SMR-Net: Semantic-Guided Mutually Reinforcing Network for Cross-Modal Image Fusion and Salient Object Detection

Guobao Xiao[1, *, †], **Xinyu Liu**[2, *], **Zebin Lin**[3], **Rui Ming**[3]

[1]School of Computer Science and Technology, Tongji University,
[2]College of Computer and Data Science, Fuzhou University,
[3]School of Computer and Data Science, Minjiang University
x-gb@163.com, lxyklmyt@163.com, zebin_lin@stu.mju.edu.cn, rming@mju.edu.cn

Abstract

This paper introduces a lightweight Semantic-guided Mutually Reinforcing network (SMR-Net) for the tasks of cross-modal image fusion and salient object detection (SOD). The core concept of SMR-Net is to leverage semantics for directing the mutual reinforcing between image fusion and SOD. Specifically, a Progressive Cross-modal Interaction (PCI) image fusion subnetwork is designed to exploit local interactions via convolution operations and extend to global interactions utilizing spatial and channel attention mechanisms. Subsequently, a cross-modal Bit-Plane Slicing-based SOD subnetwork (BPS) is developed by incorporating the fused image as a third modality. This component employs bit-plane slicing and the deformable convolution technique to effectively extract irregular semantic information embedded in fusion features. The refined semantic information then guides the feature extraction process of the source modalities in a reweighted fashion. By cascading these two subnetworks, BPS leverages final semantic results to direct PCI towards focusing more on semantic information. Ultimately, through this semantic-guided mutual enhancement process, SMR-Net excels in both producing high-quality fused images and achieving effective salient object detection. Our extensive experiments on image fusion and SOD tasks convincingly demonstrate the superiority of our network over existing state-of-the-art alternatives without introducing noticeable computational costs. Compared to nearest competitors, our method demonstrates a stronger generalization ability with 26% fewer parameters.

Introduction

Due to theoretical and hardware limitations, neither infrared nor visible images alone can effectively and comprehensively represent real imaging scenes (Li et al. 2023). Conventional visible imaging devices excel at generating richly textured images by recording reflected light from objects. Nevertheless, they struggle to acquire information from critical targets in extreme conditions, such as nighttime, occlusion, fog, and camouflage. In contrast, infrared imaging devices generate images by detecting thermal radiation information emitted by targets, emphasizing significant targets in extreme conditions but potentially neglecting environmental texture details. The complementary characteristics of imaging devices in the two modalities prompt researchers to fuse visible and infrared images, creating an informative composite image through image fusion technology (Xu, Yuan, and Ma 2023). This composite image effectively highlights significant targets while preserving environmental texture details. Furthermore, image fusion technology plays a crucial role in a wide array of high-level visual tasks, including object detection (Xiao et al. 2024), object tracking (Liu et al. 2023b), semantic segmentation (Liu et al. 2023c), and pedestrian re-identification (Qi et al. 2024).

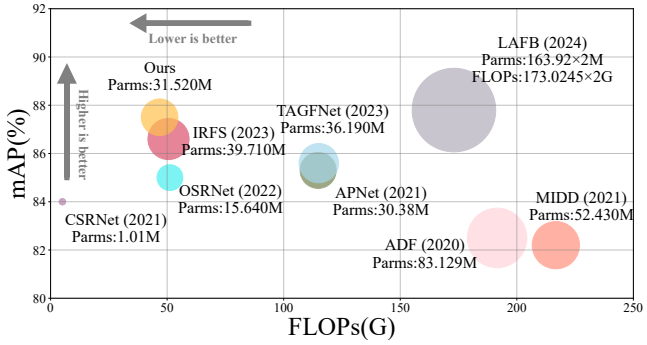


Figure 1: Mean Average Precision (mAP) and Floating Point Operations (FLOPs) of various algorithm on the VT5000 (Tu et al. 2022) dataset.

In the pursuit of advancing the application of image fusion, a notable trend has emerged, with researchers highlighting the requirement to establish connections between image fusion and high-level visual tasks (Sun et al. 2022; Liu et al. 2023a; Hong, Zhang, and Ma 2024). Note that these existing methods primarily concentrate on integrating semantic segmentation or object detection with image fusion, in contrast, this paper focuses on the combination of cross-modal image fusion and salient object detection (SOD). Cross-modal SOD, representing a significant branch of high-level visual tasks, is designed to detect the most visually distinctive objects or regions from different modal images. Thus, it would be highly beneficial to develop a more integrated and optimized approach that effectively combines cross-modal image fusion with SOD tasks.

[*]These authors contributed equally.
[†]Corresponding author.

This paper introduces a lightweight Semantic-guided Mutually Reinforcing network (SMR-Net), to boost the effectiveness of cross-modal image fusion and SOD by leveraging complementary information from both source images. Specifically, a specialized Progressive Cross-modal Interaction based image fusion subnetwork (PCI) is designed to facilitate a stepwise interaction between the two source modalities for image fusion. PCI incorporates two interaction stages: the local interaction stage, utilizing convolution operations for local interactions, and the global interaction stage, leveraging attention mechanisms for broader contextual interactions. In the global interaction stage, the spatial attention and the channel attention are simultaneously exploited to promote interaction between the two modalities across both spatial and channel dimensions. This stepwise interaction process ensures a nuanced exchange of information between the features of two source modalities, effectively preserving complementary details while adeptly filtering out redundant information.

Furthermore, a SOD subnetwork, named Bit-Plane Slicing based SOD subnetwork (BPS), is introduced to utilize the fused image as the third modality. This subnetwork is designed to extract both semantic and high-frequency information of salient objects from fused images through bit-plane slicing, and capture irregular shapes from refined fused features using deformable convolution for enhanced SOD. Within BPS, a semantic and high-frequency extraction mechanism is strategically embedded behind each feature extraction layer to amplify semantic information in the fusion features. The resultant high-semantic fusion features are then assigned appropriate weights to the visible and infrared features through the sigmoid function, thereby aiding in the extraction of semantic information from the two source modalities. It is noteworthy to highlight that our BPS outperforms existing approaches in both performance and efficiency, as evidenced in Fig. 1.

