MRFS: Mutually Reinforcing Image Fusion and Segmentation

Hao Zhang[1], Xuhui Zuo[1], Jie Jiang[2], Chunchao Guo[2], Jiayi Ma[1]*
[1]Electronic Information School, Wuhan University, Wuhan, China
[2]Data Platform Department, Tencent, Shenzhen, China
{zhpersonalbox,jyma2010}@gmail.com, xuhuizuo2001@whu.edu.cn, {zeus, chunchaoguo}@tencent.com

Abstract

This paper proposes a coupled learning framework to break the performance bottleneck of infrared-visible image fusion and segmentation, called MRFS. By leveraging the intrinsic consistency between vision and semantics, it emphasizes mutual reinforcement rather than treating these tasks as separate issues. First, we embed weakened information recovery and salient information integration into the image fusion task, employing the CNN-based interactive gated mixed attention (IGM-Att) module to extract high-quality visual features. This aims to satisfy human visual perception, producing fused images with rich textures, high contrast, and vivid colors. Second, a transformer-based progressive cycle attention (PC-Att) module is developed to enhance semantic segmentation. It establishes single-modal self-reinforcement and cross-modal mutual complementarity, enabling more accurate decisions in machine semantic perception. Then, the cascade of IGM-Att and PC-Att couples image fusion and semantic segmentation tasks, implicitly bringing vision-related and semantics-related features into closer alignment. Therefore, they mutually provide learning priors to each other, resulting in visually satisfying fused images and more accurate segmentation decisions. Extensive experiments on public datasets showcase the advantages of our method in terms of visual satisfaction and decision accuracy. The code is publicly available at https://github.com/HaoZhang1018/MRFS.

1. Introduction

Due to limitations in imaging principles, neither infrared nor visible images alone can provide a complete depiction of the imaging scene [17, 38, 42]. In this context, infrared and visible modality fusion (IVMF) combines their specific advantages, aiding both humans and machines in better understanding scenes [5, 10, 19, 37]. Owing to such practical benefits, IVMF has found applications in various fields, *e.g.*, security monitoring, and night-assisted driving [16, 22].

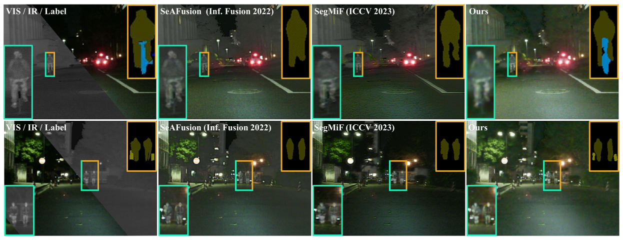


Figure 1. Comparison with the state-of-the-art SeAFusion [27] and SegMiF [14] that combine image fusion and semantic segmentation. Our method presents more salient visual objects and more accurate segmentations.

From the intended use, IVMF can be categorized into **image fusion** and **semantic fusion**. The former aims to generate high-quality visualizations consistent with human visual perception, typically featuring rich textures, significant contrast, and vivid colors [13, 23, 36]. The latter focuses on achieving more advantageous scene feature expression, enabling machines to make semantic-level decisions and descriptions for scenes, *e.g.*, semantic segmentation and object detection [12, 49]. *This paper specifically concentrates on semantic segmentation [9, 45] as a representative of semantic fusion.*

In recent decades, deep learning advancements have propelled significant progress in IVMF. In image fusion, where ground truth is lacking, research emphasizes designing unsupervised loss functions to preserve information. A common approach involves similarity losses based on various representation variables between fused and source images, aiming to extract desired image characteristics aligned with human visual perception [18, 32, 35]. For example, PMGI [41] maintains intensity and gradient proportionally, favoring high-contrast, high-frequency texture features. Differently, in semantic fusion, where labeled data is available, researchers focus on innovative network structures to ensure effective interaction between features for mutual complementation and integration [2, 31, 43, 51]. For instance, LASNet [9] integrates multi-modal features with

*Corresponding author

collaborative location, complementary activation, and edge sharpening modules, facilitating expressive feature extraction crucial for semantic decision-making.

Current methods show promising results in visual outcomes and segmentation, yet challenges persist. Firstly, the *feature neutralization drawback* and *low-light information loss* hinder image fusion. The former refers to the attenuation of thermal contrast and textures caused by unsupervised loss with a fixed proportion [18, 28, 41], while the latter denotes the erroneous discard of details in low-light visible images during fusion due to low intensity [35, 46]. Secondly, in the implementation of feature interaction, current IVMF modules neglect the problem of misjudging the importance of pooled features. Additionally, there is a paucity of discussion on the inherent logical transmission relationship between single-modal self-reinforcement and cross-modal mutual complementarity. These oversights lead to issues like *feature validity score mismatch* [3, 21] and *insufficient feature aggregation* [39, 51], constraining image fusion and segmentation accuracy. Lastly, most methods treat image fusion and semantic segmentation as isolated problems [44, 45], overlooking their intrinsic mutually beneficial relationship. Although SeAFusion [27] and SegMiF [14] seek improvement by incorporating image fusion as a precursor task to semantic segmentation, such a sequential connection may establish interdependencies and impose performance limitations on both tasks, as depicted in Fig. 1.

To address these challenges, we propose a coupled learning framework to enhance both image fusion and semantic segmentation performance. This framework leverages the intrinsic consistency between vision and semantics, making the tasks mutually reinforcing rather than independent. Firstly, we integrate salient information integration and weakened information recovery into image fusion to address the feature neutralization drawback and low-light information loss. On the one hand, instead of the unsupervised loss with a fixed proportion, we dynamically select salient contrast and textures as optimization goals for consistent preservation. On the other hand, we employ a series of data augmentation strategies to construct positive samples and incorporate them into fusion guidance, adaptively enabling low-intensity texture recovery and thermal object saliency enhancement. Secondly, we introduce a CNN-based interactive gated mix attention (IGM-Att) module for visual completion and a transformer-based progressive cycle attention (PC-Att) module for semantic completion. These modules correct mismatched feature validity scores and alleviate insufficient feature aggregation. Specifically, IGM-Att integrates a trainable gating mechanism to identify misjudged positional features and corrects mismatched high or low scores, thereby refining the conventional pooling-based attention. Meanwhile, PC-Att achieves single-modal self-reinforcement and cross-modal mutual complementarity, forming a closed loop for effective fusion between infrared and visible features. Finally, we unite image fusion and semantic segmentation by using IGM-Att and PC-Att as intermediaries. Vision-related features aligned with human perception in image fusion (*e.g.*, restored details, significant thermal contrast) benefit semantic segmentation. Conversely, semantics-related features in segmentation contribute favorable appearance characteristics to the fused image (*e.g.*, outlines of objects). Collaborative optimization of these aspects is anticipated to yield higher returns than individual optimizations.

