb{R}^{h\times w\times c}$ denoting the output feature maps, where h, w, and c denote the height, width, and number of channels of the input feature maps, respectively, and $i \in \{1,2,3,4\}$. As depicted in Fig. 3, the input feature maps $\boldsymbol{R}_i$ and $\boldsymbol{T}_i$ are initially fed separately into their corresponding Channel Attention (CA) modules to emphasize channel importance. Subsequently, the obtained outputs undergo element-wise multiplication with the original inputs.

Following this, the resulting outputs are passed into the Spatial Attention (SA) modules of the opposite modality to enhance cross-modal spatial importance. Finally, element-wise multiplication and addition operations are performed with the original inputs on their respective feature streams, yielding the outputs $\boldsymbol{R}_p^i$ and $\boldsymbol{T}_p^i$. The entire process can be mathematically defined using the following equations:

$$\begin{aligned} \boldsymbol{T}_p^i &= \boldsymbol{T}_i + \boldsymbol{T}_i \otimes \mathrm{SA}\left(\boldsymbol{R}_i \otimes \mathrm{CA}\left(\boldsymbol{R}_i\right)\right), \\ \boldsymbol{R}_p^i &= \boldsymbol{R}_i + \boldsymbol{R}_i \otimes \mathrm{SA}\left(\boldsymbol{T}_i \otimes \mathrm{CA}\left(\boldsymbol{T}_i\right)\right), \end{aligned} \tag{1}$$

where $\mathrm{CA}(\cdot)$ and $\mathrm{SA}(\cdot)$ represent the CA mechanism and SA mechanism, respectively. $\otimes$ denotes element-wise multiplication.

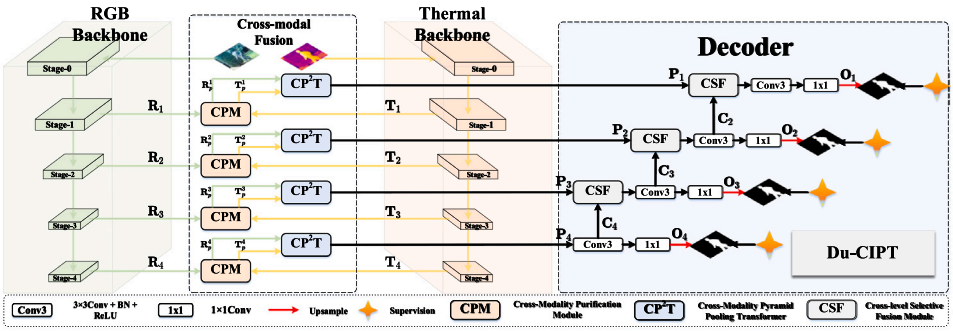

Fig. 2. Overall architecture of Du-CIPT.

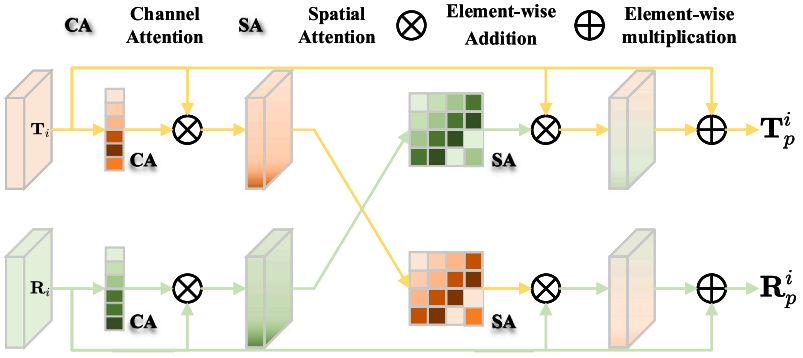

Fig. 3. The structure of the CPM.

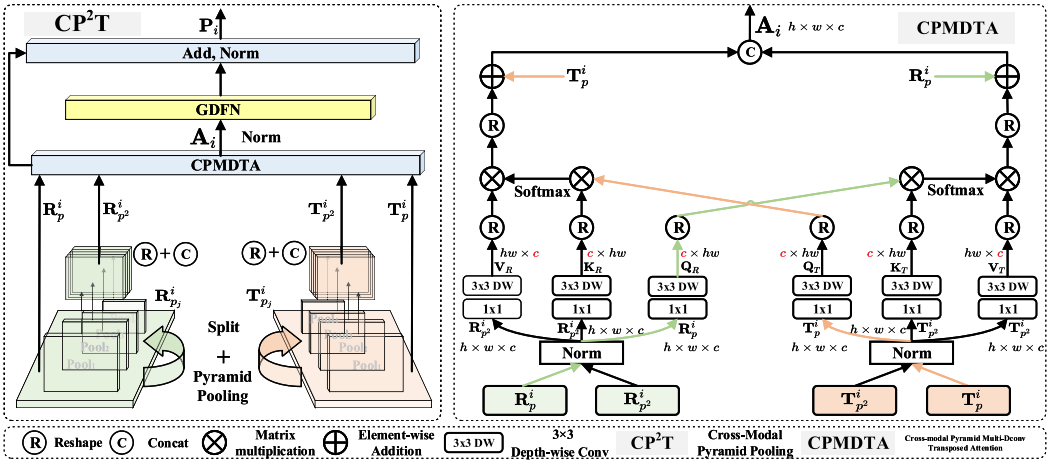

Fig. 4. The details of CP²T.

3.3. Cross-modal pyramid pooling module

Cross-modal interactive fusion learning plays a pivotal role in RGB-T dense prediction tasks. While previous studies have commonly employed convolutional blocks and attention mechanisms for interactive fusion learning [2,41,49–52], they often overlook the importance of incorporating multi-scale and long-range interactions, crucial for effective cross-modal dense prediction tasks. Inspired by the idea of P2T [53], we introduce the Cross-modal Pyramid Pooling Transformer (CP²T) to tackle this gap. Unlike existing approaches, CP²T is designed to concurrently facilitate multi-scale and long-range interactions, thereby enhancing the performance of cross-modal dense prediction tasks.

Fig. 4 illustrates the details of CP²T. Specifically, $\boldsymbol{R}_p^i$ and $\boldsymbol{T}_p^i$ are taken as the input feature maps to CP²T. These feature maps are subsequently divided into four equal parts along the channel dimension to enable pyramid pooling, resulting in multi-scale features $\boldsymbol{R}_{p_j}^i \in \mathbb{R}^{h_j \times w_j \times c/4}$ and $\boldsymbol{T}_{p_j}^i \in \mathbb{R}^{h_j \times w_j \times c/4}$, which can be expressed as follows:

$$\begin{aligned} \boldsymbol{R}_{p_j}^i &= \mathrm{AvgPool}\left(\mathrm{Split}\left(\boldsymbol{R}_p^i\right)\right), \quad j \in \{1,2,3,4\}, \\ \boldsymbol{T}_{p_j}^i &= \mathrm{AvgPool}\left(\mathrm{Split}\left(\boldsymbol{T}_p^i\right)\right), \quad j \in \{1,2,3,4\}, \end{aligned} \tag{2}$$

where $\mathrm{Split}(\cdot)$ and $\mathrm{AvgPool}(\cdot)$ represent the splitting operation and adaptive average pooling operation, respectively. For specific pooling ratios, please refer to the experimental details of Section 4. Then, the

obtained pyramid features are reshaped to the same size and concatenated along the channel dimension to obtain new multi-scale features $\boldsymbol{R}^i_{p^2} \in \mathbb{R}^{h\times w\times c}$ and $\boldsymbol{T}^i_{p^2} \in \mathbb{R}^{h\times w\times c}$, which can be expressed as follows:

$$\begin{aligned}\boldsymbol{R}^i_{p^2} &= \left[\mathrm{RP}\left(\boldsymbol{R}^i_{p_1}\right), \mathrm{RP}\left(\boldsymbol{R}^i_{p_2}\right), \mathrm{RP}\left(\boldsymbol{R}^i_{p_3}\right), \mathrm{RP}\left(\boldsymbol{R}^i_{p_4}\right)\right],\\ \boldsymbol{T}^i_{p^2} &= \left[\mathrm{RP}\left(\boldsymbol{T}^i_{p_1}\right), \mathrm{RP}\left(\boldsymbol{T}^i_{p_2}\right), \mathrm{RP}\left(\boldsymbol{T}^i_{p_3}\right), \mathrm{RP}\left(\boldsymbol{T}^i_{p_4}\right)\right],\end{aligned} \tag{3}$$

where $\mathrm{RP}(\cdot)$ and $[\cdots]$ represent the reshaping operation and concatenation operation, respectively. After these operations, multi-scale pyramid pooling uni-modal features are obtained. In particular, our design of the CP2T modules is partly inspired by P2T, in which progressive, hierarchical feature aggregation proves effective for multi-scale representation. The pyramid pooling ratios in CP2T are also inspired by P2T – $\{12, 16, 20, 24\}$, $\{6, 8, 10, 12\}$, $\{3, 4, 5, 6\}$, and $\{1, 2, 3, 4\}$ – were determined through empirical studies and domain knowledge. It is well known that lower backbone layers preserve fine structural details, while higher layers provide rich semantic cues; therefore, the pooling ratios gradually decrease toward deeper stages to balance local and global information.

3.3.1. Cross-modal Pyramid Multi-Dconv Transposed Attention

To facilitate cross-modal multi-scale long-range interaction learning, we introduce the Cross-modal Pyramid Multi-Dconv Transposed Attention (CPMDTA), depicted on the right side of Fig. 4. This module enables efficient computation while simultaneously learning multi-scale and long-range interactions.

Firstly, $\boldsymbol{R}^i_p$, $\boldsymbol{T}^i_p$, $\boldsymbol{R}^i_{p^2}$, and $\boldsymbol{T}^i_{p^2}$ after layer normalization (LN) are taken as input feature maps into the CPMDTA module. Then, CPMDTA generates two sets of queries ($\boldsymbol{Q}_R$, $\boldsymbol{Q}_T \in \mathbb{R}^{c\times hw}$), keys ($\boldsymbol{K}_R$, $\boldsymbol{K}_T \in \mathbb{R}^{c\times hw}$), and values ($\boldsymbol{V}_R$, $\boldsymbol{V}_T \in \mathbb{R}^{hw\times c}$). The generation process can be represented by the following equation:

$$\begin{aligned}&\boldsymbol{Q}_R, \boldsymbol{K}_R, \boldsymbol{V}_R = \boldsymbol{W}^d_q \boldsymbol{W}^p_q \mathrm{LN}\left(\boldsymbol{R}^i_p\right),\\ &\boldsymbol{W}^d_k \boldsymbol{W}^p_k \mathrm{LN}\left(\boldsymbol{R}^i_{p^2}\right), \boldsymbol{W}^d_v \boldsymbol{W}^p_v \mathrm{LN}\left(\boldsymbol{R}^i_{p^2}\right),\\ &\boldsymbol{Q}_T, \boldsymbol{K}_T, \boldsymbol{V}_T = \boldsymbol{W}'^d_q \boldsymbol{W}^p_q \mathrm{LN}\left(\boldsymbol{T}^i_p\right),\\ &\boldsymbol{W}'^d_k \boldsymbol{W}^p_k \mathrm{LN}\left(\boldsymbol{T}^i_{p^2}\right), \boldsymbol{W}'^d_v \boldsymbol{W}^p_v \mathrm{LN}\left(\boldsymbol{T}^i_{p^2}\right),\end{aligned} \tag{4}$$

where $\boldsymbol{W}^p$ and $\boldsymbol{W}'^p$ represent 1×1 pointwise convolutions. $\boldsymbol{W}^d$ and $\boldsymbol{W}'^d$ represent 1×1 depth-wise convolutions, and $\mathrm{LN}(\cdot)$ represents layer normalization.