Importantly, PCI serves as the link between these subnetworks by furnishing high-saliency information to steer the SOD task. Simultaneously, BPS constructs a SOD loss to facilitate parameter updates of PCI through back-propagation, prioritizing the emphasis on salient objects. Leveraging these two subnetworks, SMR-NET excels in generating high-quality fused images and accomplishing effective object detection, even in situations involving targets with closely resembling colors.

To be concrete, our contributions are summarized as:

- A lightweight semantic-guided mutually reinforcing network (SMR-Net) is introduced for cross-modal image fusion and salient object detection. SMR-Net consists of image fusion subnetwork (PCI) and SOD network (BPS). PCI furnishes crucial high-saliency information to steer the SOD task, while BPS establishes an SOD loss for updating the parameters of PCI through back-propagation, with a specific focus on enhancing the representation of salient objects.
- A novel cross-modal image fusion network is designed to seamlessly integrate source modalities through local and global stages. This progressive interaction not only preserves intricate details but also effectively filters redundant information. In addition, a cross-modal SOD network is introduced to adeptly extract irregular semantic and high-frequency information through bit-plane slicing and deformable convolution.
- Extensive experiments on publicly available datasets demonstrate the superiority of our SMR-Net over the existing image fusion and SOD algorithms in terms of visual effect and quantitative metrics. Note that our BPS demonstrates a noteworthy 6.25% enhancement in the mean absolute error (MAE) metric when compared to the state-of-the-art ResNet-based SOD methods for the VT5000 dataset, and achieves an impressive 16.67% improvement in the MAE metric for the VT821 dataset (Wang et al. 2018).

It is noteworthy that IRFS (Wang et al. 2023a) and our SMR-Net both employ a cascade structure to facilitate mutual optimisation between the image fusion network and the SOD network. However, they differ significantly: 1) IRFS directly reweights original features with fusion features, potentially introducing irrelevant information. In contrast, our method extracts semantic information from fusion features via bit-plane slicing and deformable convolution, enhancing effectiveness. 2) IRFS relies on attention mechanisms for modality interaction, while our approach solely utilizes implicit semantic information in fused features, improving performance and reducing time overhead. 3) IRFS employs manual fusion strategies in the image fusion sub-network, while our progressive cross-modal interaction module seamlessly integrates source modalities through local and global stages. Thus, our framework provides superior and efficient prediction results.

Related Work

Cross-Modal Image Fusion

Previous studies (Guan et al. 2023; Ma et al. 2022; Xu et al. 2022) concentrate on improving the visual quality and metrics of fused images, yielding promising results. Recently, researchers have made significant strides in the development of multi-task image fusion methods, which can be classified into two categories: semantic-driven methods and semantic feature compensation methods. Semantic-driven methods (Hong, Zhang, and Ma 2024; Liu et al. 2022) construct the guidance function for advanced visual tasks within the fusion network through cascading operations. The semantic feature compensation methods (Liu et al. 2023a; Tang et al. 2023b) offer semantic guidance at the feature level by integrating high-level semantic features into the fused features. In addition, Liu et al. (Liu et al. 2023d) introduced a adaptive adversarial training scheme to strengthen segmentation robustness in adversarial scenes. While existing methods primarily concentrate on the integration of image fusion and semantic segmentation, they allocate limited attention to the combination of image fusion and salient object detection.

Cross-Modal Salient Object Detection

With thermal sensors readily available, thermal and visible salient object detection (RGB-T SOD) has been extensively

studied in recent years (Wang et al. 2018; Gao et al. 2021). RGB-T SOD methods leverage the complementary capabilities of multimodal sensors to generate robust fusion features across different modalities. Tu et al. (Tu et al. 2019) suggested a collaborative graph learning algorithm and introduced VT1000 dataset for RGB-T SOD, comprising 1000 aligned pairs of RGB and thermal images along with their corresponding labels. Zhang et al. (Zhang et al. 2019) proposed to capture semantic information and visual details from RGB-T images at various depths through the fusion of multi-level CNN features. MIDD (Tu et al. 2021) introduces a multi-interactive dual-decoder for both high-level and low-level feature fusion, aimed at generating robust and efficient RGB-T features. Huo et al. (Huo et al. 2021) proposed a context-guided stacking refinement network that progressively refines features from top to bottom by leveraging the interaction between semantic and spatial information.

In contrast, existing RGB-T SOD approaches have traditionally emphasized cross-modal interaction and feature-level fusion, neglecting pixel-level fusion. In real-world scenarios, the fused image can effectively highlight object structures, crucial for distinguishing salient objects. Hence, a logical strategy entails delving into the amalgamation of image fusion and SOD tasks within a unified framework, aiming to harness mutual benefits.

Method

The overall framework presented is depicted in Fig. 2. This network is comprised of a cross-modal image fusion subnetwork (PCI) and a cross-modal salient object detection subnetwork (BPS). These two tasks synergistically reinforce each other through an alternating training method, thereby achieving significantly improved overall performance.

Progressive Cross-Modal Interaction Image Fusion Network

A Progressive Cross-modal Interaction image fusion network (PCI) is proposed to optimize the extraction of complementary information from two source modalities. PCI incorporates two key interaction stages: the local interaction stage, where convolution operations enable localized interactions, and the global interaction stage, which leverages attention mechanisms for more expansive contextual interactions.

