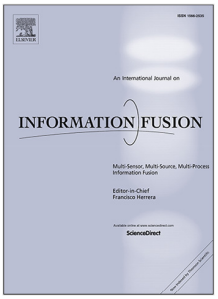
Journal Pre-proof



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

Linfeng Tang, Hao Zhang, Han Xu, Jiayi Ma

PII: S1566-2535(23)00186-0
DOI: https://doi.org/10.1016/j.inffus.2023.101870
Reference: INFFUS 101870

To appear in: *Information Fusion*

Received date : 17 April 2023
Revised date : 31 May 2023
Accepted date : 31 May 2023

Please cite this article as: L. Tang, H. Zhang, H. Xu et al., 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, *Information Fusion* (2023), doi: https://doi.org/10.1016/j.inffus.2023.101870.

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

Linfeng Tang, Hao Zhang, Han Xu and Jiayi Ma*

Electronic Information School, Wuhan University, Wuhan 430072, China

ARTICLE INFO

Keywords:
Image fusion
High-level vision task
Progressive semantic injection
Scene fidelity
Feature-level fusion

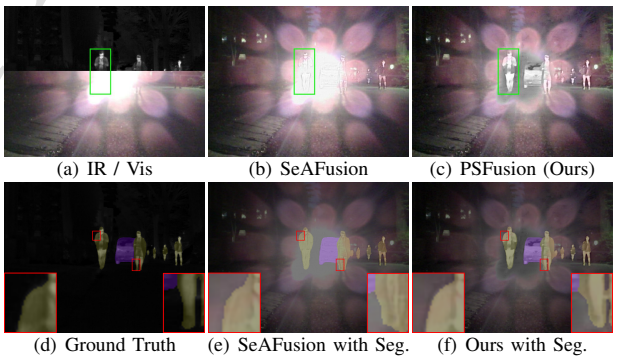
ABSTRACT

Image fusion aims to integrate complementary characteristics of source images into a single fused image that better serves human visual observation and machine vision perception. However, most existing image fusion algorithms primarily focus on improving the visual appeal of fused images. Although there are some semantic-driven methods that consider semantic requirements of downstream applications, none of them have demonstrated the potential of image-level fusion compared to feature-level fusion, which fulfills high-level vision tasks directly on multi-modal features rather than on a fused image. To overcome these limitations, this paper presents a practical infrared and visible image fusion network based on progressive semantic injection and scene fidelity constraints, termed PSFusion. First of all, the sparse semantic perception branch extracts sufficient semantic features, which are then progressively integrated into the fusion network using the semantic injection module to fulfill the semantic requirements of high-level vision tasks. The scene fidelity path within the scene restoration branch is devised to ensure that the fusion features contain complete information for reconstructing the source images. Additionally, the contrast mask and salient target mask are employed to construct the fusion loss to maintain impressive visual effects of fusion results. In particular, we provide quantitative and qualitative analyses to demonstrate the potential of image-level fusion compared to feature-level fusion for high-level vision tasks. With the rapid advancement of large-scale models, image-level fusion can expeditiously leverage the advantages of multi-modal data and state-of-the-art (SOTA) unimodal segmentation to achieve superior performance. Furthermore, extensive comparative experiments demonstrate the superiority of our PSFusion over SOTA image-level fusion alternatives in terms of visual appeal and high-level semantics. Even under harsh circumstances, our method offers satisfactory fusion results to serve subsequent high-level vision applications. The source code is available at https://github.com/Linfeng-Tang/PSFusion.

1. Introduction

Sensors with diverse imaging modalities are typically capable of describing the imaging scene from various perspectives [1, 2]. For instance, the popular visible camera generates images by capturing reflected light. It can vividly portray objects under well-defined circumstances but is susceptible to harsh environments, such as nighttime, fog, or occlusion. In contrast, the infrared sensor captures thermal radiation information, which is effective in emphasizing prominent targets and is robust in extreme environments. However, the thermal images generated by the infrared sensor typically suffer from low resolution and fail to depict detailed information. The complementary properties of the two sensors prompt researchers to fuse infrared and visible images and synthesize an informative fused image.

A typical example is shown in Figure 1. At night, the visible image, although providing a better perception of the surroundings, drowns out pedestrians and vehicles in the light. Benefiting from the unique imaging principle, the infrared image can clearly present pedestrians and vehicles. As illustrated in Figure 1 (c), an outstanding image fusion approach can effectively integrate prominent targets and environmental details to synthesize a fused image that contributes to both visual perception and machine vision. Sufficient information aggregation and excellent visual effects enable image fusion, to be broadly applied in various practical applications, such as nighttime driving assistance, video surveillance [4], object detection [5], tracking [6], semantic segmentation [7], and more.