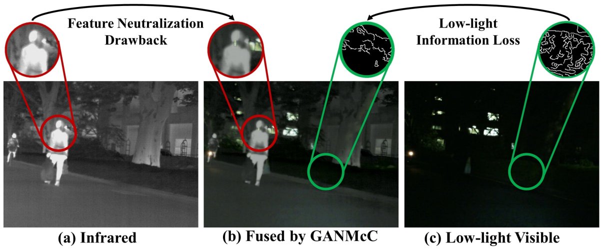


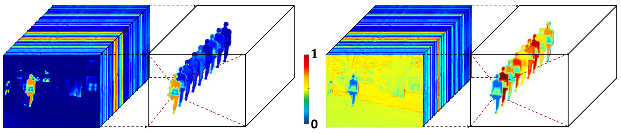
Figure 2. Illustration of the feature neutralization drawback and the low-light information loss. GANMcC [18] is a classic deep image fusion method using unsupervised loss with a fixed proportion. The edge contours are calculated using the Canny operator [4].

In summary, our contributions are as follows:

- We enhance image fusion with the capabilities of salient information integration and weakened information recovery, effectively mitigating the feature neutralization drawback and low-light information loss. Consequently, significant contrast and rich textures will be effectively transferred to the fused image.
- We design a CNN-based interactive gated mix attention module for visual completion, and a transformer-based progressive cycle attention module for semantic completion. They respectively solve the problem of mismatched feature validity scores and strengthen the sufficiency of feature aggregation.
- The strategic coupling of image fusion and semantic segmentation establishes a mutually reinforcing relationship, leading to a dual improvement in their performance.

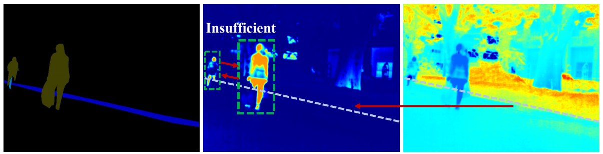
2. Background and Motivations

Feature Neutralization Drawback and Low-light Information Loss. Infrared and visible images often exhibit significant differences in intensity and gradient distribution. In such cases, traditional unsupervised loss with a fixed proportion can diminish valuable characteristics, such as contrast, illustrated in Fig. 2. A potential solution is to dynamically prioritize superior features from source images during optimization, ensuring the preservation of significant contrast and textures. Besides, under poor lighting conditions, low-intensity visible images tend to obscure valu-



(a) Positive features surrounded by negative ones (b) Negative features surrounded by positive ones

Figure 3. Illustration of the feature validity score mismatch. Red denotes positive features while blue indicates negative ones. Average pooling suppresses positive features when they are surrounded by negative ones. Average and max pooling both overestimate negative features when they are surrounded by positive ones.



(a) Segmentation Label (b) Feature from Infrared Image (c) Feature from Visible Image

Figure 4. Illustration of insufficient feature aggregation. Objects of the same class within the single modality can act as cues for mutually reinforcing, achieving single-modal self-reinforcement. Different-modal images can offer information that the other modality lacks, fulfilling cross-modal mutual complementarity.

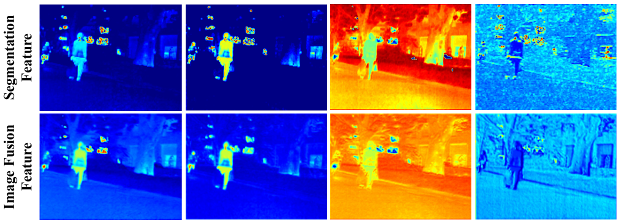


Figure 5. Features used for segmentation and those used for image fusion show a strong correlation.

able details. During fusion, these details are often overlooked and discarded, leading to the loss of crucial information, as shown in Fig. 2. Digging out more details from low-intensity regions would effectively improve the overall quality of the fused image.

Feature Validity Score Mismatch. In conventional pooling-based attention, the score of specific features is typically determined using statistics from all features in the corresponding channel or spatial location [3, 21]. Mainstream statistical operations like average pooling and max pooling [7, 30, 33] compute the average and maximum response of features in specific dimensions. However, average pooling may misassign a low score to a positive feature surrounded by negative ones in the same spatial location or channel, as shown in Fig. 3 (a). Similarly, both average and max pooling may incorrectly assign a high score to a negative feature surrounded by positive ones, as depicted in Fig. 3 (b). Introducing a correction mechanism to redistribute scores by evaluating each feature's impact on others would be highly desirable.

Insufficient Feature Aggregation. During feature fusion, many methods use cross-attention for mutual query and embedding, enhancing information complementarity [39, 51]. However, relying solely on cross-modal complementarity is insufficient, as it doesn't fully exploit its own modality information, as shown in Fig. 4. Introducing self-attention during feature extraction partially addresses this issue. Yet, it is not specifically tailored for feature fusion, and therefore lacks optimal compatibility with cross-attention. Hence, considering both single-modal self-reinforcement and cross-modal complementarity in feature fusion is beneficial. They naturally create a closed loop, enhancing the expressiveness of aggregated features.

Intrinsic Consistency between Vision and Semantics. For low-level vision tasks like image fusion, humans prioritize vision-related features for perception, while in high-level tasks such as semantic segmentation, machines focus on semantics-related features for accurate decision-making. This raises an intriguing philosophical question: do machines and humans perceive the world in similar ways? In other words, is there consistency between the features humans rely on for vision and those machines rely on for semantics? Visualizing channel features in SegMiF [14] and EAEFNet [11], used for image fusion and segmentation, reveals a strong correlation, as shown in Fig. 5. This observation leads us to explore the intrinsic consistency between vision and semantics, establishing a mutually reinforcing mechanism for both tasks.

3. Method

3.1. Overview

Our proposed MRFS elegantly couples image fusion and semantic segmentation tasks into a unified framework by implementing feature interaction through the proposed IGM-Att and PC-Att. We illustrate the overall framework in Fig. 6. It can be seen that MRFS continuously extracts and refines multi-modal features through n blocks. In each block, the encoder initially conducts feature extraction through $\Phi_n^{vi} = E_n^{vi}(\Theta_{n-1}^{vi})$ and $\Phi_n^{ir} = E_n^{ir}(\Theta_{n-1}^{ir})$, where E_n^{vi} and E_n^{ir} denote the nth encoder on the visible and infrared branches, respectively. Θ_{n-1}^{vi} and Θ_{n-1}^{ir} represent the visible and infrared features refined by the $(n-1)$th IGM-Att module. These features are initialized with the visible image I^{vi} and infrared image I^{ir} when $n = 1$. Then, we design an IGM-Att module to facilitate interaction between Φ_n^{vi} and Φ_n^{ir}. This module employs the CNN paradigm to predominantly emphasize vision-related local features. Additionally, it incorporates a gating mechanism designed to rectify misjudgments in the conventional pooling-based attention. Thus, a more effective mutual refinement of visible and infrared features is obtained by $\{\Theta_n^{vi}, \Theta_n^{ir}\} = V_n(\Phi_n^{vi}, \Phi_n^{ir})$, where $V_n(\cdot)$ denotes the func-