We begin by detailing the local interaction stage. Initially, for a given pair of visible image $I_{vi} \in \mathbb{R}^{H\times W\times 3}$ and infrared image $I_{ir} \in \mathbb{R}^{H\times W\times 1}$, two fundamental convolution blocks are applied as a feature extractor F_E. This feature extractor extracts coarse features from the source images:

$$\{F_{vi}, F_{ir}\} = \{F_E(I_{vi}), F_E(I_{ir})\}, \tag{1}$$

Next, we utilize convolution operations to capture complementary cues between the two modalities. The enhanced outputs, denoted as $\hat{F}_{vi}$ and $\hat{F}_{ir}$ through these convolution

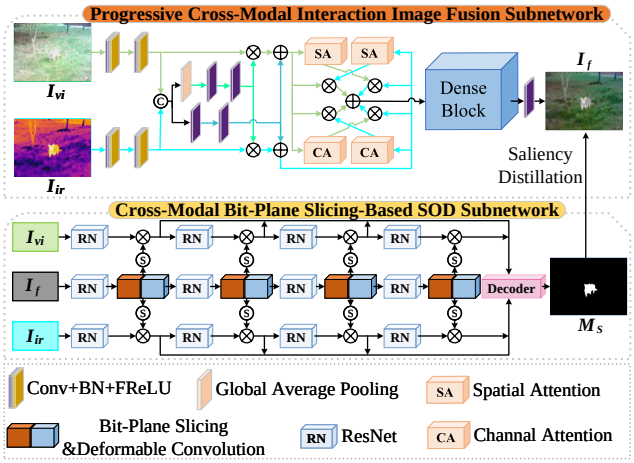


Figure 2: The overall architecture of SMR-Net.

operations, can be expressed as follows:

$$F'_{vi} = F_{vi} \otimes S(Conv(GAP(Cat(F_{vi}, F_{ir})))), \tag{2}$$

$$F'_{ir} = F_{ir} \otimes S(Conv(GAP(Cat(F_{vi}, F_{ir})))), \tag{3}$$

$$\hat{F}_{vi} = F'_{vi} \oplus S(Conv(Cat(F_{vi}, F_{ir}))), \tag{4}$$

$$\hat{F}_{ir} = F'_{ir} \oplus S(Conv(Cat(F_{vi}, F_{ir}))), \tag{5}$$

where $\oplus$ refers to element-wise summation, $Conv$ indicates convolution operation. S and Cat represent the split and concatenate operation, respectively. where $\oplus$ refers to element-wise summation, $Conv$ indicates convolution operation. S and Cat represent the split and concatenate operation, respectively.

Following the local interaction stage, we introduce the global cross-modal interaction stage to alleviate redundancy in both spatial and channel dimensions of the features obtained from the local stage and accentuate essential components. To elaborate, we concurrently feed the two modality features into spatial and channel attention mechanisms, generating spatial and channel weights. Subsequently, these weights are applied to reweight the original features from the other branch through element-wise multiplication. The resulting features are then separately subjected to element-wise addition along both spatial and channel dimensions to derive the final results. The feature filtering process is summarized as:

$$\hat{\Phi}_{vi} = (\hat{F}_{vi} \otimes SA(\hat{F}_{ir})) \oplus (\hat{F}_{vi} \otimes CA(\hat{F}_{ir})), \tag{6}$$

$$\hat{\Phi}_{ir} = (\hat{F}_{ir} \otimes SA(\hat{F}_{vi})) \oplus (\hat{F}_{ir} \otimes CA(\hat{F}_{vi})), \tag{7}$$

where SA and CA respectively represent the spatial attention and the channel attention. This sequential interaction mechanism in PCI is designed to ensure a nuanced exchange of information between the features of the two source modalities.

Then, we obtain the fused map F_{fu} as:

$$F_{fu} = \hat{\Phi}_{vi} \oplus \hat{\Phi}_{ir}. \tag{8}$$

Finally, we adopt the dense block (Huang et al. 2017) and convolution operation to reconstruct the fused image.

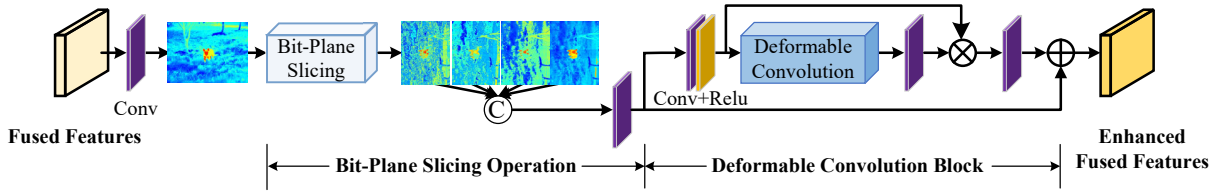


Figure 3: The overall architecture of the proposed bit-plane slicing (BS) and deformable convolution (DC) module.

Cross-Modal Bit-Plane Slicing Based SOD Network

We introduce a novel Cross-modal Bit-Plane Slicing-based SOD network (BPS), integrating bit-plane slicing into the multi-modal SOD task to enhance the exploitation of the potential of fused image as a guide for SOD.

Specifically, BPS incorporates the fusion image as a third modality to guide the multi-modal SOD network, leveraging its ability to recognize salient objects and high contrast. Using a semantic encoder for cross-modal feature extraction, we seamlessly integrate a bit-plane slicing module after each feature extraction layer. This module effectively extracts semantic information from the fusion features, capitalizing on the high contrast differences inherent in these features. The obtained features, enriched with semantic information, play a crucial role in weighting the visible and infrared features. This intricate process is visually depicted in Fig. 3. After each feature extraction layer, the fusion features undergo a convolution operation to condense them into a single channel. Subsequently, the single-channel fusion map is meticulously split into four bit planes through bit-plane slicing. This meticulous bit-slicing procedure can be succinctly summarized as follows:

$$M_{fu}^{i} = Conv(F_{fu}^{i}), \tag{9}$$

where F_{fu}^{i} and M_{fu}^{i} represent the fusion features and fusion map at the i-th scale of the backbone network. For the single-channel fusion map M_{fu}^{i}, we employ Min-Max Normalization to ensure data stability, scaling it to the range of $[0, 1]$. The normalization process can be formulated as follows:

$$M_{norm}^{i} = \frac{M_{fu}^{i} - M_{min}^{i}}{M_{max}^{i} - M_{min}^{i}}. \tag{10}$$