Next, CPMDTA performs cross-modal self-attention for cross-modal interaction to generate fusion features $\boldsymbol{A}^i \in \mathbb{R}^{h\times w\times c}$. This process can be defined by the following equation:

$$\begin{aligned}\mathbf{A}^i = \mathrm{Conv}_3\Bigg(\mathrm{Concat}\Bigg(&\mathrm{Softmax}\left(\frac{\mathbf{Q}_R\mathbf{K}^{\intercal}_T}{\alpha}\right)\mathbf{V}_T + \mathbf{R}^i_p,\\ &\mathrm{Softmax}\left(\frac{\mathbf{Q}_T\mathbf{K}^{\intercal}_R}{\alpha}\right)\mathbf{V}_R + \mathbf{T}^i_p\Bigg)\Bigg),\end{aligned} \tag{5}$$

where $\mathrm{Softmax}(\cdot)$, $\mathrm{Conv}_3(\cdot)$ and $\mathrm{Concat}(\cdot)$ represent the softmax function, 3×3 convolution, and concatenation operation, respectively. α is a hyperparameter used to control the smoothness of tensor calculations. It is not a manually tuned hyperparameter but a 'learnable adaptive parameter' that is optimized jointly with the rest of the network during training. We adopt the transposed self-attention mechanism to reduce computational complexity, which can reduce the original computational complexity from $\mathcal{O}\left(h^2w^2\right)$ to $\mathcal{O}\left(c^2\right)$. For simplicity, the concept of multi-head attention is omitted in Eq. (5).

Finally, similar to the classic Transformer [46,54], CP^{2}T feeds $\boldsymbol{A}^i$ into a feed-forward network for further feature enhancement. We adopt the Gated-Dconv Feed-forward Network (GDFN) proposed by Restormer [55] as the feed-forward network of CP^{2}T, which has advantages such as controlling information flow and forcing the model to focus on different details. The entire forward process can be represented by the following equation:

$$\boldsymbol{P}_i = \boldsymbol{A}_i + \mathrm{GDFN}\left(\mathrm{LN}\left(\boldsymbol{A}_i\right)\right), \tag{6}$$

where $\mathrm{GDFN}(\cdot)$ represents the GDFN feed-forward network. $\boldsymbol{P}_i \in \mathbb{R}^{h\times w\times c}$ is the final output fusion feature of CP^{2}T, which fully integrates features from both modalities for final prediction.

3.4. Cross-level selective fusion module

While previous works commonly utilize simple fusion strategies such as element-wise addition or concatenation for the obtained fused features at each stage, these strategies may introduce noisy features that impact dense prediction tasks. Motivated by this concern, we introduce a Cross-level Selective Fusion (CSF) module. The CSF module is designed to efficiently and effectively fuse features from different levels, addressing the challenges associated with noisy feature integration in dense prediction tasks. The detailed structure of the CSF module is shown in Fig. 5.

Specifically, the CSF module takes the feature $\boldsymbol{P}_i$ from the current level and the fused feature $\boldsymbol{C}_{i+1}$ from the previous level as input features. It begins by performing an element-wise addition operation on these two features. Subsequently, a 3×3 convolution followed by a Sigmoid function is applied to output a weight map. The input features are then multiplied element-wise by the weight map to obtain weighted features $\boldsymbol{P}_i$ and $\boldsymbol{C}_{i+1}$. Next, the module learns features through separate residual connections and a 3×3 convolution. Finally, an element-wise addition operation and a 3×3 Convolution-Batch Normalization-ReLU (CBR) component yield the final output $\boldsymbol{C}_i$. The forward process can be represented by the following equations:

$$\begin{aligned}\boldsymbol{W}_i &= \sigma\left(\mathrm{Conv}_3\left(\boldsymbol{P}_i + \delta_{\uparrow}\left(\boldsymbol{C}_{i+1}\right)\right)\right),\\ \boldsymbol{I}_i &= \mathrm{Conv}_3\left(\boldsymbol{P}_i + \boldsymbol{P}_i \otimes \boldsymbol{W}_i\right),\\ \boldsymbol{I}_{i+1} &= \mathrm{Conv}_3\left(\boldsymbol{C}_{i+1} + \boldsymbol{C}_{i+1} \otimes \boldsymbol{W}_i\right),\\ \boldsymbol{C}_i &= \mathrm{CBR}_3\left(\boldsymbol{I}_i + \boldsymbol{I}_{i+1}\right),\end{aligned} \tag{7}$$

where $\boldsymbol{W}_i$, $\boldsymbol{I}_i$, $\boldsymbol{I}_{i+1}$, and $\boldsymbol{C}_i$ are intermediate variables. $\mathrm{Conv}_3(\cdot)$ denotes a 3×3 convolution, $\delta_{\uparrow}(\cdot)$ represents upsampling operation, CBR_3 denotes a 3×3 CBR component, and $\sigma(\cdot)$ is the Sigmoid function. The final $\boldsymbol{C}_i$ is further fed into a CBR_3 component to generate new $\boldsymbol{C}_i$ for the fusion input of the next stage.

Finally, Du-CIPT employs a 1×1 convolution for pixel-level prediction on $\boldsymbol{C}_i$:

$$\boldsymbol{O}_i = \mathrm{Conv}_1\left(\boldsymbol{C}_i\right), \quad i \in \{1, 2, 3, 4\}, \tag{8}$$

where $\boldsymbol{O}_i$ represents the final output.

3.5. Loss functions

We focus on two types of dense prediction tasks: RGB-T salient object detection and RGB-T semantic segmentation. The former is a binary classification task, dividing pixels into foreground or background, and the latter is a multi-class task, assigning each pixel to a fixed category. Different loss functions should be adopted for these two tasks. We will introduce the loss functions used for both tasks.

3.5.1. Loss function for RGB-T salient object detection

For the RGB-T salient object detection task, we adhere to the deep supervision strategy proposed by Hou et al. [56], where each side output is supervised. Thus, the overall loss is formulated as:

$$\mathcal{L}_{sod} = \sum_{i=1}^{4} \alpha_i \mathcal{L}_{sod_i}, \tag{9}$$

Here, $\mathcal{L}_{sod_i}$ represents the loss of the ith side output, and α_i denotes the weight for each loss. In our study, α_i is set to 1. $\mathcal{L}_{sod}$ indicates the

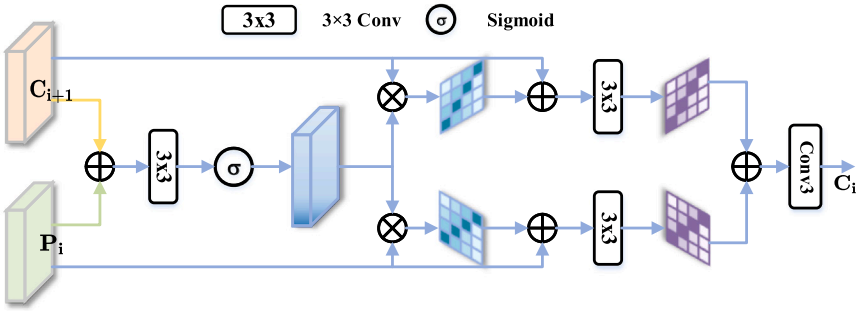

Fig. 5. The details of CSF.

total loss. Additionally, following previous work by Zhao et al. [50], the loss is defined as follows:

$$\mathcal{L}_{sod_i} = \mathcal{L}^i_{bce}\left(\boldsymbol{O}_i, \boldsymbol{G}\right) + \mathcal{L}^i_{smooth}\left(\boldsymbol{O}_i, \boldsymbol{G}\right) + \mathcal{L}^i_{iou}\left(\boldsymbol{O}_i, \boldsymbol{G}\right), \tag{10}$$

where $\mathcal{L}^i_{bce}(\cdot)$ represents binary cross-entropy loss function, $\mathcal{L}^i_{smooth}(\cdot)$ represents Smooth loss function [57], and $\mathcal{L}^i_{iou}(\cdot)$ represents IoU loss function [58]. The three loss functions respectively supervise the training model at the pixel, region, and global levels. $\boldsymbol{G}$ denotes the binary ground truth (GT).

3.5.2. Loss function for RGB-T semantic segmentation

For the RGB-T semantic segmentation task, we also adopt the multi-side deep supervision strategy and follow previous work by Zhang et al. [51,59] using a composite loss function composed of weighted binary cross-entropy loss and Lovasz-softmax [60] loss functions. Therefore, the loss function for the entire training process can be expressed as follows:

$$\mathcal{L}_{ss} = \sum_{i=1}^{4} \mathcal{L}^i_{wbce}\left(\boldsymbol{O}_i, \boldsymbol{G}_{SS}\right) + \mathcal{L}^i_{Lovasz}\left(\boldsymbol{O}_i, \boldsymbol{G}_{SS}\right). \tag{11}$$

Here, $\mathcal{L}_{wbce}$ and $\mathcal{L}_{Lovasz}$ respectively represent weighted binary cross-entropy loss and Lovasz-softmax loss functions. $\boldsymbol{G}_{SS}$ is the ground truth segmentation map. $\mathcal{L}_{ss}$ represents the total loss for RGB-T semantic segmentation.

4. Experiments

4.1. Settings

In this section, we introduce the datasets, experimental details, and evaluation metrics for both RGB-T salient object detection and RGB-T semantic segmentation.

4.1.1. RGB-T salient object detection: Datasets, experimental details, comparison methods, and evaluation metrics

Datasets. We follow the work of previous researchers [25,28,41,50,61], which uses 2500 pairs of RGB-T image sets from the VT5000 dataset [27] for training, while the remaining 2500 pairs, along with VT1000 [61] and VT821 [35] datasets, are used for testing.