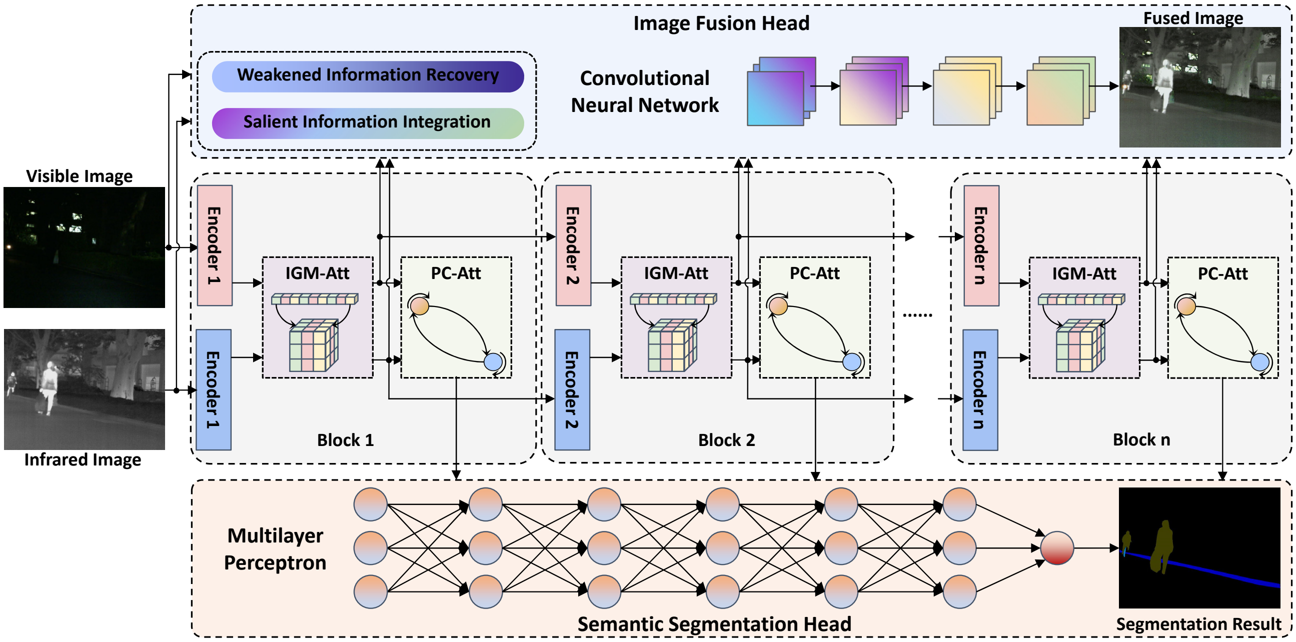


Figure 6. The overall framework of our proposed MRFS.

tion of the nth IGM-Att. Besides, PC-Att is developed to further facilitate interaction and aggregation of multi-modal features. In contrast to IGM-Att, PC-Att places greater emphasis on semantics-related features that necessitate a global understanding. Consequently, we establish a feature cyclic transfer and fusion architecture within the transformer paradigm. The aggregated features obtained after PC-Att processing can be expressed as $\Omega_n^f = S_n(\Theta_n^{vi}, \Theta_n^{ir})$, where $S_n(\cdot)$ represents the function of the nth PC-Att. Due to the comprehensive consideration of single-modal self-reinforcement and cross-modal mutual complementarity, the aggregated features Ω_n^f encompass complete information describing the scene. Afterward, the output of IGM-Att and source images are fed into the image fusion head F to generate the fused image: $I^f = F(\Theta_n^{vi}, \Theta_n^{ir}, I^{vi}, I^{ir})$. It adopts CNN architecture, incorporating functions of salient information integration and weakened information recovery, thereby effectively enhancing the visual quality of the fused image. Meanwhile, we feed the output of PC-Att into the MLP-based semantic segmentation head G to attain pixel-level classification decisions $I^s = G(\Omega_n^f)$. Through the synergistic optimization of image fusion and semantic segmentation, their performance can be improved in a mutually reinforcing manner.

3.2. IGM-Att

The IGM-Att module integrates a gating mechanism into the conventional pooling-based attention for visual completion, as shown in Fig. 7. First, it adopts conventional channel and spatial attention [33] to model the relative relationship between features. In the channel attention, $\Phi_n^{vi} \in \mathbb{R}^{H\times W\times C}$ and $\Phi_n^{ir} \in \mathbb{R}^{H\times W\times C}$ are concatenated along the channel dimension, and then reduced by the max and average pooling to capture the response properties of each channel. Next, we use MLP to project these maximum and average responses separately. The resulting vectors are summed and activated by the Sigmoid function to generate channel-attention weights, denoted as $W_n^c \in \mathbb{R}^{2\times 1\times 1\times C}$. This process can be formalized as $W_n^c = Sigmoid(MLP(AP(C(\Phi_n^{vi}, \Phi_n^{ir}))) + MLP(MP(C(\Phi_n^{vi}, \Phi_n^{ir}))))$. In the spatial attention, we use max and average pooling on both Φ_n^{vi} and Φ_n^{ir}, yielding four maps that capture their pixel-wise response properties. These maps are concatenated along the channel dimension, and then processed by two convolutions with the Sigmoid function to produce the spatial-attention weights, denoted as $W_n^s \in \mathbb{R}^{H\times W\times 1}$. We formalize this process as $W_n^s = Sigmoid(Conv(C(AP(\Phi_n^{vi}, \Phi_n^{ir}), MP(\Phi_n^{vi}, \Phi_n^{ir}))))$. We derive hybrid weights $W_n^{cs} \in \mathbb{R}^{2\times H\times W\times C}$, by multiplying channel-attention weights with spatial-attention weights, expressed as $W_n^{cs} = W_n^c \times W_s^c$. Nevertheless, as previously noted, such pooling-based attention may lead to mismatched feature validity scores. To address this, we introduce a gating mechanism to correct the obtained hybrid weights W_n^{cs}. Specifically, we first flatten Φ_n^{vi} and Φ_n^{ir}. The flattened features are then concatenated along the channel dimension and input into an MLP with a Sigmoid function to derive the gated correction scores $W_n^g \in \mathbb{R}^{H\times W\times C}$. We encapsulate the entire gating mechanism process as $W_n^g = Sigmoid(MLP(C(FL(\Phi_n^{vi}), FL(\Phi_n^{ir}))))$. Next,

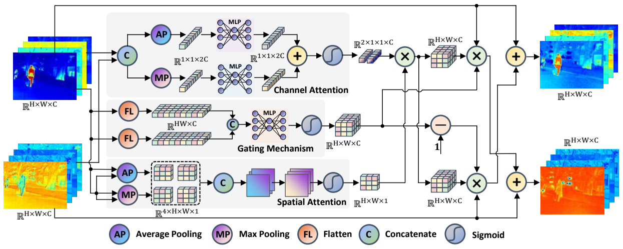

$\mathbb{R}^{H\times W\times C}$
C
AP
MP
MLP
$\mathbb{R}^{1\times1\times2C}$
$\mathbb{R}^{2\times1\times1\times C}$
Channel Attention
FL
$\mathbb{R}^{HW\times C}$
Gating Mechanism
1
Spatial Attention
$\mathbb{R}^{4\times H\times W\times1}$
$\mathbb{R}^{H\times W\times1}$
AP Average Pooling
MP Max Pooling
FL Flatten
C Concatenate
Sigmoid

Figure 7. Architecture of our proposed IGM-Att module.