Subsequently, we implement the bit-plane slicing algorithm. The first bit-plane corresponds to the most significant bit, capturing approximately half of M_{norm}^{i}, while the second bit-plane encompasses the remaining bits. This iterative pattern continues until we obtain four distinct bit-planes. This bit-slicing procedure can be summarized as follows:

$$bit_{l}^{i} = \left\lfloor \frac{M_{norm}^{i} \mod 2^{-l+1}}{2^{-l}} \right\rfloor, \quad l = 1, 2, 3, 4 \tag{11}$$

where bit_{l}^{i} indicates the l-th bit-plane. In this paper, we focus specifically on 4-bit planes.

Furthermore, recognizing that each bit-plane contains different levels of semantic information, we concatenate the four obtained bit-planes. These concatenated bit-planes are then subjected to a convolution operation, serving to extract features and restore the original channel count:

$$\hat{F}_{bs}^{i} = Conv(Cat(bit_{1}^{i}, bit_{2}^{i}, bit_{3}^{i}, bit_{4}^{i})). \tag{12}$$

In addition, by introducing deformable convolution (Dai et al. 2017), we augment the spatial awareness of the original convolution operation, empowering it to effectively capture irregular shapes. To specifically target the extraction of irregular salient objects, the deformable convolution is seamlessly integrated after the bit-plane slicing process. As depicted in Fig. 3, when presented with the input $\hat{F}_{bs}^{i}$, the resulting output is denoted as $\hat{F}_{out}^{i}$. This deformable convolution operation is succinctly formulated as follows:

$$\hat{F}_{enh}^{i} = Conv(ReLu(\hat{F}_{bs}^{i})), \tag{13}$$

$$\hat{F}_{dc}^{i} = (Conv(DC(\hat{F}_{enh}^{i})))), \tag{14}$$

$$\hat{F}_{out}^{i} = Conv(\hat{F}_{dc}^{i} \otimes \hat{F}_{enh}^{i}) \oplus \hat{F}_{bs}^{i}, \tag{15}$$

where $\hat{F}_{enh}^{i}$ signifies the enhanced features, and $\hat{F}_{dc}^{i}$ represents the features obtained through deformable convolution operations, with DC denoting deformable convolution.

Following this, the fusion features, enriched with semantic-salient information, are employed to weight the pixels of visible and thermal feature maps using a sigmoid function. Through a continuous iteration of feature extraction and enhancement operations, we acquire refined features at various scales. To effectively reconstruct and synthesize the final salient object map from visible and thermal features, we employ MSGD (Wang et al. 2023a) as our decoder module.

Loss Function

For the fusion task, we introduce the intensity loss $\mathcal{L}_{int} = \frac{1}{HW}\left\|I_f - max(I_{ir}, I_{vi})\right\|_1$ to retain the prominent saliency and intensity distribution in the fused images, while the gradient loss $\mathcal{L}_{texture} = \frac{1}{HW}\left\||\nabla I_f| - max(|\nabla I_{ir}|, |\nabla I_{vi}|)\right\|_1$ is employed to preserve the texture details, where H and W are the height and width of an image, respectively, $\|\cdot\|_1$ denotes the $l_1 - norm$ and $max(\cdot)$ stands for the element-wise maximum selection. The total loss is defined as $\mathcal{L}_{fusion} = \mathcal{L}_{int} + \alpha\mathcal{L}_{texture} + \mathcal{L}_{sod}$, where α are hyperparameters employed to balance the weights of the loss values. In our paper, we set $\alpha = 1$. Notice that we extraordinarily introduce $\mathcal{L}_{SOD}$ to force the fused image to stand out the saliency-related information.

Figure 4: Qualitative comparison of our method with 9 state-of-the-art image fusion methods on typical 3 image pairs from the TNO, RoadScene and MSRS datasets.

For the SOD task, we utilize the weighted binary cross-entropy (wBCE) loss and weighted IoU (wIoU) as our SOD loss functions. The corresponding loss function is defined as $\mathcal{L}_{sod} = \mathcal{L}^{\omega}_{bce}(M, GT) + \mathcal{L}^{\omega}_{iou}(M, GT)$, where M represents the predicted saliency map, and GT indicates the ground truth.

Experiments

Experimental Configurations

Datasets To accommodate the joint framework of image fusion and salient object detection tasks, we select the VT5000 (Tu et al. 2022) training set for training our network. In addition, we also individually evaluate the generalization capabilities of our SMR-Net in the tasks of image fusion and SOD.

In the image fusion sub-task, we verify the generalization ability of our PCI subnetwork on the TNO (Toet 2017) (15 pairs), RoadScene (Xu et al. 2020) (40 pairs) and MSRS (Tang et al. 2022) (80 pairs) datasets, comparing it against the the state-of-the-art image fusion algorithm as follow: FuisonGAN (Ma et al. 2019), PMGI (Zhang et al. 2020), GANMcC (Ma et al. 2020), SDNet (Zhang and Ma 2021), STDFusionNet (Ma et al. 2021), CUFD (Xu et al. 2022), TarDAL (Liu et al. 2022), DIVFusion (Tang et al. 2023a) and IRFS (Wang et al. 2023a).

In the SOD sub-task, we verify the generalization ability of our BPS subnetwork on the VT5000 test set ($2,500$ pairs), VT1000 (Tu et al. 2019) ($1,000$ pairs), and VT821 (Wang et al. 2018) (821 pairs) datasets, comparing it with the state-of-the-art SOD methods, including MTMR (Wang et al. 2018), SGDL (Tu et al. 2019), ADF (Zhang et al. 2019), MIDD (Tu et al. 2021), APNet (Zhou et al. 2021), CSRNet (Huo et al. 2021), OSRNet (Huo et al. 2022), TAGFNet (Wang et al. 2023b), IRFS (Wang et al. 2023a) and LAFB (Wang et al. 2024).