Experiments details. The Du-CIPT model proposed is trained on an NVIDIA GTX 3080Ti GPU with 12 GB of memory, using the PyTorch framework. The backbone is initialized with a ResNet-50 [48] pre-trained on ImageNet [62]. AdamW optimizer [63] is employed with an initial learning rate of $8e-5$, following a StepLR decay strategy where the learning rate is multiplied by 0.1 every 40 epochs, with weight decay set to $1e-6$. All input images are resized to 352×352. The batch size is 10 and is trained for a total of 150 epochs. Empirical studies determine pyramid pooling rates as $\{12, 16, 20, 24\}$, $\{6, 8, 10, 12\}$, $\{3, 4, 5, 6\}$, $\{1, 2, 3, 4\}$. Data augmentation strategies such as random flipping, random cropping, and color enhancement are employed during training.

Baselines. To further validate the effectiveness of the proposed method, Du-CIPT is compared with 22 state-of-the-art (SOTA) methods, which can be categorized into four groups: (1) Three RGB salient object detection (SOD) methods, including PoolNet [64], R3Net [65], and CPD [66]; (2) Four RGB-D SOD methods, including MMCI [67], TANet [68], S2MA [69], and JL-DCF [70]; (3) Three traditional RGB-T SOD methods, including SGDL [61], M3S [71], and MTMR [35]; (4) Twelve deep learning-based RGB-T SOD methods, including FMCF [49], APNet [24], MIDD [25], MIA [26], ADF [27], ECFFNet [28], MMNet [72], CSRNet [50], CGFNet [40], LSNet [29], MSEDNet (ResNet-50 Version) [73], and SFMNet (ResNet-50 Version) [74]. All RGB and RGB-D methods are retrained on RGB-T datasets for fair comparison. Prediction maps are obtained from their open codes and pre-trained weights.

Evaluation metrics. Following the existing works, we employ five widely used evaluation metrics to assess the proposed model, including S-measure (S_m) [75], maximum F-measure (F_β) [76], maximum E-measure (E_ξ) [77], mean absolute error (MAE) [78], and Precision–Recall (PR) and F-measure curves.

4.1.2. RGB-T semantic segmentation: Datasets, experimental details, comparison methods, and evaluation metrics

Datasets. Currently, two widely used RGB-T semantic segmentation datasets include MFNet dataset [17] and PST900 dataset [14].

Experimental details. The proposed Du-CIPT is trained on an NVIDIA GTX 3080Ti GPU with 12 GB memory, implemented using the PyTorch framework. Empirical studies determine pyramid pooling rates as $\{12, 16, 20, 24\}$, $\{6, 8, 10, 12\}$, $\{3, 4, 5, 6\}$, $\{1, 2, 3, 4\}$. The backbones are initialized with a ResNet-50 pre-trained on ImageNet [62]. For Du-CIPT based ResNet-50, AdamW optimizer [63] is utilized with an initial learning rate of $8e-5$ and follows a StepLR decay strategy where the learning rate is multiplied by 0.5 every 20 epochs, with weight decay set to $5e-4$. The batch size is 3 and is trained for a total of 150 epochs. Data augmentation strategies such as color jittering, random horizontal flipping, and scaling data augmentation strategies are used during training.

Baselines. In this section, Du-CIPT is compared with relevant state-of-the-art RGB/RGB-D/RGB-T semantic segmentation methods. To ensure a fair comparison, following previous works [16,23], several semantic segmentation methods are modified to adapt to RGB-T datasets. RGB/RGB-D/RGB-T semantic segmentation methods are retrained under the same RGB-T semantic segmentation datasets settings. For the MFNet dataset, 16 advanced methods are selected for comparison, including two RGB semantic segmentation methods, DANet [79], HRNet [80], and their revised versions; four RGB-D semantic segmentation methods, namely FuseNet [81], D-CNN [82], ACNet [83], and SA-Gate [84]; ten RGB-T semantic segmentation methods, including MFNet [17], RTFNet [18], PSTNet [14], MLFNet [45], FuseSeg [19], ABMDRNet [23], MMNet [21], EGFNet [16], SFAF-MA [85], and MFNet [86].

Table 1
Du-CIPT and 22 advanced methods are compared on these three datasets regarding the $S_m \uparrow$, maximum F-measure ($F_\beta \uparrow$), maximum E-measure ($E_\xi \uparrow$), and MAE $\downarrow$ values. The best and second-best results in each column are highlighted in red and blue, respectively. "–" indicates that the result is not available.

Type	Method	Publication	VT5000 [27]				VT1000 [61]				VT821 [35]			
			$S_m \uparrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	MAE $\downarrow$	$S_m \uparrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	MAE $\downarrow$	$S_m \uparrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	MAE $\downarrow$
RGB SOD	PoolNet [64]	CVPR2019	0.770	0.725	0.856	0.089	0.834	0.828	0.904	0.067	0.751	0.690	0.823	0.109
	R3Net [65]	AAAI2018	0.757	0.671	0.825	0.084	0.842	0.805	0.896	0.055	0.787	0.709	0.844	0.073
	CPD [66]	ICCV2019	0.847	0.800	0.899	0.050	0.906	0.889	0.944	0.032	0.827	0.758	0.867	0.057
RGB-D SOD	MMCI [67]	PR2019	0.820	0.759	0.884	0.056	0.879	0.853	0.930	0.040	0.759	0.661	0.804	0.089
	TANet [68]	TIP2019	0.843	0.795	0.897	0.047	0.899	0.879	0.939	0.031	0.816	0.734	0.855	0.052
	S2MA [69]	CVPR2020	0.854	0.808	0.894	0.054	0.919	0.911	0.952	0.030	0.811	0.763	0.840	0.098
	JL-DCF [70]	CVPR2020	0.862	0.830	0.912	0.050	0.913	0.913	0.957	0.030	0.839	0.817	0.889	0.076
Traditional RGB-T SOD	SGDL [61]	TMM2020	0.750	0.695	0.829	0.089	0.787	0.770	0.859	0.090	0.765	0.735	0.839	0.085
	M3S [71]	MIPR2020	0.652	0.596	0.760	0.168	0.726	0.735	0.828	0.145	0.723	0.738	0.837	0.140
	MTMR [35]	IGTA2018	0.680	0.613	0.792	0.114	0.706	0.715	0.836	0.119	0.725	0.690	0.812	0.108
Deep Learning-Based RGB-T SOD	FMCF [49]	TIP2020	0.814	0.756	0.866	0.056	0.874	0.850	0.921	0.037	0.761	0.667	0.811	0.081
	APNet [24]	TETCI2021	0.876	0.848	0.920	0.035	0.921	0.913	0.955	0.021	0.867	0.825	0.909	0.034
	MIDD [25]	TIP2021	0.868	0.849	0.920	0.043	0.915	0.913	0.957	0.027	0.871	0.851	0.918	0.045
	MIA [26]	NC2022	0.879	0.865	0.930	0.040	0.924	**0.928**	**0.964**	0.025	0.844	0.831	0.902	0.070
	ADF [27]	TMM2022	0.864	0.837	0.911	0.048	0.910	0.908	0.951	0.034	0.810	0.752	0.839	0.077
	ECFFNet [28]	TCSVT2022	0.874	0.848	0.921	0.038	0.923	0.917	0.959	0.021	0.877	0.834	0.910	0.034
	MMNet [72]	TCSVT2022	0.863	–	–	0.043	0.917	–	–	0.027	0.874	–	–	0.040
	CSRNet [50]	TCSVT2022	0.868	0.837	0.914	0.042	0.918	0.908	0.953	0.024	**0.885**	**0.858**	**0.923**	0.038
	CGFNet [40]	TCSVT2022	**0.883**	**0.869**	**0.927**	**0.035**	0.923	0.923	0.959	0.023	0.881	**0.866**	0.920	0.038
	LSNet [29]	TIP2023	0.877	0.851	0.924	0.037	**0.925**	0.922	0.963	**0.023**	0.878	0.844	0.921	**0.033**
	MSEDNet [73]	NN2024	0.831	0.887	0.868	0.048	0.889	0.894	0.924	0.026	0.853	0.799	0.896	0.034
	SFMNet [74]	NN2024	0.812	0.797	0.857	0.078	0.811	0.857	0.901	0.066	0.826	0.836	0.837	0.071
	Du-CIPT	2025	**0.888**	**0.865**	**0.933**	**0.032**	**0.932**	**0.931**	**0.968**	**0.019**	**0.887**	0.856	**0.925**	**0.033**

For the PST900 dataset, 12 advanced methods are selected for comparison, including three semantic segmentation methods, ERFNet [87], CCNet [88], Efficient-FCN [89], and their revised versions; two RGB-D semantic segmentation methods, ACNet [83] and SA-Gate [84]; six RGB-T semantic segmentation methods, including MFNet [17], PST-Net [14], MFFENet [90], EGFNet [16], MTANet [91], DSGBINet [92], and MFNet [86]. Prediction maps are obtained from their open codes and pre-trained weights.

We note that recent advanced fusion methods, such as EGFNet [31], CMNeXt [93], and CMX [47], are not included in our direct experimental comparison because they rely on substantially stronger backbones (e.g., MiT-B2/B4 [94], ConvNeXt [95], or ResNet-152 [48]) than the ResNet-50 used in our framework, which would make a direct comparison less fair. In addition, considering our hardware constraint of a single NVIDIA RTX 3080Ti GPU with 12 GB memory, we chose ResNet-50 as a more practical and balanced backbone to control computational complexity while maintaining competitive performance. Nevertheless, unlike these methods, Du-CIPT particularly emphasizes cross-modal purification before fusion and multi-scale cross-modal interaction for dense prediction.

Evaluation metrics. Two widely used evaluation metrics, mean Accuracy (mACC) and mean Intersection over Union (mIoU), are employed to evaluate the proposed and comparison models.

4.2. Performance comparison

4.2.1. RGB-T salient object detection performance comparison

Quantitative comparison. The results of Du-CIPT compared to other advanced methods regarding S_m, maximum F-measure (F_β), maximum E-measure (E_ξ), and MAE values are shown in Table 1. From the results in the table, it can be observed that Du-CIPT outperforms other methods in almost all four metrics, except for slightly lower F_β values compared to CGFNet and CSRNet on the VT5000 and VT821 datasets. In terms of S_m values, Du-CIPT achieves performance gains of 0.57%, 0.43%, and 0.57% compared to the second-best methods on the VT5000, VT1000, and VT821 datasets, respectively. Regarding the MAE metrics, Du-CIPT achieves the best MAE values of 0.032, 0.019, and 0.033, with notable performance gains of 8.57% and 17.39% on the VT5000 and VT1000 datasets, respectively. In terms of the E_ξ metrics, Du-CIPT also achieves performance gains of 0.65%, 0.41%, and 0.22% compared to the second-best methods.