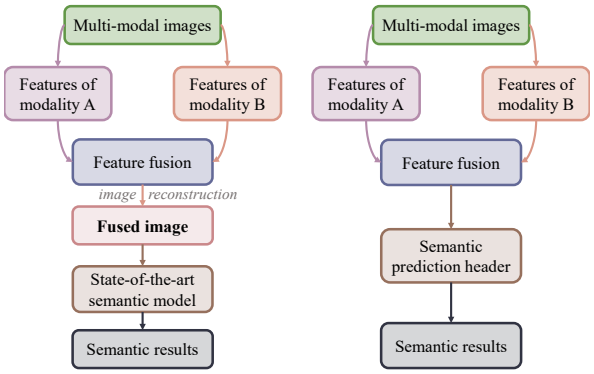


(a) IR / Vis (b) SeAFusion (c) PSFusion (Ours)
(d) Ground Truth (e) SeAFusion with Seg. (f) Ours with Seg.

Figure 1: Comparison of fusion and segmentation results between SeAFusion [3] and our method under harsh conditions.

In recent years, the infrared and visible image fusion technology continues to attract increasing attention because of its practicability. Infrared and visible image fusion can be roughly divided into vision perception-oriented approaches

*Corresponding author
linfeng0419@gmail.com (L. Tang); zhpersonalbox@gmail.com (H. Zhang); xu_han@whu.edu.cn (H. Xu); jyma2010@gmail.com (J. Ma)
ORCID(s): 0000-0003-3264-3265 (J. Ma)



(a) Image-level fusion-based scheme (b) Feature-level fusion-based scheme

Figure 2: Schematic of image-level fusion and feature-level fusion for high-level vision tasks.

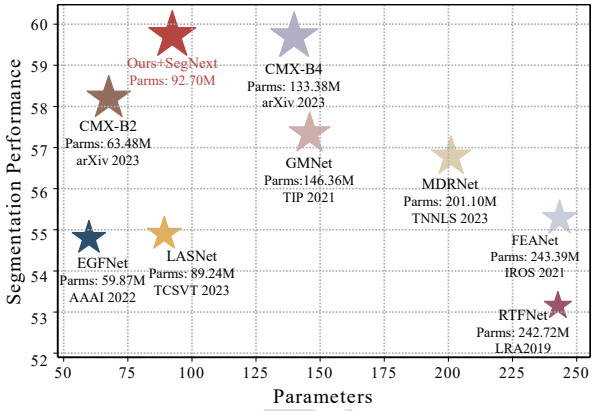


Figure 3: Comparison of the computational complexity between feature-level fusion and image-level fusion for the semantic segmentation task.

and semantic-driven approaches. The early image fusion methods primarily pursue better visual effects by introducing multi-scale transforms [8], subspace transforms [9], sparse representations [10], and saliency analysis [11]. With the rise of deep learning, researchers have also developed deep models to improve the visual performance of fusion results by incorporating Convolutional Neural Network (CNN) [12], Auto-Encoder (AE) [13], Generative Adversarial Network (GAN) [14], and Transformer [15]. Although the aforementioned approaches, especially the deep learning-based ones, achieve satisfactory fusion results, none of them has considered how to facilitate subsequent high-level vision tasks.

To address this gap, Tang *et al.* first proposed a semantic-aware fusion framework named SeAFusion, by attaching a segmentation model behind the fusion network, as presented in Figure 2 (a). SeAFusion aims to maintain excellent apparent effects while reinforcing semantic information in fused images [3]. Following this, Liu *et al.* [16] and Sun *et al.* [17] also devised object detection-driven image fusion approaches to impose the fusion network to retain more semantic information from the detection perspective. However, these methods utilize specific high-level models to constrain the fusion results, which may limit the generalization of the fused images to other models. Besides, SeAFusion solely relies on the maximum selection strategy to constrain the intensity of fused images, which may limit the potential of image fusion in some extreme circumstances. As depicted in Figure 1 (b), the fused image of SeAFusion does not effectively present pedestrians in the highlighted region due to interference of the vehicle light, although they are remarkable in the infrared image. The deficiency in the fusion result also leads to the segmentation model being unable to accurately segment objects, especially the edge regions.

It is worth noting that feature-level fusion is a more mainstream and straightforward solution for high-level vision tasks [18, 19]. As shown in Figure 2 (b), it fulfills the high-level vision tasks directly on the multi-modal fusion features and does not need to generate a fused image in advance. Feature-level fusion typically deploys well-established backbones, *i.e.*, feature extraction networks, to extract informative semantic features from source images. Subsequently, specific fusion modules are devised to integrate complementary features based on the employed backbone, and task-specific prediction headers are applied to accomplish the desired tasks. This scheme utilizes either a modality-shared feature extraction network or two independent feature extraction networks to extract semantic features. On the one hand, the modality-shared feature extraction network is unable to adapt effectively to the domain variation between infrared and visible images, which leads to significant performance degradation, as evidenced by EGFNet [20] and LASNet [21] in Figure 3. On the other hand, two independent feature extraction branches (especially large-scale backbones) usually lead to twice the computational load, as demonstrated by GMNet [18] and MDRNet [22].