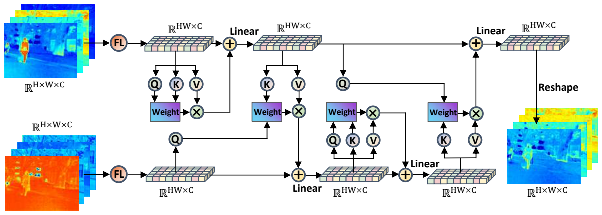


Figure 8. Architecture of our proposed PC-Att module.

the hybrid weights undergo refinement using the correction scores: $W_n^{csg-vi} = W_n^{cs}[0] \times W_n^g$ and $W_n^{csg-ir} = W_n^{cs}[1] \times (1 - W_n^g)$. This sophisticated strategy enhances the precision of feature importance assessments, reinforcing valuable features and effectively suppressing redundant ones. Finally, we refine features for visual completion using the obtained weights: $\Theta_n^{vi} = \Phi_n^{vi} + W_n^{csg-vi} \times \Phi_n^{ir}$ and $\Theta_n^{ir} = \Phi_n^{ir} + W_n^{csg-ir} \times \Phi_n^{vi}$.

3.3. PC-Att

The IGM-Att module leverages CNN-based attention, emphasizing the refinement of local visual features. In contrast, semantic segmentation necessitates global scene understanding capabilities. Therefore, we develop a PC-Att module for refined semantic completion, illustrated in Fig. 8. In the PC-Att module, we employ two information reinforcement strategies, namely, single-modal self-reinforcement and cross-modal mutual complementarity. For single-modal self-reinforcement, we compute Query ($Q^X \in \mathbb{R}^{HW\times C}$), Key ($K^X \in \mathbb{R}^{HW\times C}$), and Value ($V^X \in \mathbb{R}^{HW\times C}$) using linear layers from the feature $\Upsilon_X \in \mathbb{R}^{HW\times C}$. Q^X is then employed to query K^X for allocation scores, and utilize these scores to modulate V^X to obtain the refined value: $V^{X^m} = softmax(\frac{Q^X K^{X^T}}{\sqrt{d_k}})V^X$. Subsequently, the refined value V^{X^m} is used to enhance Υ^X, resulting in improved features $\Upsilon^{X^{self}} = Linear(\Upsilon^X + V^{X^m})$. This entire process is formally expressed as $\Upsilon^{X^{self}} = Self(\Upsilon^X)$. With this design, valuable cues from the single-modal image itself (*e.g.*, objects of the same class) can be leveraged for feature refinement. However, the ability of self-reinforcement is inherently limited, so we introduce the cross-modal mutual complementarity. Without loss of generality, we describe its process by using the X-modal feature Υ^X to enhance the Y-modal feature Υ^Y. Q^Y is derived from Υ^Y to represent the demand, while K^X and V^X are extracted from Υ^X. Following a procedure similar to self-reinforcement, we leverage Q^Y and K^X to compute the allocation score and modulate V^X: $V^{Y\leftarrow X^m} = softmax(\frac{Q^Y K^{X^T}}{\sqrt{d_k}})V^X$. The resulting cross-modal refined feature is obtained by $\Upsilon^{Y^{cross}} = Linear(\Upsilon_Y + V^{Y\leftarrow X^m})$. This process is formally simplified as $\Upsilon^{Y^{cross}} = Cross(\Upsilon^Y, \Upsilon^X)$. Referring to Fig. 8, in our PC-Att module, we flatten Θ_n^{vi} and Θ_n^{ir} and apply the self-reinforcement and cross-modal complementarity strategies in a closed loop: $\Upsilon_n^{vi^{self}} = Self(FL(\Theta_n^{vi}))$, $\Upsilon_n^{ir^{cross}} = Cross(FL(\Theta_n^{ir}), \Upsilon_n^{vi^{self}})$, $(\Upsilon_n^{ir^{cross}})^{self} = Self(\Upsilon_n^{ir^{cross}})$, $\Omega_n^f = (\Upsilon_n^{vi^{self}})^{cross} = Cross(\Upsilon_n^{vi^{self}}, (\Upsilon_n^{ir^{cross}})^{self})$. This process seamlessly integrates long-distance complete semantic information from infrared and visible images into the generated fused feature Ω_n^f.

3.4. Task Heads

Image Fusion Head. We adopt the CNN architecture in our image fusion head, as it has demonstrated effectiveness in low-level vision tasks, producing visually satisfying images [15, 20]. Initially, we integrate the function of weakened information recovery into it. Specifically, we apply common data augmentation strategies (*e.g.*, gamma transformation, contrast stretching) to process I^{vi}, I^{ir}, creating positive samples $\widetilde{I^{vi}}, \widetilde{I^{ir}}$. These samples can recover information lost due to weak reflected light and low contrast, offering valuable guidance for optimizing anchor points (*i.e.*, fused image I^f). Secondly, we introduce the function of salient information integration to preserve crucial characteristics. Concretely, the maximum function is applied to process positive samples in both gradient and intensity domains, constructing an explicit optimization goal. Anchor points are then directed to approach this goal, ensuring the preservation of significant contrast and rich textures. The corresponding integration loss is defined as $\mathcal{L}_{\text{inte}} = \left\|I^f - max(I^{vi}, I^{ir})\right\|_1 + \left\|\nabla I^f - max(\nabla I^{vi}, \nabla I^{ir})\right\|_1$. Besides, we enforce color consistency in YCbCr space to maintain scene colors, and the color loss is defined as $\mathcal{L}_{\text{color}} = \left\|Cb^f - Cb^{\widetilde{vi}}\right\|_1 + \left\|Cr^f - Cr^{\widetilde{vi}}\right\|_1$. Therefore, the total image fusion loss is summarized as $\mathcal{L}\text{fusion} = \mathcal{L}\text{inte} + \mathcal{L}\text{color}$.

Semantic Segmentation Head. We employ the MLP decoder from SegFormer [34] as the semantic segmentation head for its simplicity, lightweight nature, and effective understanding of global scene semantics. The conventional cross-entropy loss is used to constrain semantic segmentation, formalized as $\mathcal{L}_{\text{seg}} = -\sum P \log I^s$. Here, P refers to the label, and I^s indicates the classification probability output by the semantic segmentation head.

4. Experiments

4.1. Datasets and Implementation

Datasets. We evaluate the performance of our MRFS on semantic segmentation and image fusion tasks using the MFNet [6], PST900 [25], and FMB [14] datasets. These datasets consist of 1569, 1038, and 1500 pairs of infrared and visible images, each with sizes of 480×640, 720×1280, and 600×800, respectively. The number of image pairs used for testing is 393, 288, and 280, respectively.

Implementation Details. The semantic segmentation and image fusion tasks are jointly trained for 500 epochs to achieve adequate coupled learning. The initial learning rate is set to $6e^{-5}$, and the batch size is 3, utilizing the Adam optimizer. The encoders in Fig. 6 are configured as transformer-based structures following SegFormer[34], with the number of blocks set to 4 based on its original configuration. All experiments are conducted on the NVIDIA GeForce RTX 3090 GPU with 24GB memory and the AMD EPYC 7H12 64-Core Processor CPU.