Furthermore, compared to methods based solely on RGB or RGB-D data, RGB-T SOD methods leveraging deep learning techniques have shown substantial performance enhancements. Despite the strong similarities between RGB-D and RGB-T SOD methods, the latter exhibit a significant advantage in handling RGB-T SOD tasks. For example, the widely used RGB-D SOD method JL-DCF [70] achieves impressive results in RGB-D SOD tasks. However, when adapted to RGB-T SOD tasks, it often performs poorly, primarily due to the inherent differences between depth maps and thermal infrared maps. In RGB-D SOD tasks, pixel values in depth maps typically correspond to salient objects, a characteristic absent in thermal infrared maps. Directly transferring RGB-D SOD methods to RGB-T SOD tasks results in sub-optimal performance. Notably, some RGB-D SOD methods, such as MMCI [67] and TANet [68], even underperform compared to RGB SOD methods. This observation underscores the need for future research to explore joint training approaches for RGB-D and RGB-T SOD tasks.

Compared to traditional RGB-T SOD methods, deep learning-based RGB-T SOD methods exhibit significant advantages, indicating the powerful feature representation capability of deep learning. Among all deep learning-based RGB-T SOD methods, Du-CIPT achieves the best performance.

To quantitatively evaluate the performance of the proposed method, PR and F-measure curves are presented in Fig. 6. A higher curve indicates better performance of the corresponding model. It is evident that Du-CIPT surpasses competing advanced methods in terms of performance. Additionally, the minimum recall value of the PR curve can indicate the robustness of the model. From the figure, it can be observed that Du-CIPT maintains relatively high recall rates even near the maximum threshold, suggesting that its predicted saliency maps closely resemble the ground truth. Moreover, Du-CIPT achieves the best trade-off performance between precision and recall across these three datasets.

Qualitative comparison. To further demonstrate the effectiveness of the proposed Du-CIPT, we conduct a qualitative comparison with 20 other advanced methods on three datasets, as depicted in Fig. 7. From the figure, it is evident that Du-CIPT accurately identifies salient objects while effectively removing non-salient parts. Notably, it exhibits better robustness in challenging scenarios, such as handling large objects (as illustrated in the 2nd row of VT5000, the 1st row of VT1000, and the 1st row of VT821), small objects (as demonstrated in the 2nd row of VT1000), and multiple objects (as depicted in the 1st row of VT5000). Moreover, even in the presence of irregularly shaped complex objects and low-light conditions, Du-CIPT accurately segments objects, as evidenced by the 2nd and 3rd rows of VT821 and the third row of VT5000. These visual results showcase that under the interaction of RGB and thermal infrared modalities and leveraging the proposed

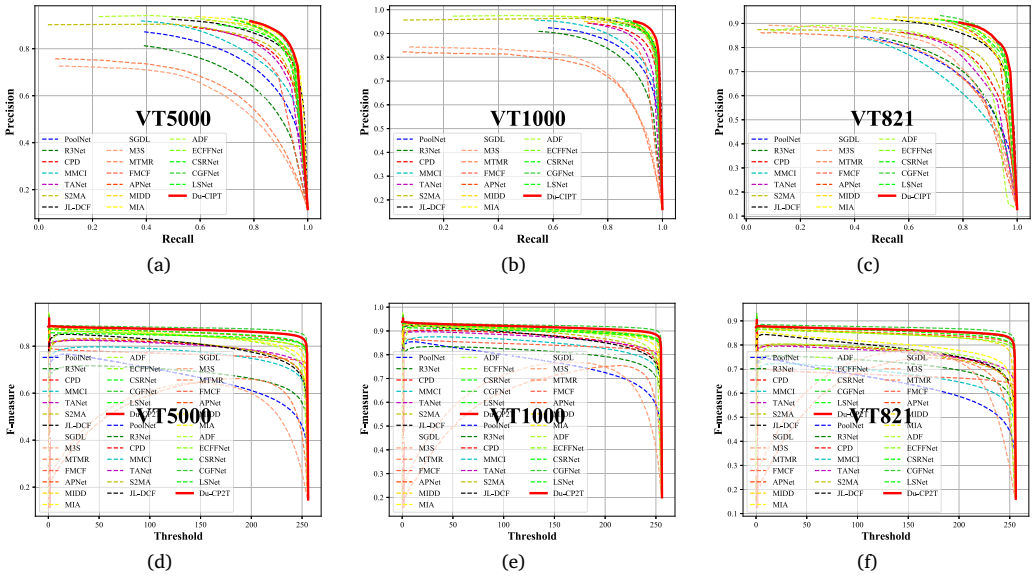

Fig. 6. The PR and F-measure curves of Du-CIPT compared to other SOTA methods on the three datasets.

Fig. 7. Qualitative comparison between Du-CIPT and 20 state-of-the-art methods on these three datasets.

multi-scale long-range interaction method, our model accurately identifies objects of various scales and complex shapes, demonstrating superior performance compared to other methods.

4.2.2. RGB-T semantic segmentation performance comparison

Quantitative comparison. The quantitative results of Du-CIPT compared to 16 other state-of-the-art methods on the MFNet dataset are presented in Table 2. These results primarily include the ACC and IoU metrics for eight categories, as well as mACC and mIoU metrics. It is evident from the table that Du-CIPT achieves the best performance in terms of mACC and mIoU, surpassing other methods. Specifically, out of the 18 metrics, Du-CIPT attains the highest value in seven and the second-highest value in one, indicating its strong adaptability across different scenarios. Moreover, Du-CIPT outperforms other methods comprehensively in overall performance, with performance gains of 0.69% and 2.74% in mACC and mIoU compared to the state-of-the-art EGFNet, respectively. The comparison confirms the significant superiority of Du-CIPT over existing methods. Notably, when considering all categories, Du-CIPT excels in the Person and Color Cone categories, achieving performance gains of 4.75% and 11.35% compared to the MFNet method, respectively. Furthermore, the Color Cone category has a relatively low proportion among all categories, suggesting that Du-CIPT demonstrates robust long-tail learning capabilities, outperforming other methods in this category.

In addition, we provide a performance comparison between our method and other advanced methods in Daytime and Nighttime scenes, as illustrated in Table 3. The table shows that Du-CIPT achieves the best performance in both scenarios, particularly excelling in the Nighttime scene. This underscores the success of the proposed multi-scale long-range pyramid pooling method in fully leveraging the advantages of thermal infrared maps for image segmentation in low-light conditions. The comprehensive analysis presented confirms the effectiveness of Du-CIPT on the MFNet dataset.

For the PST900 dataset, Du-CIPT is compared with 12 advanced methods, and the comparative results are presented in Table 4. The results primarily include the performance of five categories and the overall performance. Notably, Du-CIPT achieves the best performance on the PST900 dataset, particularly surpassing the state-of-the-art method DSGBINet [92] by 2.33% in terms of mIoU, further affirming the effectiveness of the proposed method.

Moreover, across all 12 metrics, Du-CIPT attains the highest values in four metrics and the second-highest values in three metrics, indicating its comprehensive superior performance. Specifically, Du-CIPT achieves the highest IoU in the Background and Fire-Extinguisher categories, notably surpassing DSGBINet [92] and achieving a performance gain of 7.25% in the Fire-Extinguisher category. This underscores the effectiveness of the proposed method, particularly in segmenting small objects that are not easily detected, such as Fire-Extinguishers. The comprehensive analysis presented further confirms the effectiveness of Du-CIPT and its ability to adapt to multiple datasets.

Table 2
Du-CIPT and 16 methods are compared on the MFNet dataset in terms of mean Accuracy (mACC, %) and mean Intersection over Union (mIoU, %). The best and second-best results in each column are highlighted in red and blue respectively. "–" indicates that the result is not available. "C-S" stands for Car Stop, "G-u" stands for Guardrail, "C-C" stands for Color Cone.

Type	Method	Publication	Car		Person		Bike		Curve		C-S		G-u		C-C		Bump		mAcc	mIoU
			Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU		
RGB SS	DANet [79]	CVPR2019	89.8	84.5	65.2	55.0	76.5	62.6	44.2	33.4	32.7	27.4	2.8	0.9	46.6	41.9	56.0	44.5	57.0	49.7
RGB-T SS	DANet [79]	CVPR2019	91.3	71.3	82.7	48.1	79.2	51.8	48.0	30.2	25.5	18.2	5.2	0.7	47.6	30.3	19.9	18.8	55.2	41.3
RGB SS	HRNet [80]	CVPR2019	91.1	84.9	66.6	55.4	76.6	60.3	42.6	33.3	37.9	28.3	11.5	2.5	44.8	40.3	62.6	46.9	59.2	49.9
RGB-T SS	HRNet [80]	CVPR2019	90.8	86.9	75.1	67.3	70.2	59.2	39.1	35.3	28.0	23.1	12.1	1.7	50.4	46.6	55.8	47.3	57.9	51.7
RGB-D SS	FuseNet [81]	ACCV2016	81.0	75.6	75.2	66.3	64.5	51.9	51.0	37.8	28.7	15.0	0.0	0.0	31.1	21.4	51.9	45.0	52.4	45.6
	D-CNN [82]	ECCV2018	85.2	77.0	61.7	53.4	76.0	56.5	40.2	30.9	9.9	29.3	22.8	6.4	32.9	30.1	36.5	32.3	55.1	46.1
	ACNet [83]	ICIP2019	93.7	79.4	86.8	64.7	77.8	52.7	57.2	32.9	51.5	28.4	7.0	0.8	57.5	16.9	49.8	44.4	64.3	46.3
	SA-Gate [84]	ECCV2020	86.0	73.8	80.8	59.2	69.4	51.3	56.7	38.4	24.7	19.3	0.0	0.0	56.9	24.5	52.1	48.8	58.3	45.8
RGB-T SS	MFNet [17]	IROS2017	77.2	65.9	67.0	58.9	53.9	42.9	36.2	29.9	19.1	9.9	0.1	8.5	30.3	25.2	30.0	27.7	45.1	39.7
	RTFNet50 [18]	RAL2019	91.3	86.3	78.2	67.8	71.5	58.2	69.8	43.7	32.1	24.3	13.4	3.6	40.4	26.0	73.5	57.2	62.2	51.7
	RTFNet152 [18]	RAL2019	93.0	87.4	79.3	70.3	76.8	62.7	60.7	45.3	38.5	29.8	0.0	0.0	45.5	29.1	74.7	55.7	63.1	53.2
	PSTNet [14]	ICRA2020	–	76.8	–	52.6	–	55.3	–	29.6	–	25.1	–	15.1	–	39.4	–	45.0	–	48.4
	MLFNet [45]	Meas2021	–	82.3	–	68.1	–	67.3	–	27.3	–	30.4	–	15.7	–	55.6	–	40.1	–	53.8
	FuseSeg [19]	TASE2021	93.1	87.9	81.4	71.7	78.5	64.6	68.4	44.8	29.1	22.7	63.7	6.4	55.8	46.9	66.4	47.9	70.6	54.5
	ABMDRNet [23]	CVPR2021	94.3	84.8	90.0	69.6	75.7	60.3	64.0	45.1	44.1	33.1	31.0	5.1	61.7	47.4	66.2	50.0	69.5	54.8
	MMNet [21]	APIN2022	–	83.9	–	69.3	–	59.0	–	43.2	–	24.7	–	4.6	–	42.2	–	50.7	62.7	52.8
	EGFNet [16]	AAAI2022	95.8	87.6	89.0	69.8	80.6	58.8	71.5	42.8	48.7	33.8	33.6	7.0	65.3	48.3	71.1	47.1	72.7	54.8
	SFAF-MA [85]	TIM2023	94.3	87.8	83.9	72.4	72.0	59.5	64.4	46.0	34.0	24.7	35.6	4.3	55.8	39.1	67.9	52.6	67.5	53.8
	MFNet [86]	OLE2024	95.7	87.8	91.7	69.5	82.3	64.2	71.1	43.8	36.6	27.4	39.3	6.2	66.8	50.2	68.7	52.8	72.3	55.5
	Du-CIPT	2025	93.4	86.5	92.1	72.8	82.4	61.6	66.4	44.9	34.5	26.2	59.3	6.6	67.0	55.9	69.0	55.3	73.2	56.3