Evaluation Metrics In the image fusion sub-task, we select five common evaluation metrics to quantify the evaluation, including peak signal-to-noise ratio (PSNR), mean squared error (MSE), visual information fidelity (VIF) and quality with no reference based on the fusion of similarity and information (Qabf). In the SOD sub-task, we also adopt common metrics to evaluate the performance of saliency map, including S-measure (Tu et al. 2019), F-measure (Achanta et al. 2009), E-measure (Fan et al. 2018) and MAE (Perazzi et al. 2012).

Implementation Details We alternately train our image fusion subnetwork (PCI) and SOD subnetwork (BPS), setting the total number of training epochs to 10 and the batch size to 4. Both PCI and BPS use the Adam optimizer to update the training parameters. Additionally, the siamese encoder of BPS relies on the pre-trained ResNet-34 (He et al. 2016) backbone. It is worth mentioning that, for data in the RGB color space, we convert it to the YCbCr color space to process color information. All experiments are conducted on the NVIDIA GeForce RTX 3090 GPU with 24GB memory and 3.20 GHz Intel(R).

Cross-Modal Image Fusion

The visual results in the TNO, RoadScene and MSRS datasets are displayed in Fig. 4. To visualize the superior comparison of the fusion images, we zoom in on an area with salient object in the red box and an area with rich texture details in the green box. As can be seen in the first row that FusionGAN, GANMcC, PMGI, SDNet, CUFD and TarDAL fail to reflect the texture details of the branch from the visible image. DIVFusion and IRFS introduce higher artifacts. From the second row, we can see that FusionGAN, GANMcC, STDFusionNet, CUFD, DIVFusion and IRFS all badly lose the texture of the car logo due to the effect of exposure from the visible image. From the third row, it can be noted that STDFusion, CUFD, DIVFusion and IRFS are all contaminated by the exposure from visible image, and present unrealistic scenarios. Only our method successfully highlights salient objects from infrared images as well as preserves texture details from visible images. Further, we present quantitative results in Table 1. In the TNO dataset, our method achieves the best performance on PSNR, MSE and Qabf metrics, and the second best performance on VIF metrics, demonstrating that our fusion results closely resemble the source images and contain more saliency information. In the RoadScene dataset, PCI gets the best performance on VIF and Qabf metrics, and second-best performance on PSNR and MSE metrics, indicating that our method retains more saliency information and edge details from the source images. In the MSRS dataset, our method obtains the best performance on all metrics, implying that our fused images retain high contrast and texture details

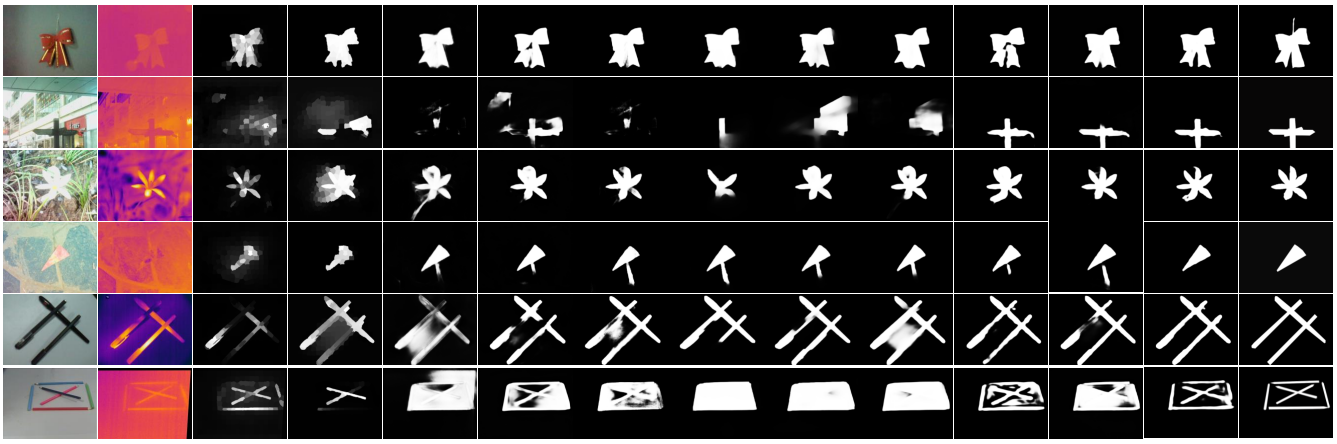


Figure 5: Qualitative comparison of our method with 10 state-of-the-art RGB-T SOD methods on typical 6 image pairs from the VT5000, VT1000 and VT821 datasets.

Datasets	Metric	FusionGAN	PMGI	GANMcC	SDNet	CUFD	STDFusionNet	TarDAL	DIVFusion	IRFS	PCI
TNO	**PSNR**↑	62.332	63.779	63.696	65.425	64.603	63.597	60.434	61.379	66.149	66.728
	MSE ↓	0.051	0.033	0.038	0.024	0.028	0.037	0.067	0.059	0.020	0.020
	VIF ↑	0.403	0.654	0.534	0.604	0.641	0.838	0.394	0.627	0.632	0.743
	Qabf ↑	0.221	0.413	0.298	0.418	0.399	0.473	0.181	0.326	0.390	0.547
RoadScene	**PSNR**↑	61.572	63.567	62.849	65.382	64.492	60.839	60.785	63.997	67.093	65.953
	MSE ↓	0.047	0.032	0.036	0.021	0.025	0.055	0.056	0.028	0.014	0.018
	VIF ↑	0.375	0.631	0.503	0.639	0.627	0.654	0.456	0.582	0.610	0.662
	Qabf ↑	0.247	0.485	0.358	0.541	0.470	0.382	0.264	0.318	0.440	0.583
MSRS	**PSNR**↑	66.484	59.645	67.109	66.612	65.025	66.417	65.173	55.970	66.512	67.730
	MSE ↓	0.016	0.076	0.015	0.015	0.022	0.016	0.024	0.169	0.016	0.012
	VIF ↑	0.477	0.767	0.707	0.529	0.587	0.529	0.141	0.821	0.797	0.837
	Qabf ↑	0.163	0.384	0.373	0.366	0.401	0.423	0.116	0.273	0.467	0.583

Table 1: PSNR, MSE, VIF and Qabf metrics comparisons with different image fusion models on TNO, RoadScene and MSRS datasets (unit: RED indicates the best result and BLUE represents the second best result).

while providing good visual effects.