4.2. Semantic Segmentation

We conduct comparative experiments for semantic segmentation, evaluating against nine state-of-the-art competitors: SeAFusion [27], SegFormer [34], EGFNet [48], LASNet [9], SegMiF [14], MDRNet+[45], SGFNet[31], MMSMCNet [50], and EAEFNet [11]. Due to unavailable training codes for EGFNet and MMSMCNet, we use their original reported results on the MFNet and PST900 datasets, and do not retrain them on the FMB dataset. We provide visualizations in Figs. 9 and 10 for objective comparison, excluding the PST900 dataset due to its sparse and poor image quality. Clearly, our method yields superior visual segmentation, characterized by optimal classification precision and comprehensive object delineation. For instance, our method effectively preserves the fine contours of pedestrians, yielding vivid postures, while other methods can only determine approximate areas. Further, we present quantitative results in Tables 1, 2, and 3. Our MRFS attains the highest mean pixel intersection-over-union (mIOU) across all three datasets. Overall, these findings substantiate that our method achieves state-of-the-art excellence in semantic segmentation.

4.3. Image Fusion

Comparative experiments are conducted to assess image fusion performance, involving eight state-of-the-art competitors: SDNet [40], U2Fusion [35], SeAFusion [27], DetFusion [26], DATFuse [29], CDDFuse [46], TGFuse [23], and SegMiF [14]. Due to the low quality of the PST900 dataset, comparisons for image fusion are performed on the MFNet

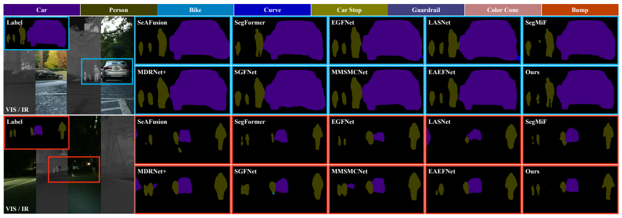


Figure 9. Qualitative segmentation on the MFNet dataset.

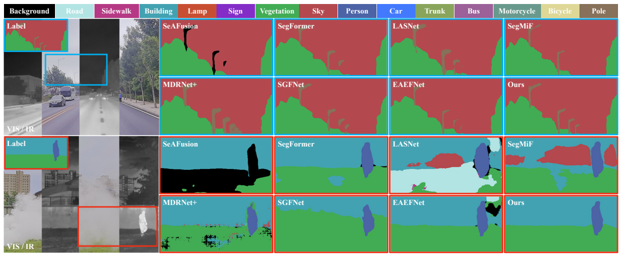


Figure 10. Qualitative segmentation on the FMB dataset.

Table 1. Quantitative segmentation on the MFNet dataset.

MFNet	Car	Person	Bike	Curve	Car Stop	Guar.	Cone	Bump	mIoU
SeAFusion	84.2	71.1	58.7	33.1	20.1	0.0	40.4	33.9	48.8
SegFormer	89.5	73.2	63.8	45.9	20.8	4.14	44.8	51.5	54.7
EGFNet	87.6	69.8	58.8	42.8	33.8	7.0	48.3	47.1	54.8
LASNet	84.2	67.1	56.9	41.1	39.6	18.9	48.8	40.1	54.9
SegMiF	87.8	71.4	63.2	47.5	31.1	0.0	48.9	50.3	56.1
MDRNet+	87.1	69.8	60.9	47.8	34.2	8.2	50.2	55.0	56.8
SGFNet	88.4	77.6	64.3	45.8	31.0	6.0	57.1	55.0	57.6
MMSMCNet	89.2	69.1	63.5	46.4	41.9	8.8	48.8	57.6	58.1
EAEFNet	87.6	72.6	63.8	48.6	35.0	14.2	52.4	58.3	58.9
Ours	89.4	75.4	65.0	49.0	37.2	5.4	53.1	58.8	**59.1**

Table 2. Quantitative segmentation on the PST900 dataset.

PST900	Hand-Drill	BackPack	Frie-Extinguisher	Survivor	mIoU
SeAFusion	65.6	59.6	41.1	29.5	58.9
SegFormer	74.3	86.4	61.1	69.3	78.1
EGFNet	64.7	83.1	71.3	74.3	78.5
LASNet	77.8	86.5	82.8	75.5	84.4
MDRNet+	63.0	76.3	63.5	71.3	74.6
SegMiF	66.0	81.4	76.3	75.5	79.7
MMSMCNet	62.4	89.2	73.3	74.7	79.8
SGFNet	82.8	75.8	79.9	72.7	82.1
EAEFNet	80.4	87.7	84.0	76.2	85.6
Ours	79.7	87.4	88.0	79.6	**86.9**

Table 3. Quantitative segmentation on the FMB dataset.

FMB	Car	Person	Truck	T-Lamp	T-Sign	Buil.	Vege.	Pole	mIoU
SeAFusion	76.2	59.6	15.1	34.4	68.0	80.1	83.5	38.4	51.9
SegFormer	76.5	68.4	38.7	20.9	70.6	81.4	83.8	43.9	56.3
LASNet	73.2	58.3	33.1	32.6	68.5	80.8	83.4	41.0	55.7
SegMiF	78.7	65.5	42.4	35.6	71.7	80.1	85.1	35.7	58.5
MDRNet+	75.4	67.0	27.0	41.4	68.4	79.8	82.7	45.3	55.5
SGFNet	75.0	67.2	34.6	45.8	71.4	78.2	82.7	42.8	56.0
EAEFNet	79.7	61.6	22.5	34.3	74.6	82.3	86.6	46.2	58.0
Ours	76.2	71.3	34.4	50.1	75.8	85.4	87.0	53.6	**61.2**

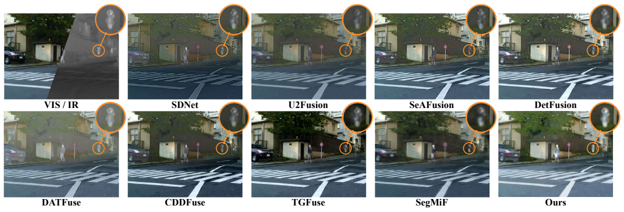


Figure 11. Qualitative image fusion on the MFNet dataset.

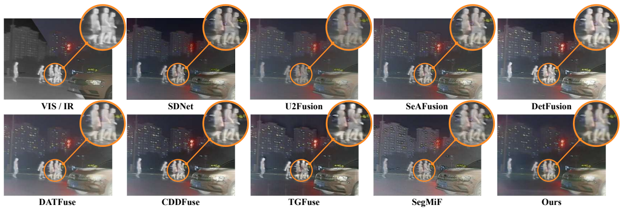


Figure 12. Qualitative image fusion on the FMB dataset.