Table 3
Du-CIPT and 12 methods are compared on the MFNet dataset for Daytime and Nighttime scenes in terms of mean accuracy (Accuracy, ACC, %) and mean Intersection over Union (IoU, %). The best results in each column are highlighted in red. "–" indicates that the result is not available.

Type	Method	Publication	Daytime		Nighttime	
			Acc	IoU	Acc	IoU
RGB SS	DANet [79]	CVPR2019	61.0	46.3	52.6	47.0
RGB SS	DANet [79]	CVPR2019	50.9	37.5	52.4	40.1
RGB SS	HRNet [80]	CVPR2019	64.7	46.7	54.0	47.3
RGB-T SS	HRNet [80]	CVPR2019	54.4	46.1	55.1	50.7
RGB-D SS	FuseNet [81]	ACCV2016	49.5	41.0	48.9	43.9
	D-CNN [82]	ECCV2018	50.6	42.4	50.7	43.2
	ACNet [83]	ICIP2019	60.7	41.6	63.9	47.4
	SA-Gate [84]	ECCV2020	49.3	37.9	56.9	45.6
RGB-T SS	MFNet [17]	IROS2017	42.6	36.1	48.9	43.9
	RTFNet50 [18]	RAL2019	57.3	44.4	59.4	52.0
	RTFNet152 [18]	RAL2019	60.0	45.8	60.7	54.8
	MLFNet [45]	Meas2021	–	45.6	–	54.9
	FuseSeg [19]	TASE2021	62.1	47.8	67.3	54.6
	ABMDRNet [23]	CVPR2021	58.4	46.7	68.3	55.5
	EGFNet [16]	AAAI2022	74.4	47.3	68.0	55.0
	Du-CIPT		73.2	56.3	73.2	56.3

Table 4
Du-CIPT and 12 advanced methods are compared on the PST900 dataset in terms of mean Accuracy (mACC, %) and mean Intersection over Union (mIoU, %). The best and second-best results in each column are highlighted in red and blue respectively. "–" indicates that the result is not available. "H-D" stands for Hand-Drill, "Fi-Ex" stands for Fire-Extinguisher.

Type	Method	Publication	Background		H-D		Backpack		Fi-Ex		Survivor		mAcc	mIoU
			Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU		
RGB SS	ERFNet [87]	TITS2018	–	98.69	–	42.40	–	65.28	–	61.18	–	41.69	–	61.85
RGB-T SS	ERFNet [87]	TITS2018	–	98.73	–	52.76	–	68.08	–	58.79	–	34.38	–	62.55
RGB SS	CCNet [88]	ICCV2019	99.86	99.05	51.77	32.27	68.30	66.42	67.79	51.84	60.84	57.50	69.71	61.42
RGB-T SS	CCNet [88]	ICCV2019	99.59	98.74	54.09	51.01	75.96	72.95	88.06	73.80	49.45	33.52	73.43	66.00
RGB SS	Efficient-FCN [89]	ECCV2020	99.81	98.63	32.08	30.12	60.06	58.15	78.87	39.96	32.76	28.00	60.72	50.98
RGB-T SS	Efficient-FCN [89]	ECCV2020	99.80	98.85	48.75	38.58	69.90	67.59	76.45	46.28	38.86	35.06	66.75	57.27
RGB-D SS	ACNet [83]	ICIP2019	99.83	99.25	53.59	51.46	85.56	83.19	84.88	59.95	69.10	65.19	78.67	71.81
	SA-Gate [84]	ECCV2020	99.74	99.25	89.88	81.01	89.03	79.77	80.70	72.97	64.19	62.22	84.71	79.05
RGB-T SS	MFNet [17]	IROS2017	–	98.63	–	41.13	–	64.27	–	60.35	–	20.70	–	57.02
	PSTNet [14]	ICRA2020	–	98.85	–	53.60	–	69.20	–	70.12	–	50.03	–	68.36
	MFFENet [90]	TMM2022	–	99.40	–	72.50	–	81.02	–	66.38	–	75.60	–	78.98
	EGFNet [16]	AAAI2022	99.48	99.26	97.99	64.67	94.17	83.05	95.17	71.29	83.30	74.30	94.02	78.51
	MTANet [91]	TIV2022	–	99.33	–	62.05	–	87.50	–	64.95	–	79.14	–	78.60
	DSGBINet [92]	TCSVT2023	99.73	99.39	94.53	74.99	88.65	85.11	94.78	79.31	81.37	75.56	91.81	82.87
	MFNet [86]	OLE2024	99.88	99.42	90.75	84.47	80.54	78.78	91.10	79.13	78.93	70.58	88.24	82.47
	Du-CIPT	2025	99.78	99.47	91.80	76.54	89.58	83.63	92.72	85.06	84.02	79.11	91.60	84.80

Qualitative comparison. To further demonstrate the effectiveness of the proposed Du-CIPT, we provide qualitative comparisons between Du-CIPT and advanced methods on the two datasets in Fig. 8. In terms of the MFNet dataset, the first four rows depict examples from the Daytime scene, while the latter four rows showcase examples from the Nighttime scene. In the Daytime scene examples, all cases show complex backgrounds. However, Du-CIPT successfully segments the target objects, including less conspicuous objects like the "Bump" object in the fourth case, which other methods fail to segment. This success is attributed to the proposed CP^{2}T module, facilitating multi-scale and long-range interactions for accurate localization and identification of less conspicuous objects.

In contrast, the latter four cases from the Nighttime scene encounter low-light challenges. Even in RGB images, the expected targets might not be discernible to the human eye. However, these targets are clearly visible in the thermal infrared maps. Thanks to the proposed CPM module, Du-CIPT effectively calibrates features from RGB images under low-light conditions, generating efficient cross-modal fusion features to predict targets. From the images, it is evident that Du-CIPT accurately segments all objects, even in cases where the targets are very small,

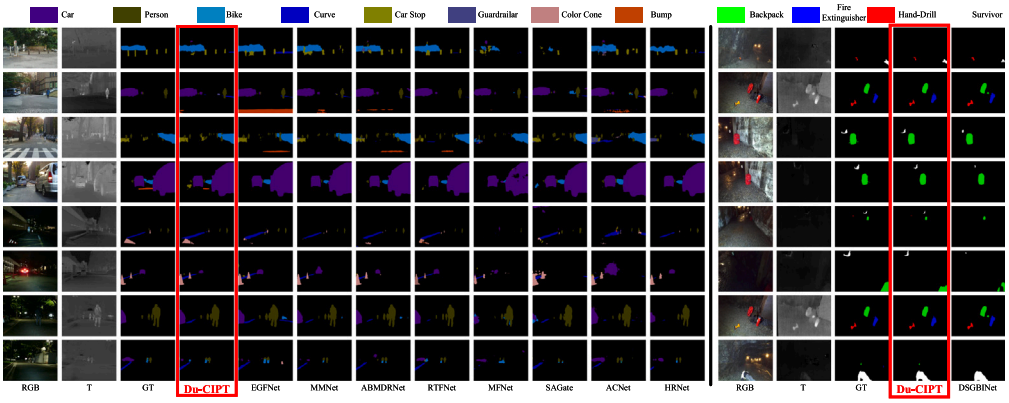

Fig. 8. Du-CIPT and the state-of-the-art methods are qualitatively compared through visualization of the two datasets.

Table 5
Du-CIPT and advanced methods are compared in terms of complexity, including FLOPs and parameters, as well as in terms of performance metrics, specifically F_β and mean Intersection over Union (mIoU, %), on the VT1000 dataset, MFNet, and PST900 dataset. The comparison results are highlighted with the best result in each column marked in red. "–" indicates that the result is not available.

RGB-T SOD					
Method	Publication	Input Size	Param (M)	FLOPs (G)	F_β
CGFNet [40]	TCSVT2022	352 × 352	–	139.97	0.923
APNet [24]	TETCI2021	352 × 352	–	46.48	0.913
MIDD [25]	TIP2021	352 × 352	–	57.44	0.913
ADF [27]	TMM2022	352 × 352	–	474.71	0.908
SFMNet [74]	NN2024	352 × 352	52.36	146.92	0.857
Du-CIPT		352 × 352	50.89	**29.20**	**0.931**
RGB-T SS					
Method	Publication	Input Size	Param (M)	FLOPs (G)	mIoU(%)
RTFNet50 [18]	RAL2019	640 × 480	185.24	245.71	51.70
RTFNet152 [18]	RAL2019	640 × 480	254.51	337.04	53.20
PSTNet [14]	ICRA2020	640 × 480	**20.38**	129.37	48.40
FuseSeg [19]	TASE2021	640 × 480	141.52	193.40	54.50
ABMDRNet [23]	CVPR2021	640 × 480	64.60	194.33	54.80
EGFNet [16]	AAAI2022	640 × 480	62.82	201.29	54.80
MTANet [91]	TIV2022	640 × 480	121.58	264.69	56.10
MMNet [21]	APIN2022	640 × 480	23.90	95.00	52.80
Du-CIPT		640 × 480	55.00	**73.05**	**56.30**

such as the "Color Cone" object in the first case. In terms of the PST900 dataset, our method exhibits greater robustness in detecting both small objects and multiple objects compared to DSGBINet.

In summary, Du-CIPT demonstrates the ability to handle various complex scenarios and produce reliable segmentation results, showcasing excellent generalization capabilities.