Cross-Modal Salient Object Detection

The qualitative comparisons on the VT5000, VT1000 and VT821 datasets are shown in Fig. 5. As we can see, our method achieves better detection performance than other methods in some challenging cases: complex object (1-th and 3-th rows), complex scene (2-th row), small object (4-th row) and multiple objects (5-th and 6-th rows). The excellent visual examples indicate that our approach can better locate salient objects and produce more accurate saliency maps. Additionally, the quantitative results on the VT5000, VT1000 and VT821 datasets are exhibited in Table 1. As can be clearly found, our BPS achieves the best performance on all metrics in the VT5000 and VT821 datasets. In VT5000, our method performs 44.4% better in the F-measure and has 73.7% lower MAE than MTMR. In VT821, our approach outperforms the advanced ADF, with improvements of 17.7% and 61.0% on F-measure and MAE metrics, respectively. In VT1000, our algorithm merely performs 0.2% lower than the better on the S-measure and F-measure metrics and 5.2% higher than the better on the MAE metric.

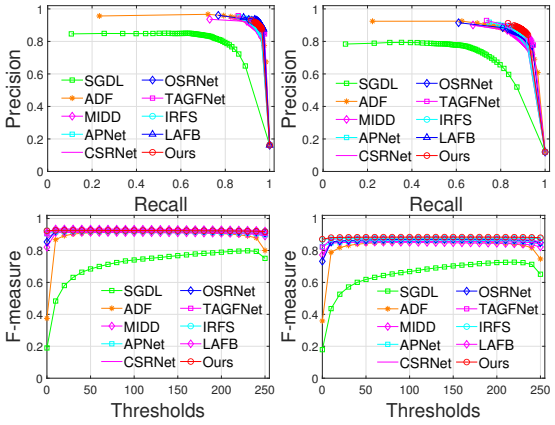
Further, we introduce Precision-Recall (PR) and F-measure curves to present a comprehensive comparison of the model performance, as shown in Fig. 6. From these curves, we can observe that our algorithm consistently outperforms all other models under different thresholds, indicating that the proposed method is more robust than other models.

Ablation Study

BS and DC blocks To verify the necessity of BS block and DC block in the BPS subnetwork, we successively remove BS and DC blocks, and the visual comparison are displayed in Fig. 7. Clearly, the version W/O BS loses abundant edge and texture details of the salient object. Meanwhile, the version W/O DC retains too much meaningless information. Only our fully model has higher sensitivity to salient object detection. Similarly, the quantitative analysis are presented in Table 3. Obviously, using BS block brings an improvement of 20.1% in the F-measure metric, and using the DC block brings an improvement of 54.5% in MAE score on the VT5000 dataset. When combining the two blocks, the performance of the four metrics gets the best, which brings an

Datasets	Metric	MTMR	SGDL	ADF	MIDD	APNet	CSRNet	OSRNet	TAGFNet	IRFS	LAFB	BPS
VT821	$S_\alpha \uparrow$	.725	.764	.810	.871	.867	.885	.875	.880	.863	.884	.888
	$F_\beta \uparrow$	.662	.731	.717	.805	.816	.831	.814	.822	.813	.843	.844
	$E_\epsilon \uparrow$	.815	.846	.843	.895	.907	.909	.896	.905	.901	.915	.920
	MAE$\downarrow$	.108	.085	.077	.045	.034	.038	.043	.035	.036	.034	.030
VT1000	$S_\alpha \uparrow$	.706	.787	.910	.907	.921	.918	.926	.926	.924	.932	.924
	$F_\beta \uparrow$	.715	.764	.847	.871	.883	.877	.892	.890	.901	.905	.899
	$E_\epsilon \uparrow$	.836	.856	.921	.928	.938	.925	.935	.935	.943	.945	.945
	MAE$\downarrow$	.119	.090	.034	.029	.021	.024	.022	.021	.019	.018	.020
VT5000	$S_\alpha \uparrow$	.680	.750	.864	.856	.876	.868	.875	.884	.883	.893	.891
	$F_\beta \uparrow$	.595	.672	.778	.789	.820	.811	.824	.827	.851	.857	.859
	$E_\epsilon \uparrow$	.795	.824	.891	.891	.938	.905	.908	.913	.928	.931	.935
	MAE$\downarrow$	.114	.089	.048	.046	.035	.042	.040	.036	.032	.030	.030

Table 2: S-measure, adaptive F-measure, adaptive E-measure and MAE metrics comparisons with 10 state-of-the-art RGB-T SOD models on VT5000, VT1000 and VT821 datasets.



(a) VT1000 (b) VT5000

Figure 6: Comparison of PR curves (Top) and F-measure curves (Down) across various RGB-T SOD methods.

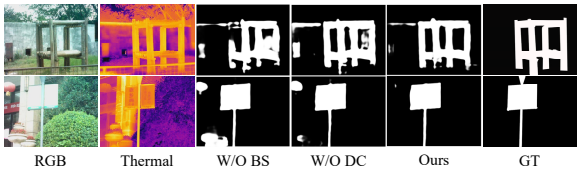


Figure 7: Visual comparison for the proposed BS and DC blocks on VT5000 dataset

improvement of 21.3% in the F-measure metric. This substantial increase further reinforces the necessity of employing two blocks.