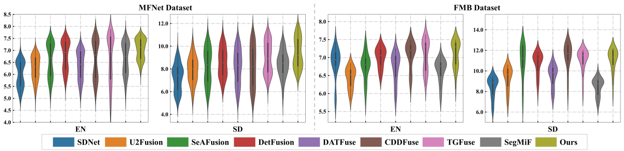


Figure 13. Quantitative fusion on the MFNet and FMB datasets.

and FMB datasets only. Visual results for objective evaluation are presented in Figs. 11 and 12. Our MRFS effectively recovers weak details, enhancing overall visual quality. For example, in Fig. 11, our fused image improves the visibility of the tree canopy and highlights faint thermal objects better than other methods, crucial for human visual perception. Additionally, Fig. 12 demonstrates another strength of our method. Leveraging salient information integration, MRFS effectively suppresses artifacts around thermal objects in infrared images, while other methods cannot. Moreover, due to the absence of ground truth in image fusion, existing reference metrics assessing similarity between fused and source images are biased by source image degradation. Hence, we employ non-reference metrics, entropy (EN) [24] and standard deviation (SD) [1], to evaluate the amount of information and contrast. As shown in Fig. 13, MRFS achieves the highest scores on most metrics, demonstrating its advanced image fusion performance.

4.4. Ablation Studies

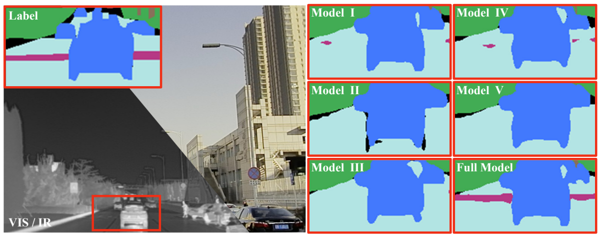


Figure 14. Qualitative segmentation of ablation studies.

Table 4. Quantitative segmentation of ablation studies.

Model	Car	Person	Truck	T-Lamp	T-Sign	Buil.	Vege.	Pole	mIoU
I	78.9	70.1	33.4	48.2	74.9	84.9	87.2	50.8	60.5
II	77.3	67.7	42.3	43.7	72.6	84.9	87.2	50.9	60.9
III	78.3	70.0	36.6	51.5	74.0	84.0	86.6	52.5	60.2
IV	78.8	68.7	31.9	43.5	72.8	85.3	87.7	52.3	59.5
V	75.4	69.7	24.6	51.4	73.8	85.3	97.0	53.6	59.8
Full model	76.2	71.3	34.4	50.1	75.8	85.4	87.0	53.6	**61.2**

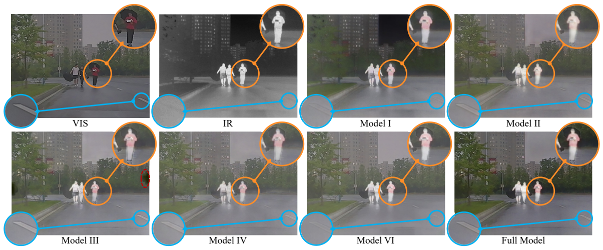


Figure 15. Qualitative fusion of ablation studies.

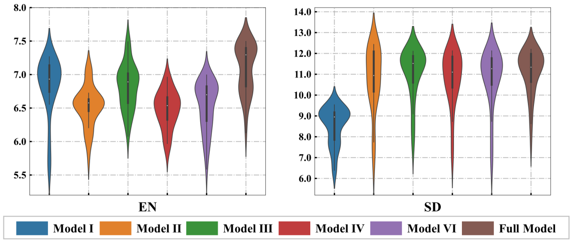


Figure 16. Quantitative fusion of ablation studies.

We conduct extensive ablation studies to evaluate specific designs in our method, examining six variants. **Model I:** replaces salient information integration with a proportional strategy [41]; **Model II:** omits weakened information recovery; **Model III:** substitutes IGM-Att with conventional pooling-based attention [3]; **Model IV:** replaces PC-Att with cross-attention-based feature integration [47]; **Model V:** removes the image fusion head; **Model VI:** eliminates the semantic segmentation head. Fig. 14 and Table 4 show the impact of these variants on semantic segmentation, where removing any design leads to decreased performance. Particularly, replacing PC-Att and removing the image fusion head cause significant performance degradation, emphasizing PC-Att's role in multi-modal feature fusion and the importance of image fusion in enhancing semantic segmentation. Figs. 15 and 16 illustrate these designs' roles in image fusion, where removing any diminishes performance. For instance, replacing IGM-Att introduces black artifacts, and removing the segmentation head reduces the visibility of ground traffic markings. These results underscore the efficacy of these designs in improving image fusion and highlight semantic segmentation's positive influence.

Table 5. Gains of IGM-Att & PC-Att to dual-stream competitors.

FMB	Car	Person	Truck	T-Lamp	T-Sign	Buil.	Vege.	Pole	mIoU
LASNet	73.2	58.3	33.1	32.6	68.5	80.8	83.4	41.0	55.7
LASNet+	75.1	56.6	37.0	45.1	73.2	80.2	83.2	43.3	**58.1**
SGFNet	75.0	67.2	34.6	45.8	71.4	78.2	82.7	42.8	56.0
SGFNet+	74.4	67.2	32.9	43.5	75.3	79.8	84.1	42.7	**57.7**

Table 6. Quantitative semantic verification with object detection.

Method	Segmentation	Image Fusion	FLOPs (G)	Params (M)
SegFormer	✓	✗	185.47	64.01
LASNet	✓	✗	371.03	93.58
MDRNet+	✓	✗	891.82	210.87
SGFNet	✓	✗	225.63	125.12
EAEFNet	✓	✗	316.49	147.21
SDNet	✗	✓	64.55	0.07
U2Fusion	✗	✓	633.09	0.659
DetFusion	✗	✓	278.62	83.04
DATFuse	✗	✓	8.68	0.01
CDDFuse	✗	✓	863.22	1.19
TGFuse	✗	✓	137.34	19.34
SeAFusion	✓	✓	102.53	13.06
SegMiF	✓	✓	526.20	45.60
Ours	✓	✓	219.16	134.97

4.5. Application Gain of IGM-Att and PC-Att

Our proposed IGM-Att and PC-Att are used to solve the problems of the feature validity score mismatch and insufficient feature aggregation, respectively. Therefore, it is plug-and-play for approaches using a dual-stream architecture similar to our MRFS. To evaluate the application gain of our IGM-Att and PC-Att, we transfer them to two dual-stream competitors, *i.e.*, LASNet and SGFNet. The quantitative gains on the FMB dataset are reported in Table 5. It can be seen that after using IGM-Att and PC-Att, LASNet and SGFNet achieved gains of 2.4 and 1.7, respectively. These results demonstrate the broad gains of our IGM-Att and PC-Att on the semantic segmentation task.

4.6. Complexity Discussion

We quantify the number of parameters and the FLOPs (tested on images with the size of 600 × 800) for various semantic segmentation and image fusion methods to gauge their complexity, as detailed in Table 6. While our method may not attain optimal performance on these two indicators, it remains acceptable and even outperforms some dedicated semantic segmentation and image fusion methods.

4.7. Semantic Verification with Object Detection

The ablation studies have demonstrated that semantic segmentation can promote image fusion performance. Can this promotion be understood as an injection of broad semantics? To investigate, we initially generate fused results using various fusion methods on the LLVIP dataset [8]. Subsequently, we retrain the YOLO-v5 detector using these fused images to assess their embedded semantic attributes. Each retraining involves 400 fused images for training and 120 for testing. The results in Fig. 17 and Table 7 demonstrate that our MRFS outperforms most other methods in detection performance. The only exception is DetFusion, which shows comparable performance, as it follows an object detection-driven technical route within the image fusion domain. Overall, our MRFS effectively enhances the semantic attributes of fused images, offering potential value for applications in other high-level semantic tasks.