4.2.3. Complexity

From the results presented in Table 5, it is shown that in the RGB-T SOD task, Du-CIPT achieves the highest F_β value with the lowest computational complexity of 29.20G FLOPs. Notably, it outperforms popular advanced methods such as CGFNet and ADF while costing half the floating-point operations, demonstrating superior efficiency without compromising performance. In the RGB-T SS task, Du-CIPT achieves the highest performance with a computational complexity of 73.05G FLOPs and a parameter count of 55.00 M, surpassing methods like PSTNet and MMNet. This indicates that the proposed method achieves superior performance with lower computational complexity, striking a favorable trade-off between complexity and performance. In addition to FLOPs and parameter counts, we further report the inference speed of Du-CIPT on a single NVIDIA RTX 3080Ti GPU. Specifically, the model runs at 41.3 FPS for RGB-T SOD with an input size of 352 × 352, and 18.7 FPS for RGB-T semantic segmentation with an input size of 640 × 480. These results indicate that Du-CIPT provides a favorable accuracy-efficiency trade-off, although the current implementation is mainly designed for high-quality dense prediction rather than strict real-time deployment.

4.3. Ablation study

The results of the ablation experiments on two RGB-T tasks are presented in Table 6. In this section, the effectiveness of each module is investigated by removing a single module from the entire model. Additionally, the results of each ablation experiment are visualized in Fig. 9.

4.3.1. Effectiveness of CPM

To demonstrate the effectiveness of the CPM module, we conducted an ablation study on this module, denoted as V4, as shown in the penultimate row of Table 6. Results indicate that without CPM, the performance of Du-CIPT decreases compared to the entire model. For instance, in the RGB-T SOD task, the F_β metric on the VT5000 dataset drops from 0.865 to 0.859, and on the VT821 dataset, it drops from 0.856 to 0.852. Similarly, in the RGB-T SS task, the mIoU on the MFNet dataset decreases from 56.3 to 55.8. These results underscore the importance of the CPM.

Furthermore, the ablation visualization results of this module are shown in the 5th row of Fig. 9. Notably, even in scenarios where RGB images are mostly noise or in low-light scenes, the CPM module effectively identifies targets, as evidenced by the examples of the person and bike in the 1st and 3rd columns. Despite the RGB image containing considerable noise, the complete model (denoted as V5) successfully recognizes the targets, while the model performs poorly in the absence of this module. In RGB-T scenes, RGB images often face challenges such as low-light conditions, while thermal infrared maps can provide excellent features. The CPM module enhances the features of RGB

Table 6
The ablation experiments on the VT5000, VT1000, VT821, and MFNet datasets. The best result in each column is highlighted in red. "B" denotes the baseline model. "w/o" indicates without the component, and "PP" denotes pyramid pooling.

Number	Method	RGB-T SOD												RGB-T SS
		VT5000				VT1000				VT821				MFNet
		$S_m \uparrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	MAE ↓	$S_m \uparrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	MAE ↓	$S_m \uparrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	MAE ↓	mIoU (%)
V1	w/o CSF + w/o PP	0.883	0.858	0.925	0.035	0.928	0.925	0.964	0.020	0.874	0.839	0.908	0.039	55.4
V2	w/o CSF	0.884	0.861	0.929	0.035	0.928	0.924	0.963	0.020	0.879	0.849	0.913	0.039	55.9
V3	w/o CP^{2}T	0.885	0.860	0.929	0.035	0.927	0.924	0.963	0.021	0.879	0.848	0.913	0.035	55.1
V4	w/o CPM	0.884	0.859	0.930	0.034	0.932	0.930	0.967	0.019	0.882	0.852	0.916	0.037	55.8
V5	Du-CIPT	**0.888**	**0.865**	**0.933**	**0.032**	**0.932**	**0.931**	**0.968**	**0.019**	**0.887**	**0.856**	**0.925**	**0.033**	**56.3**

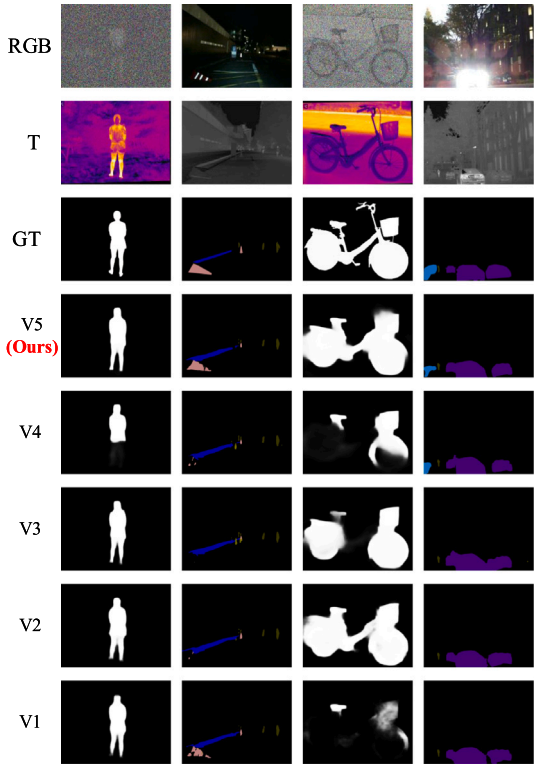

Fig. 9. Visualization of the ablation experiments.

images through mutual purification and calibration, thereby improving overall performance.

4.3.2. Effectiveness of CP^2 T module

To demonstrate the effectiveness of the proposed CP^{2}T module, we remove the CP^{2}T module from the entire model. The ablation results (denoted as V3) are presented in the third from the last row of Table 6. The table illustrates that the CP^{2}T module plays a crucial role, in achieving performance gains. For instance, on the VT1000 dataset, it increases the F_β score from 0.924 to 0.931. Similarly, the mIoU on the MFNet dataset rises from 55.1 to 56.3. These results suggest that demonstrate its effectiveness.

Moreover, the ablation visualization results of this module are depicted in the 6th rows of Fig. 9. Notably, with the effect of multi-scale and long-range interactions, the model can further predict the completeness of objects, albeit with some noise.

4.3.3. Effectiveness of CSF module

To confirm the effectiveness of the proposed CSF module, we remove the CSF module from the entire model and replace with direct addition. The ablation results (denoted as V2) are presented in the fourth row from the bottom of Table 6. The results in the table show that removing the CSF module will hurt the model performance. For instance, on the VT821 dataset, it decreases the F_β score from 0.856 to 0.849. Similarly, the mIoU on the MFNet dataset decreases from 56.3 to 55.9. These results suggest that demonstrate its effectiveness. It efficiently fuses features critical for final prediction while filtering out unimportant interference noise, thereby achieving performance gains. The ablation visualization results of this module can be seen in the 7th row of Fig. 9. Compared to the entire model, it contains noise, further confirming the effectiveness of this module.

4.3.4. Effectiveness of pyramid pooling

To assess the effectiveness of the Pyramid Pooling, we further remove the component from the V2 model (denoted as V1). The ablation results are presented in the 3rd row of Table 6. The table indicates a slight performance degradation with the absence of pyramid pooling. For instance, compared to the V2 model, the F_β score on the VT5000 dataset decreases from 0.861 to 0.858, and on the VT821 dataset, it decreases from 0.849 to 0.839. Similarly, the mIoU on the MFNet dataset increases from 55.4 to 55.9. These results highlight the effectiveness of multi-scale interaction in enhancing model performance. Through multi-scale interaction, better fusion features can be generated for final prediction.

It is worth noting that we use pyramid pooling to obtain multi-scale features, mainly because this method does not introduce additional parameters and therefore avoids increasing computational complexity. In contrast, methods such as dilated/atrous convolution (ASPP [96]), while capable of obtaining multi-scale features, may negatively impact the inference speed.

4.4. Discussion

4.4.1. Examples of failure

We show several failure cases in Fig. 10, where the model encounters difficulties in predicting salient objects and semantic segmentation: (1) In situations where an RGB image depicts both near and far objects yet fails to distinguish between them clearly, coupled with a thermal infrared map lacking distinctive highlights corresponding to object distances, depth perception becomes notably challenging. This difficulty can impede accurate object identification and segmentation. For instance, in the 1st scenario (illustrated in the top row of Fig. 10), a person in the distance appears more prominent in the thermal infrared image, while a nearby pole stands out more prominently in the RGB image. Consequently, the model struggles to correctly predict nearby objects, often favoring those more prominent in the infrared image and thus misclassifying nearby objects into incorrect categories.

(2) While RGB images are plagued by significant noise, the quality of thermal infrared maps is often subpar. In scenarios depicted in the 3rd row of Fig. 10, our model struggles to identify salient objects, such as the word "EN" in the image. However, when the quality of the thermal infrared map is higher, as demonstrated in the second row, the model can still accurately predict salient objects even when the RGB image is noisy. This observation underscores the effectiveness of the CPM module in successfully calibrating the other modality, particularly when one modality provides clearer information.

(3) When both RGB and thermal infrared maps suffer from poor quality, as depicted in the 4th row of Fig. 10, the performance diminishes significantly. This indicates that the model struggles when both modalities are simultaneously of low quality.

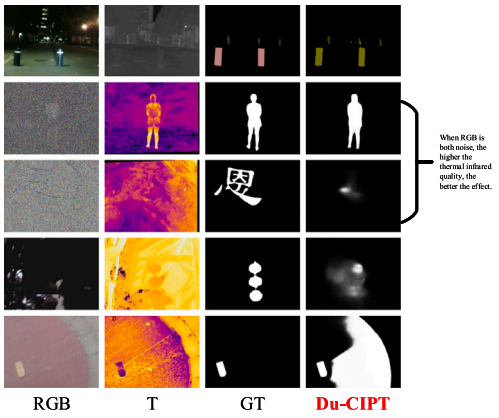

Fig. 10. Visualization failure cases of Du-CIPT.

(4) In cases where multiple salient regions are simultaneously presented in both RGB and thermal infrared maps, but only one salient object exists in the ground truth, as illustrated in the 5th row of Fig. 10, the model encounters difficulty. The model identifies both the red area and the highlighted region in the thermal infrared image as salient despite only one object being present in the ground truth. This discrepancy highlights the struggle to discern the final output object when faced with multiple salient regions.

4.4.2. Limitations and future work

Two limitations have been identified that warrant further investigation. Firstly, while Du-CIPT has achieved excellent performance, its reliance on pixel-level labels renders it unsuitable for semi-supervised or unsupervised learning in RGB-T dense prediction tasks. Recent research indicates a growing interest in these approaches, underscoring the need to adapt methods to accommodate such scenarios.