Analysis of BS with Different Layers The proposed BS module utilizes bit-plane slicing to extract high-frequency and semantic information from fused images. This section concentrates on the selection of layers for bit-plane. As shown in Table 4, Bit4-only performs best on all metrics, followed by the case where all four layers are utilized.

#	BS	DC	VT5000 $S_\alpha \uparrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	$\mathcal{MAE} \downarrow$
0			0.810	0.708	0.833	0.066
1	✓		0.888	0.848	0.931	0.031
2		✓	0.886	0.847	0.928	0.032
3	✓	✓	0.891	0.859	0.935	0.030

Table 3: Ablation studies of with and without the proposed BS and DC modules.

#	Bit1	Bit2	Bit3	Bit4	VT5000 $S_\alpha \uparrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	$\mathcal{MAE} \downarrow$
0					0.886	0.847	0.928	0.032
1	✓				0.888	0.851	0.932	0.030
2		✓			0.889	0.855	0.932	0.030
3			✓		0.888	0.851	0.929	0.031
4				✓	0.891	0.859	0.935	0.030
5	✓	✓			0.888	0.853	0.933	0.030
6	✓	✓	✓		0.889	0.855	0.933	0.030
7	✓	✓	✓	✓	0.889	0.857	0.933	0.030

Table 4: Ablation studies of the selection of different bit-plane. The bit4-only configuration achieves the best performance, indicating higher semantic consistency.

Conclusion

The paper presents SMR-Net, a lightweight network for cross-modal image fusion and SOD tasks, consisting of the fusion subnetwork (PCI) and the SOD subnetwork (BPS). The PCI enhances cross-modal interactions using convolution and attention mechanisms, while the BPS employs bit-plane slicing and deformable convolution for saliency extraction. The interconnected subnetworks allow PCI to guide SOD and BPS to refine PCI through back-propagation, focusing on salient objects. SMR-Net excels in generating high-quality fused images and effectively detecting objects without significant computational costs, as demonstrated by extensive experiments that show its superiority over existing algorithms in both qualitative and quantitative metrics.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grants 62472312 and 32201679.

References

Achanta, R.; Hemami, S.; Estrada, F.; and Susstrunk, S. 2009. Frequency-tuned salient region detection. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 1597–1604.

Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; and Wei, Y. 2017. Deformable Convolutional Networks. In *Proc. IEEE Int. Conf. Comput. Vis.*

Fan, D.-P.; Gong, C.; Cao, Y.; Ren, B.; Cheng, M.-M.; and Borji, A. 2018. Enhanced-alignment measure for binary foreground map evaluation. In *Proc. Int. Jt. Conf. Artif. Intell*, 698–704.

Gao, W.; Liao, G.; Ma, S.; Li, G.; Liang, Y.; and Lin, W. 2021. Unified information fusion network for multi-modal RGB-D and RGB-T salient object detection. *IEEE Trans. Circuits Syst. Video Technol.*, 32(4): 2091–2106.

Guan, Y.; Xu, R.; Yao, M.; Wang, L.; and Xiong, Z. 2023. Mutual-guided dynamic network for image fusion. In *Proc. 31st ACM Int. Conf. Multimedia*, 1779–1788.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 770–778.

Hong, W.; Zhang, H.; and Ma, J. 2024. MERF: A Practical HDR-Like Image Generator via Mutual-Guided Learning Between Multi-Exposure Registration and Fusion. *IEEE Trans. Image Process.*, 33: 2361–2376.

Huang, G.; Liu, Z.; Van Der Maaten, L.; and Weinberger, K. Q. 2017. Densely Connected Convolutional Networks. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2261–2269.

Huo, F.; Zhu, X.; Zhang, L.; Liu, Q.; and Shu, Y. 2021. Efficient context-guided stacked refinement network for RGB-T salient object detection. *IEEE Trans. Circuits Syst. Video Technol.*, 32(5): 3111–3124.

Huo, F.; Zhu, X.; Zhang, Q.; Liu, Z.; and Yu, W. 2022. Real-time one-stream semantic-guided refinement network for RGB-thermal salient object detection. *IEEE Trans. Instrum. Meas.*, 71: 1–12.

Li, J.; Chen, J.; Liu, J.; and Ma, H. 2023. Learning a graph neural network with cross modality interaction for image fusion. In *Proc. 31st ACM Int. Conf. Multimedia*, 4471–4479.

Liu, J.; Fan, X.; Huang, Z.; Wu, G.; Liu, R.; Zhong, W.; and Luo, Z. 2022. Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 5802–5811.

Liu, J.; Liu, Z.; Wu, G.; Ma, L.; Liu, R.; Zhong, W.; Luo, Z.; and Fan, X. 2023a. Multi-interactive feature learning and a full-time multi-modality benchmark for image fusion and segmentation. In *Proc. IEEE Int. Conf. Comput. Vis.*, 8115–8124.

Liu, L.; Li, C.; Xiao, Y.; and Tang, J. 2023b. Quality-aware rgbt tracking via supervised reliability learning and weighted residual guidance. In *Proc. 31st ACM Int. Conf. Multimedia*, 3129–3137.

Liu, Z.; Liu, J.; Zhang, B.; Ma, L.; Fan, X.; and Liu, R. 2023c. PAIF: Perception-aware infrared-visible image fusion for attack-tolerant semantic segmentation. In *Proc. 31st ACM Int. Conf. Multimedia*, 3706–3714.

Liu, Z.; Liu, J.; Zhang, B.; Ma, L.; Fan, X.; and Liu, R. 2023d. PAIF: Perception-aware infrared-visible image fusion for attack-tolerant semantic segmentation. In *Proc. 31st ACM Int. Conf. Multimedia*, 3706–3714.