Table 7. Quantitative semantic verification with object detection.

	Precision	Recall	mAP@0.5	mAP@0.75	mAP@[0.5:0.95]
VIS	97.0	94.1	96.3	73.4	63.8
IR	97.6	97.3	98.9	88.1	72.8
SDNet	97.6	97.3	98.8	90.7	73.9
U2Fusion	96.2	98.5	97.8	91.0	73.7
SeAFusion	97.7	99.1	99.1	91.1	73.4
DetFusion	97.4	98.5	99.2	89.6	74.2
DATFuse	96.5	99.1	99.3	90.1	73.6
CDDFuse	97.9	98.2	99.1	91.4	73.6
TGFuse	98.2	97.6	98.9	90.3	73.7
SegMiF	98.2	97.3	98.8	91.0	73.4
Ours	98.2	97.6	99.3	92.9	74.2

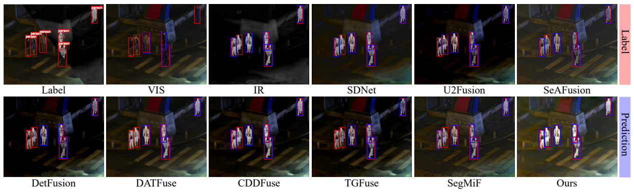


Figure 17. Qualitative semantic verification with object detection.

5. Conclusion

This work proposes a coupled learning framework for infrared and visible image fusion and semantic segmentation. Leveraging the intrinsic consistency between vision and semantics, it makes these tasks mutually reinforcing, achieving dual improvement. Firstly, weakened information recovery and salient information integration are incorporated into the image fusion task, ensuring fused results align with human perception. Secondly, the IGM-Att and PC-Att modules address challenges from mismatched feature validity scores and enhance feature aggregation sufficiency. Cascading these modules facilitates implicit convergence of vision-related and semantics-related features, enabling mutual guidance in learning processes and leading to improved solutions. Extensive experiments on public datasets demonstrate that our MRFS achieves both visual satisfaction and decision accuracy.

Acknowledgments

This work was supported by the NSFC (62276192) and the Tencent-University cooperation project (JR2023TEG011).

References

[1] V Aslantas and Emre Bendes. A new image quality metric for image fusion: The sum of the correlations of differences. *AEU-International Journal of Electronics and Communications*, 69(12):1890–1896, 2015. 7

[2] Xiaokang Chen, Kwan-Yee Lin, Jingbo Wang, Wayne Wu, Chen Qian, Hongsheng Li, and Gang Zeng. Bi-directional cross-modality feature propagation with separation-and-aggregation gate for rgb-d semantic segmentation. In *Proceedings of the European Conference on Computer Vision*, pages 561–577, 2020. 1

[3] Fuqin Deng, Hua Feng, Mingjian Liang, Hongmin Wang, Yong Yang, Yuan Gao, Junfeng Chen, Junjie Hu, Xiyue Guo, and Tin Lun Lam. Feanet: Feature-enhanced attention network for rgb-thermal real-time semantic segmentation. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 4467–4473, 2021. 2, 3, 7

[4] Lijun Ding and Ardeshir Goshtasby. On the canny edge detector. *Pattern Recognition*, 34(3):721–725, 2001. 2

[5] Haolong Fu, Shixun Wang, Puhong Duan, Changyan Xiao, Renwei Dian, Shutao Li, and Zhiyong Li. Lraf-net: Long-range attention fusion network for visible-infrared object detection. *IEEE Transactions on Neural Networks and Learning Systems*, 2023. 1

[6] Qishen Ha, Kohei Watanabe, Takumi Karasawa, Yoshitaka Ushiku, and Tatsuya Harada. Mfnet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5108–5115, 2017. 6

[7] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7132–7141, 2018. 3

[8] Xinyu Jia, Chuang Zhu, Minzhen Li, Wenqi Tang, and Wenli Zhou. Llvip: A visible-infrared paired dataset for low-light vision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3496–3504, 2021. 8

[9] Gongyang Li, Yike Wang, Zhi Liu, Xinpeng Zhang, and Dan Zeng. Rgb-t semantic segmentation with location, activation, and sharpening. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(3):1223–1235, 2023. 1, 6

[10] Hui Li, Tianyang Xu, Xiao-Jun Wu, Jiwen Lu, and Josef Kittler. Lrrnet: A novel representation learning guided fusion network for infrared and visible images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):11040–11052, 2023. 1

[11] Mingjian Liang, Junjie Hu, Chenyu Bao, Hua Feng, Fuqin Deng, and Tin Lun Lam. Explicit attention-enhanced fusion for rgb-thermal perception tasks. *IEEE Robotics and Automation Letters*, 8(7):4060–4067, 2023. 3, 6

[12] Jinyuan Liu, Xin Fan, Zhanbo Huang, Guanyao Wu, Risheng Liu, Wei Zhong, and Zhongxuan Luo. Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5802–5811, 2022. 1

[13] Jinyuan Liu, Xin Fan, Ji Jiang, Risheng Liu, and Zhongxuan Luo. Learning a deep multi-scale feature ensemble and an edge-attention guidance for image fusion. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(1): 105–119, 2022. 1

[14] Jinyuan Liu, Zhu Liu, Guanyao Wu, Long Ma, Risheng Liu, Wei Zhong, Zhongxuan Luo, and Xin Fan. Multi-interactive feature learning and a full-time multi-modality benchmark for image fusion and segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8115–8124, 2023. 1, 2, 3, 6

[15] Yu Liu, Xun Chen, Hu Peng, and Zengfu Wang. Multi-focus image fusion with a deep convolutional neural network. *Information Fusion*, 36:191–207, 2017. 5

[16] Fuya Luo, Yunhan Li, Guang Zeng, Peng Peng, Gang Wang, and Yongjie Li. Thermal infrared image colorization for nighttime driving scenes with top-down guided attention. *IEEE Transactions on Intelligent Transportation Systems*, 23 (9):15808–15823, 2022. 1

[17] Jiayi Ma, Yong Ma, and Chang Li. Infrared and visible image fusion methods and applications: A survey. *Information Fusion*, 45:153–178, 2019. 1

[18] Jiayi Ma, Hao Zhang, Zhenfeng Shao, Pengwei Liang, and Han Xu. Ganmcc: A generative adversarial network with multiclassification constraints for infrared and visible image fusion. *IEEE Transactions on Instrumentation and Measurement*, 70:1–14, 2021. 1, 2

[19] Jiayi Ma, Linfeng Tang, Fan Fan, Jun Huang, Xiaoguang Mei, and Yong Ma. Swinfusion: Cross-domain long-range learning for general image fusion via swin transformer. *IEEE/CAA Journal of Automatica Sinica*, 9(7):1200–1217, 2022. 1