Secondly, the thermal infrared modality's pixel values solely reflect surface temperature distribution, which may not always correlate positively with object characteristics. Consequently, relying solely on RGB and thermal infrared modalities for object prediction proves inadequate. As demonstrated in Fig. 10, the simultaneous distribution of multiple objects in the RGB image at varying distances leads to poor model performance. Therefore, integrating depth information into object prediction may enhance performance, with further improvements achievable through the combined influence of all three modalities.

In conclusion, RGB-T dense prediction tasks offer fertile ground for continued exploration and refinement.

5. Conclusion

In this paper, we introduce the Du-CIPT model for RGB-T salient object detection and semantic segmentation tasks, with a focus on enhancing the fusion of RGB-T bi-modalities through long-range and multi-scale interaction learning. Unlike previous fusion methods, our approach employs a Transformer equipped with dual pyramid pooling to more effectively capture long-range and multi-scale multi-modal interactions, thereby generating more comprehensive fusion features. Furthermore, in scenarios where the RGB modality exhibits poor feature representation under low-light conditions, our proposed CPM adeptly calibrates the feature representation, leading to improved performance. Additionally, the CSF module facilitates efficient integration of cross-level features, effectively filtering out unimportant interfering noise and achieving optimal prediction. Experimental results on five common datasets for both tasks validate the effectiveness of our proposed method. Looking ahead, we aim to explore RGB-D-T fusion methods for dense prediction tasks.

CRediT authorship contribution statement

Jiesheng Wu: Writing – review & editing, Writing – original draft, Methodology, Investigation, Data curation, Conceptualization. **Ji Du:** Formal analysis. **Fangwei Hao:** Validation. **Jiankang Hong:** Visualization.

Ethical approval

This research did not involve human participants or animals. Ethical approval was therefore not required.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (NSFC) under Grants No.62541601 and No.62306010.

Data availability

We will share my code and dataset on GitHub.

References

[1] R. Cong, J. Lei, H. Fu, M.-M. Cheng, W. Lin, Q. Huang, Review of visual saliency detection with comprehensive information, IEEE Trans. Circuits Syst. Video Technol. 29 (10) (2019) 2941–2959.
[2] S. Zhao, Q. Zhang, A feature divide-and-conquer network for RGB-T semantic segmentation, IEEE Trans. Circuits Syst. Video Technol. 33 (6) (2023) 2892–2905.
[3] J.-Y. Zhu, J. Wu, Y. Xu, E. Chang, Z. Tu, Unsupervised object class discovery via saliency-guided multiple class learning, IEEE Trans. Pattern Anal. Mach. Intell. 37 (4) (2015) 862–875.
[4] M. Donoser, M. Urschler, M. Hirzer, H. Bischof, Saliency driven total variation segmentation, in: 2009 IEEE 12th International Conference on Computer Vision, 2009, pp. 817–824.
[5] M. Niu, K. Song, L. Huang, Q. Wang, Y. Yan, Q. Meng, Unsupervised saliency detection of rail surface defects using stereoscopic images, IEEE Trans. Ind. Informatics 17 (3) (2021) 2271–2281.
[6] Y. Bao, K. Song, J. Liu, Y. Wang, Y. Yan, H. Yu, X. Li, Triplet-graph reasoning network for few-shot metal generic surface defect segmentation, IEEE Trans. Instrum. Meas. 70 (2021) 1–11.
[7] Y.J. Zhao, W.H. Xu, C.Z. Xi, D.T. Liang, H.N. Li, Automatic and accurate measurement of microhardness profile based on image processing, IEEE Trans. Instrum. Meas. 70 (2021) 1–9.
[8] Z. Song, L. Zhao, J. Zhou, Learning hybrid semantic affinity for point cloud segmentation, IEEE Trans. Circuits Syst. Video Technol. 32 (7) (2022) 4599–4612.
[9] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, B. Schiele, The cityscapes dataset for semantic urban scene understanding, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2016, pp. 3213–3223.
[10] Y. Tian, S. Zhu, Partial domain adaptation on semantic segmentation, IEEE Trans. Circuits Syst. Video Technol. 32 (6) (2022) 3798–3809.
[11] D.O. Medley, C. Santiago, J.C. Nascimento, CyCoSeg: A cyclic collaborative framework for automated medical image segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 44 (11) (2022) 8167–8182.
[12] J. Zhuang, Z. Wang, B. Wang, Video semantic segmentation with distortion-aware feature correction, IEEE Trans. Circuits Syst. Video Technol. 31 (8) (2021) 3128–3139.
[13] K. Song, Y. Zhao, L. Huang, Y. Yan, Q. Meng, RGB-T image analysis technology and application: A survey, Eng. Appl. Artif. Intell. 120 (2023) 105919.
[14] S.S. Shivakumar, N. Rodrigues, A. Zhou, I.D. Miller, V. Kumar, C.J. Taylor, PST900: RGB-thermal calibration, dataset and segmentation network, in: 2020 IEEE International Conference on Robotics and Automation, ICRA, 2020, pp. 9441–9447.

[15] F. Deng, H. Feng, M. Liang, H. Wang, Y. Yang, Y. Gao, J. Chen, J. Hu, X. Guo, T.L. Lam, Feanet: Feature-enhanced attention network for RGB-thermal real-time semantic segmentation, in: 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, 2021, pp. 4467–4473.

[16] W. Zhou, S. Dong, C. Xu, Y. Qian, Edge-aware guidance fusion network for RGB–thermal scene parsing, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, (3) 2022, pp. 3571–3579.

[17] Q. Ha, K. Watanabe, T. Karasawa, Y. Ushiku, T. Harada, Mfnet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes, in: 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, 2017, pp. 5108–5115.

[18] Y. Sun, W. Zuo, M. Liu, RTFNet: RGB-thermal fusion network for semantic segmentation of urban scenes, IEEE Robot. Autom. Lett. 4 (3) (2019) 2576–2583.

[19] Y. Sun, W. Zuo, P. Yun, H. Wang, M. Liu, FuseSeg: Semantic segmentation of urban scenes based on RGB and thermal data fusion, IEEE Trans. Autom. Sci. Eng. 18 (3) (2021) 1000–1011.

[20] J. Xu, K. Lu, H. Wang, Attention fusion network for multi-spectral semantic segmentation, Pattern Recognit. Lett. 146 (2021) 179–184.

[21] X. Lan, X. Gu, X. Gu, MMNet: Multi-modal multi-stage network for RGB-T image semantic segmentation, Appl. Intell. 52 (5) (2022) 5817–5829.

[22] W. Zhou, J. Liu, J. Lei, L. Yu, J.-N. Hwang, GMNet: Graded-feature multilabel-learning network for RGB-thermal urban scene semantic segmentation, IEEE Trans. Image Process. 30 (2021) 7790–7802.

[23] Q. Zhang, S. Zhao, Y. Luo, D. Zhang, N. Huang, J. Han, ABMDRNet: Adaptive-weighted bi-directional modality difference reduction network for RGB-T semantic segmentation, in: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2021, pp. 2633–2642.

[24] W. Zhou, Y. Zhu, J. Lei, J. Wan, L. Yu, APNet: Adversarial learning assistance and perceived importance fusion network for all-day RGB-T salient object detection, IEEE Trans. Emerg. Top. Comput. Intell. 6 (4) (2022) 957–968.

[25] Z. Tu, Z. Li, C. Li, Y. Lang, J. Tang, Multi-interactive dual-decoder for RGB-thermal salient object detection, IEEE Trans. Image Process. 30 (2021) 5678–5691.

[26] Y. Liang, G. Qin, M. Sun, J. Qin, J. Yan, Z. Zhang, Multi-modal interactive attention and dual progressive decoding network for RGB-D/T salient object detection, Neurocomputing 490 (2022) 132–145.

[27] Z. Tu, Y. Ma, Z. Li, C. Li, J. Xu, Y. Liu, RGBT salient object detection: A large-scale dataset and benchmark, IEEE Trans. Multimed. 25 (2023) 4163–4176.

[28] W. Zhou, Q. Guo, J. Lei, L. Yu, J.-N. Hwang, ECFFNet: Effective and consistent feature fusion network for RGB-T salient object detection, IEEE Trans. Circuits Syst. Video Technol. 32 (3) (2022) 1224–1235.

[29] W. Zhou, Y. Zhu, J. Lei, R. Yang, L. Yu, LSNet: Lightweight spatial boosting network for detecting salient objects in RGB-thermal images, IEEE Trans. Image Process. 32 (2023) 1329–1340.

[30] W. Zhou, S. Dong, M. Fang, L. Yu, CACFNet: Cross-modal attention cascaded fusion network for RGB-T urban scene parsing, IEEE Trans. Intell. Veh. 9 (1) (2024) 1919–1929.

[31] S. Dong, W. Zhou, C. Xu, W. Yan, EGFNet: Edge-aware guidance fusion network for RGB–thermal urban scene parsing, IEEE Trans. Intell. Transp. Syst. 25 (1) (2024) 657–669.

[32] S. Dong, W. Zhou, X. Qian, L. Yu, GEBNet: Graph-enhancement branch network for RGB-T scene parsing, IEEE Signal Process. Lett. 29 (2022) 2273–2277.

[33] S. Dong, Y. Feng, Q. Yang, Y. Huang, D. Liu, H. Fan, Efficient multimodal semantic segmentation via dual-prompt learning, 2023, arXiv preprint arXiv:2312.00360.

[34] W. Zhou, T. Gong, J. Lei, L. Yu, DBCNet: Dynamic bilateral cross-fusion network for RGB-T urban scene understanding in intelligent vehicles, IEEE Trans. Syst. Man, Cybern.: Syst. 53 (12) (2023) 7631–7641.

[35] G. Wang, C. Li, Y. Ma, A. Zheng, J. Tang, B. Luo, RGB-t saliency detection benchmark: Dataset, baselines, analysis and a novel approach, in: Image and Graphics Technologies and Applications: 13th Conference on Image and Graphics Technologies and Applications, IGTA 2018, Beijing, China, April 8–10, 2018, Revised Selected Papers 13, Springer, 2018, pp. 359–369.

[36] Z. Tu, T. Xia, C. Li, X. Wang, Y. Ma, J. Tang, RGB-t image saliency detection via collaborative graph learning, IEEE Trans. Multimed. 22 (1) (2019) 160–173.

[37] C. Xu, Q. Li, Q. Zhou, X. Jiang, D. Yu, Y. Zhou, Asymmetric cross-modal activation network for RGB-T salient object detection, Knowl.-Based Syst. 258 (2022) 110047.

[38] H. Wen, K. Song, L. Huang, H. Wang, Y. Yan, Cross-modality salient object detection network with universality and anti-interference, Knowl.-Based Syst. 264 (2023) 110322.

[39] Z. Tu, Z. Li, C. Li, Y. Lang, J. Tang, Multi-interactive dual-decoder for RGB-thermal salient object detection, IEEE Trans. Image Process. 30 (2021) 5678–5691.