Ma, J.; Tang, L.; Fan, F.; Huang, J.; Mei, X.; and Ma, Y. 2022. SwinFusion: Cross-domain long-range learning for general image fusion via swin transformer. *IEEE/CAA J. Autom. Sinica*, 9(7): 1200–1217.

Ma, J.; Tang, L.; Xu, M.; Zhang, H.; and Xiao, G. 2021. STDFusionNet: An infrared and visible image fusion network based on salient target detection. *IEEE Trans. Instrum. Meas.*, 70: 1–13.

Ma, J.; Yu, W.; Liang, P.; Li, C.; and Jiang, J. 2019. FusionGAN: A generative adversarial network for infrared and visible image fusion. *Inf. Fusion*, 48: 11–26.

Ma, J.; Zhang, H.; Shao, Z.; Liang, P.; and Xu, H. 2020. GANMcC: A generative adversarial network with multiclassification constraints for infrared and visible image fusion. *IEEE Trans. Instrum. Meas.*, 70: 1–14.

Perazzi, F.; Krähenbühl, P.; Pritch, Y.; and Hornung, A. 2012. Saliency filters: Contrast based filtering for salient region detection. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 733–740.

Qi, J.; Liang, T.; Liu, W.; Li, Y.; and Jin, Y. 2024. A generative-based image fusion strategy for visible-infrared person re-identification. *IEEE Trans. Circuits Syst. Video Technol.*, 34(1): 518–533.

Sun, Y.; Cao, B.; Zhu, P.; and Hu, Q. 2022. Detfusion: A detection-driven infrared and visible image fusion network. In *Proc. 30st ACM Int. Conf. Multimedia*, 4003–4011.

Tang, L.; Xiang, X.; Zhang, H.; Gong, M.; and Ma, J. 2023a. DIVFusion: Darkness-free infrared and visible image fusion. *Inf. Fusion*, 91: 477–493.

Tang, L.; Yuan, J.; Zhang, H.; Jiang, X.; and Ma, J. 2022. PIAFusion: A progressive infrared and visible image fusion network based on illumination aware. *Inf. Fusion*, 83: 79–92.

Tang, L.; Zhang, H.; Xu, H.; and Ma, J. 2023b. Rethinking the necessity of image fusion in high-level vision tasks: A practical infrared and visible image fusion network based on progressive semantic injection and scene fidelity. *Inf. Fusion*, 101870.

Toet, A. 2017. The TNO multiband image data collection. *Data in brief*, 15: 249–251.

Tu, Z.; Li, Z.; Li, C.; Lang, Y.; and Tang, J. 2021. Multi-interactive dual-decoder for RGB-thermal salient object detection. *IEEE Trans. Image Process.*, 30: 5678–5691.

Tu, Z.; Ma, Y.; Li, Z.; Li, C.; Xu, J.; and Liu, Y. 2022. RGBT salient object detection: A large-scale dataset and benchmark. *IEEE Trans. Multimedia*, 25: 4163 – 4176.

Tu, Z.; Xia, T.; Li, C.; Wang, X.; Ma, Y.; and Tang, J. 2019. RGB-T image saliency detection via collaborative graph learning. *IEEE Trans. Multimedia*, 22: 160–173.

Wang, D.; Liu, J.; Liu, R.; and Fan, X. 2023a. An interactively reinforced paradigm for joint infrared-visible image fusion and saliency object detection. *Inf. Fusion*, 98: 101828.

Wang, G.; Li, C.; Ma, Y.; Zheng, A.; Tang, J.; and Luo, B. 2018. RGB-T saliency detection benchmark: Dataset, baselines, analysis and a novel approach. In *IGTA 2018*, 359–369. Springer.

Wang, H.; Song, K.; Huang, L.; Wen, H.; and Yan, Y. 2023b. Thermal images-aware guided early fusion network for cross-illumination RGB-T salient object detection. *Eng. Appl. Artif. Intell.*, 118: 105640.

Wang, K.; Tu, Z.; Li, C.; Zhang, C.; and Luo, B. 2024. Learning Adaptive Fusion Bank for Multi-modal Salient Object Detection. *IEEE Trans. Circuits Syst. Video Technol.*

Xiao, G.; Tang, Z.; Guo, H.; Yu, J.; and Shen, H. T. 2024. FAFusion: Learning for Infrared and Visible Image Fusion via Frequency Awareness. *IEEE Trans. Instrum. Meas.*, 73: 1–11.

Xu, H.; Gong, M.; Tian, X.; Huang, J.; and Ma, J. 2022. CUFD: An encoder–decoder network for visible and infrared image fusion based on common and unique feature decomposition. *Comput. Vis. Image Underst.*, 218: 103407.

Xu, H.; Ma, J.; Jiang, J.; Guo, X.; and Ling, H. 2020. U2Fusion: A unified unsupervised image fusion network. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(1): 502–518.

Xu, H.; Yuan, J.; and Ma, J. 2023. MURF: Mutually Reinforcing Multi-Modal Image Registration and Fusion. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(10): 12148–12166.

Zhang, H.; and Ma, J. 2021. SDNet: A versatile squeeze-and-decomposition network for real-time image fusion. *Int. J. Comput. Vis.*, 129: 2761–2785.

Zhang, H.; Xu, H.; Xiao, Y.; Guo, X.; and Ma, J. 2020. Rethinking the image fusion: A fast unified image fusion network based on proportional maintenance of gradient and intensity. In *Proc. AAAI Conf. Artif. Intell.*, volume 34, 12797–12804.

Zhang, Q.; Huang, N.; Yao, L.; Zhang, D.; Shan, C.; and Han, J. 2019. RGB-T salient object detection via fusing multi-level CNN features. *IEEE Trans. Image Process.*, 29: 3321–3335.

Zhou, W.; Zhu, Y.; Lei, J.; Wan, J.; and Yu, L. 2021. 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): 957–968.