[20] Jinshan Pan, Deqing Sun, Jiawei Zhang, Jinhui Tang, Jian Yang, Yu-Wing Tai, and Ming-Hsuan Yang. Dual convolutional neural networks for low-level vision. *International Journal of Computer Vision*, 130(6):1440–1458, 2022. 5

[21] Chengli Peng, Kaining Zhang, Yong Ma, and Jiayi Ma. Cross fusion net: A fast semantic segmentation network for small-scale semantic information capturing in aerial scenes. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–13, 2022. 2, 3

[22] Jia Qi, Tengfei Liang, Wu Liu, Yidong Li, and Yi Jin. A generative-based image fusion strategy for visible-infrared person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(1):518–533, 2024. 1

[23] Dongyu Rao, Tianyang Xu, and Xiao-Jun Wu. Tgfuse: An infrared and visible image fusion approach based on transformer and generative adversarial network. *IEEE Transactions on Image Processing*, 2023. 1, 6

[24] J Wesley Roberts, Jan A Van Aardt, and Fethi Babikker Ahmed. Assessment of image fusion procedures using entropy, image quality, and multispectral classification. *Journal of Applied Remote Sensing*, 2(1):023522, 2008. 7

[25] Shreyas S Shivakumar, Neil Rodrigues, Alex Zhou, Ian D Miller, Vijay Kumar, and Camillo J Taylor. Pst900: Rgb-thermal calibration, dataset and segmentation network. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pages 9441–9447, 2020. 6

[26] Yiming Sun, Bing Cao, Pengfei Zhu, and Qinghua Hu. Detfusion: A detection-driven infrared and visible image fusion network. In *Proceedings of the ACM International Conference on Multimedia*, pages 4003–4011, 2022. 6

[27] Linfeng Tang, Jiteng Yuan, and Jiayi Ma. Image fusion in the loop of high-level vision tasks: A semantic-aware realtime infrared and visible image fusion network. *Information Fusion*, 82:28–42, 2022. 1, 2, 6

[28] Wei Tang, Fazhi He, and Yu Liu. Ydtr: Infrared and visible image fusion via y-shape dynamic transformer. *IEEE Transactions on Multimedia*, 25:5413–5428, 2023. 2

[29] Wei Tang, Fazhi He, Yu Liu, Yansong Duan, and Tongzhen Si. Datfuse: Infrared and visible image fusion via dual attention transformer. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(7):3159–3172, 2023. 6

[30] Qilong Wang, Banggu Wu, Pengfei Zhu, Peihua Li, Wangmeng Zuo, and Qinghua Hu. Eca-net: Efficient channel attention for deep convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11534–11542, 2020. 3

[31] Yike Wang, Gongyang Li, and Zhi Liu. Sgfnet: Semanticguided fusion network for rgb-thermal semantic segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(12):7737–7748, 2023. 1, 6

[32] Zhishe Wang, Wenyu Shao, Yanlin Chen, Jiawei Xu, and Lei Zhang. A cross-scale iterative attentional adversarial fusion network for infrared and visible images. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(8):3677–3688, 2023. 1

[33] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision*, pages 3–19, 2018. 3, 4

[34] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021. 5, 6

[35] Han Xu, Jiayi Ma, Junjun Jiang, Xiaojie Guo, and Haibin Ling. U2fusion: A unified unsupervised image fusion network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1):502–518, 2022. 1, 2, 6

[36] Han Xu, Jiteng Yuan, and Jiayi Ma. Murf: Mutually reinforcing multi-modal image registration and fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):12148–12166, 2023. 1

[37] Qin Xu, Yiming Mei, Jinpei Liu, and Chenglong Li. Multimodal cross-layer bilinear pooling for rgbt tracking. *IEEE Transactions on Multimedia*, 24:567–580, 2022. 1

[38] Xunpeng Yi, Han Xu, Hao Zhang, Linfeng Tang, and Jiayi Ma. Diff-retinex: Rethinking low-light image enhancement with a generative diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12302–12311, 2023. 1

[39] Xiaowen Ying and Mooi Choo Chuah. Uctnet: Uncertaintyaware cross-modal transformer network for indoor rgb-d semantic segmentation. In *Proceedings of the European Conference on Computer Vision*, pages 20–37, 2022. 2, 3

[40] Hao Zhang and Jiayi Ma. Sdnet: A versatile squeeze-anddecomposition network for real-time image fusion. *International Journal of Computer Vision*, 129(10):2761–2785, 2021. 6

[41] Hao Zhang, Han Xu, Yang Xiao, Xiaojie Guo, and Jiayi Ma. Rethinking the image fusion: A fast unified image fusion network based on proportional maintenance of gradient and intensity. In *Proceedings of the AAAI conference on artificial intelligence*, pages 12797–12804, 2020. 1, 2, 7

[42] Hao Zhang, Han Xu, Xin Tian, Junjun Jiang, and Jiayi Ma. Image fusion meets deep learning: A survey and perspective. *Information Fusion*, 76:323–336, 2021. 1

[43] Jiaming Zhang, Huayao Liu, Kailun Yang, Xinxin Hu, Ruiping Liu, and Rainer Stiefelhagen. Cmx: Cross-modal fusion for rgb-x semantic segmentation with transformers. *IEEE Transactions on Intelligent Transportation Systems*, 24(12): 14679–14694, 2023. 1

[44] Xingchen Zhang and Yiannis Demiris. Visible and infrared image fusion using deep learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8):10535–10554, 2023. 2

[45] Shenlu Zhao, Yichen Liu, Qiang Jiao, Qiang Zhang, and Jungong Han. Mitigating modality discrepancies for rgb-t semantic segmentation. *IEEE Transactions on Neural Networks and Learning Systems*, 2023. 1, 2, 6

[46] Zixiang Zhao, Haowen Bai, Jiangshe Zhang, Yulun Zhang, Shuang Xu, Zudi Lin, Radu Timofte, and Luc Van Gool. Cddfuse: Correlation-driven dual-branch feature decomposition for multi-modality image fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5906–5916, 2023. 2, 6

[47] Heng Zhou, Chunna Tian, Zhenxi Zhang, Qizheng Huo, Yongqiang Xie, and Zhongbo Li. Multispectral fusion transformer network for rgb-thermal urban scene semantic segmentation. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2022. 7

[48] Wujie Zhou, Shaohua Dong, Caie Xu, and Yaguan Qian. Edge-aware guidance fusion network for rgb-thermal scene parsing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3571–3579, 2022. 6

[49] Wujie Zhou, Fan Sun, Qiuping Jiang, Runmin Cong, and Jenq-Neng Hwang. Wavenet: Wavelet network with knowledge distillation for rgb-t salient object detection. *IEEE Transactions on Image Processing*, 32:3027–3039, 2023. 1

[50] Wujie Zhou, Han Zhang, Weiqing Yan, and Weisi Lin. Mmsmcnet: Modal memory sharing and morphological complementary networks for rgb-t urban scene semantic segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(12):7096–7108, 2023. 6

[51] Wujie Zhou, Shaohua Dong, Meixin Fang, and Lu Yu. Cacfnet: Cross-modal attention cascaded fusion network for rgb-t urban scene parsing. *IEEE Transactions on Intelligent Vehicles*, 9(1):1919–1929, 2024. 1, 2, 3