[40] J. Wang, K. Song, Y. Bao, L. Huang, Y. Yan, CGFNet: Cross-guided fusion network for RGB-T salient object detection, IEEE Trans. Circuits Syst. Video Technol. 32 (5) (2022) 2949–2961.

[41] R. Cong, K. Zhang, C. Zhang, F. Zheng, Y. Zhao, Q. Huang, S. Kwong, Does thermal really always matter for RGB-T salient object detection? IEEE Trans. Multimed. 25 (2023) 6971–6982.

[42] Q. Wang, Y. Chi, T. Shen, J. Song, Z. Zhang, Y. Zhu, Improving RGB-infrared object detection by reducing cross-modality redundancy, Remote. Sens. 14 (9) (2022) 2020.

[43] Y. Pang, X. Zhao, L. Zhang, H. Lu, Caver: Cross-modal view-mixed transformer for bi-modal salient object detection, IEEE Trans. Image Process. 32 (2023) 892–904.

[44] X. Jiang, Y. Hou, H. Tian, L. Zhu, Mirror complementary transformer network for RGB-thermal salient object detection, IET Comput. Vis. 18 (1) (2024) 15–32.

[45] Z. Guo, X. Li, Q. Xu, Z. Sun, Robust semantic segmentation based on RGB-thermal in variable lighting scenes, Measurement 186 (2021) 110176.

[46] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, 2020, arXiv preprint arXiv:2010.11929.

[47] H. Liu, J. Zhang, K. Yang, X. Hu, R. Stiefelhagen, CMX: Cross-modal fusion for RGB-x semantic segmentation with transformers, 2022, arXiv preprint arXiv:2203.04838.

[48] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[49] Q. Zhang, N. Huang, L. Yao, D. Zhang, C. Shan, J. Han, RGB-T salient object detection via fusing multi-level CNN features, IEEE Trans. Image Process. 29 (2020) 3321–3335.

[50] F. Huo, X. Zhu, L. Zhang, Q. Liu, Y. Shu, Efficient context-guided stacked refinement network for RGB-T salient object detection, IEEE Trans. Circuits Syst. Video Technol. 32 (5) (2022) 3111–3124.

[51] G. Li, Y. Wang, Z. Liu, X. Zhang, D. Zeng, RGB-T semantic segmentation with location, activation, and sharpening, IEEE Trans. Circuits Syst. Video Technol. 33 (3) (2023) 1223–1235.

[52] Y. Lv, Z. Liu, G. Li, Context-aware interaction network for RGB-T semantic segmentation, IEEE Trans. Multimed. (2023) 1–13.

[53] Y.-H. Wu, Y. Liu, X. Zhan, M.-M. Cheng, P2T: Pyramid pooling transformer for scene understanding, IEEE Trans. Pattern Anal. Mach. Intell. 45 (11) (2022) 12760–12771.

[54] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, 2017, pp. 5998–6008.

[55] S.W. Zamir, A. Arora, S. Khan, M. Hayat, F.S. Khan, M.-H. Yang, Restormer: Efficient transformer for high-resolution image restoration, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 5728–5739.

[56] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, P.H. Torr, Deeply supervised salient object detection with short connections, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3203–3212.

[57] C. Godard, O. Mac Aodha, G.J. Brostow, Unsupervised monocular depth estimation with left-right consistency, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 270–279.

[58] G. Máttyus, W. Luo, R. Urtasun, Deeproadmapper: Extracting road topology from aerial images, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 3438–3446.

[59] Y. Wang, G. Li, Z. Liu, SGFNet: Semantic-guided fusion network for RGB-thermal semantic segmentation, IEEE Trans. Circuits Syst. Video Technol. 33 (12) (2023) 7737–7748.

[60] M. Berman, A.R. Triki, M.B. Blaschko, The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 4413–4421.

[61] Z. Tu, T. Xia, C. Li, X. Wang, Y. Ma, J. Tang, RGB-T image saliency detection via collaborative graph learning, IEEE Trans. Multimed. 22 (1) (2020) 160–173.

[62] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, ImageNet: A large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 248–255.

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

[64] J.-J. Liu, Q. Hou, M.-M. Cheng, J. Feng, J. Jiang, A simple pooling-based design for real-time salient object detection, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2019, pp. 3912–3921.

[65] Z. Deng, X. Hu, L. Zhu, X. Xu, J. Qin, G. Han, P.-A. Heng, R3net: Recurrent residual refinement network for saliency detection, in: Proceedings of the 27th International Joint Conference on Artificial Intelligence, AAAI Press Menlo Park, CA, USA, 2018, pp. 684–690.

[66] Z. Wu, L. Su, Q. Huang, Cascaded partial decoder for fast and accurate salient object detection, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2019, pp. 3902–3911.

[67] H. Chen, Y. Li, D. Su, Multi-modal fusion network with multi-scale multi-path and cross-modal interactions for RGB-D salient object detection, Pattern Recognit. 86 (2019) 376–385.

[68] H. Chen, Y. Li, Three-stream attention-aware network for RGB-d salient object detection, IEEE Trans. Image Process. 28 (6) (2019) 2825–2835.

[69] N. Liu, N. Zhang, J. Han, Learning selective self-mutual attention for RGB-D saliency detection, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2020, pp. 13753–13762.

[70] K. Fu, D.-P. Fan, G.-P. Ji, Q. Zhao, JL-DCF: Joint learning and densely-cooperative fusion framework for RGB-d salient object detection, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2020, pp. 3049–3059.
[71] Z. Tu, T. Xia, C. Li, Y. Lu, J. Tang, M3S-NIR: Multi-modal multi-scale noise-insensitive ranking for RGB-T saliency detection, in: 2019 IEEE Conference on Multimedia Information Processing and Retrieval, MIPR, 2019, pp. 141–146.
[72] W. Gao, G. Liao, S. Ma, G. Li, Y. Liang, W. Lin, Unified information fusion network for multi-modal RGB-D and RGB-T salient object detection, IEEE Trans. Circuits Syst. Video Technol. 32 (4) (2022) 2091–2106.
[73] D. Peng, W. Zhou, J. Pan, D. Wang, MSEDNet: Multi-scale fusion and edge-supervised network for RGB-T salient object detection, Neural Netw. 171 (2024) 410–422.
[74] H. Yue, J. Guo, X. Yin, Y. Zhang, S. Zheng, Salient object detection in low-light RGB-T scene via spatial-frequency cues mining, Neural Netw. 178 (2024) 106406.
[75] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, A. Borji, Structure-measure: A new way to evaluate foreground maps, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 4548–4557.
[76] R. Achanta, S. Hemami, F. Estrada, S. Susstrunk, Frequency-tuned salient region detection, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2009, pp. 1597–1604.
[77] D.-P. Fan, C. Gong, Y. Cao, B. Ren, M.-M. Cheng, A. Borji, Enhanced-alignment measure for binary foreground map evaluation, 2018, arXiv preprint arXiv:1805.10421.
[78] A. Borji, M.-M. Cheng, H. Jiang, J. Li, Salient object detection: A benchmark, IEEE Trans. Image Process. 24 (12) (2015) 5706–5722.
[79] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, H. Lu, Dual attention network for scene segmentation, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2019, pp. 3141–3149.
[80] K. Sun, B. Xiao, D. Liu, J. Wang, Deep high-resolution representation learning for human pose estimation, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2019, pp. 5686–5696.
[81] C. Hazirbas, L. Ma, C. Domokos, D. Cremers, Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture, in: Computer Vision–ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part I 13, Springer, 2017, pp. 213–228.
[82] W. Wang, U. Neumann, Depth-aware cnn for rgb-d segmentation, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 135–150.
[83] X. Hu, K. Yang, L. Fei, K. Wang, ACNET: Attention based network to exploit complementary features for RGBD semantic segmentation, in: 2019 IEEE International Conference on Image Processing, ICIP, 2019, pp. 1440–1444.
[84] X. Chen, K.-Y. Lin, J. Wang, W. Wu, C. Qian, H. Li, G. Zeng, Bi-directional cross-modality feature propagation with separation-and-aggregation gate for RGB-d semantic segmentation, in: European Conference on Computer Vision, Springer, 2020, pp. 561–577.
[85] X. He, M. Wang, T. Liu, L. Zhao, Y. Yue, SFAF-MA: Spatial feature aggregation and fusion with modality adaptation for RGB-thermal semantic segmentation, IEEE Trans. Instrum. Meas. 72 (2023) 1–10.
[86] J. Liu, W. Zhou, Y. Zhang, T. Luo, Misalignment fusion network for parsing infrared and visible urban scenes, Opt. Lasers Eng. 179 (2024) 108260.
[87] E. Romera, J.M. Álvarez, L.M. Bergasa, R. Arroyo, Erfnet: Efficient residual factorized ConvNet for real-time semantic segmentation, IEEE Trans. Intell. Transp. Syst. 19 (1) (2018) 263–272.
[88] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, W. Liu, CCNet: Criss-cross attention for semantic segmentation, in: 2019 IEEE/CVF International Conference on Computer Vision, ICCV, 2019, pp. 603–612.
[89] J. Liu, J. He, J. Zhang, J.S. Ren, H. Li, Efficientfcn: Holistically-guided decoding for semantic segmentation, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVI 16, Springer, 2020, pp. 1–17.
[90] W. Zhou, X. Lin, J. Lei, L. Yu, J.-N. Hwang, MFFENet: Multiscale feature fusion and enhancement network for RGB–thermal urban road scene parsing, IEEE Trans. Multimed. 24 (2022) 2526–2538.
[91] W. Zhou, S. Dong, J. Lei, L. Yu, MTANet: Multitask-aware network with hierarchical multimodal fusion for RGB-T urban scene understanding, IEEE Trans. Intell. Veh. 8 (1) (2023) 48–58.
[92] C. Xu, Q. Li, X. Jiang, D. Yu, Y. Zhou, Dual-space graph-based interaction network for RGB-thermal semantic segmentation in electric power scene, IEEE Trans. Circuits Syst. Video Technol. 33 (4) (2023) 1577–1592.
[93] J. Zhang, R. Liu, H. Shi, K. Yang, S. Reiß, K. Peng, H. Fu, K. Wang, R. Stiefelhagen, Delivering arbitrary-modal semantic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 1136–1147.
[94] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J.M. Alvarez, P. Luo, SegFormer: Simple and efficient design for semantic segmentation with transformers, Adv. Neural Inf. Process. Syst. 34 (2021) 12077–12090.
[95] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, S. Xie, A convnet for the 2020s, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 11976–11986.
[96] L.-C. Chen, Rethinking atrous convolution for semantic image segmentation, 2017, arXiv preprint arXiv:1706.05587.