Additionally, the existing feature-level fusion-based methods are tailored solely for the specific mission (*e.g.*, semantic segmentation, object detection, and tracking), which fail to be effectively generalized to other tasks. Furthermore, backbones based on unimodal input, such as Transformer [23] and ConvNeXt [24], are evolving rapidly. However, applying the existing feature-level fusion modules to these novel backbones often requires tedious re-designs. The gap between the feature fusion modules and backbones limits the potential of feature-level fusion-based schemes to pursue higher performance. It is worth emphasizing that the existing semantic-driven image fusion approaches only demonstrate the superiority of fused images over unimodal images (*i.e.*, infrared or visible images) for high-level vision tasks. Nevertheless, they do not excavate the potential of image-level fusion compared to feature-level fusion, which leaves the development of image-level fusion in a quandary.

To overcome the limitations of existing image fusion algorithms, we propose a practical infrared and visible image fusion network based on progressive semantic injection and scene fidelity, abbreviated as PSFusion. Our method includes the scene restoration branch with an image fusion

path and the sparse semantic prediction branch, both of which share multi-scale feature extraction networks. We progressively inject semantic information that is collectively constrained by multiple semantic perception tasks into the scene restoration branch using semantic injection modules at the feature level. This allows our fusion results to contain abundant semantic cues, which are friendly and robust for arbitrary high-level models. Moreover, we introduce a scene fidelity path in the scene restoration branch that is responsible for reconstructing source images from the fusion features to constrain the fusion features to contain all complete information of source images. Additionally, the superficial detail fusion module and profound semantic fusion module are devised to aggregate structural information in shallow features and semantic information in deep features, respectively. Importantly, we fully excavate the potential of image-level fusion for high-level vision tasks. As presented in Figure 3, when a unimodal segmentation model takes our fusion results as input, it can achieve equivalent performance to the state-of-the-art (SOTA) multi-modal segmentation models based on feature-level fusion. In particular, our solution only utilizes one feature extraction branch to extract semantic features from fused images, resulting in fewer parameters compared to SOTA feature-level fusion-based multi-modal schemes.

The main contributions of this paper can be summarized as follows:

- We demonstrate, for the first time, that for high-level vision tasks, multi-modal image-level fusion can achieve comparable performance to multi-modal feature-level fusion with a lower computational load. It justifies the necessity of image fusion in high-level vision tasks.
- We progressively inject semantic features into the fusion network at the feature level, thereby ensuring that the fusion results with abundant semantic cues are friendly and robust for arbitrary high-level backbones. Moreover, a scene fidelity path paralleling the image fusion path is devised to constrain the fusion modules to preserve complete information of source images.
- Extensive experiments demonstrate the superiority of our proposed method in terms of visual perception and high-level semantics over both image-level and feature-level fusion algorithms.

The remainder of this paper is organized as follows. In Section 2, we briefly introduce the related works of image fusion, unimodal semantic segmentation, and multi-modal segmentation. In Section 3, we elaborate on the proposed PSFusion, including the overall framework and loss function. Section 4 illustrates the impressive performance of our method in comparison with other alternatives, especially the feature-level fusion-based multi-modal segmentation models, followed by some concluding remarks in Section 5.

2. Related work

In this section, several typical works related to our method are presented. Firstly, we review some representative infrared and visible image fusion methods to introduce the development of image-level fusion. Subsequently, we provide an overview of popular backbones for unimodal semantic segmentation, as well as mainstream multi-modal segmentation approaches based on feature-level fusion.

2.1. Infrared and visible image fusion

2.1.1. Vision-perception oriented image fusion

In the earliest ages, infrared and visible image fusion methods aim solely to present the complete information from source images in a single fused image. In order to ensure that the synthesized fusion results are more consistent with the human visual perception system, several image processing techniques are introduced into the field of image fusion. These techniques include the Laplacian pyramid, discrete wavelet [25], non-subsampled contourlet transform [26], latent low-rank representation [10], sparse representation [27], principal component analysis [9], and non-negative matrix factorization [28]. Moreover, Ma *et al.* introduced the idea of total variation into image fusion and proposed GTF [29], which defines image fusion as overall intensity maintenance and texture structure preservation to obtain high-contrast fusion results. Additionally, some researchers attempted to incorporate the advantages of diverse techniques to generate more satisfying fusion results [30, 11]. However, employing traditional image processing techniques for image fusion tasks often involves complex activity level measurements and hand-crafted fusion rules. As a result, these methods fail to effectively integrate semantic information and respond to complex scenarios.

The thriving development of deep learning also offers new opportunities for image fusion. Deep learning-based approaches for infrared and visible image fusion can be broadly classified into pre-trained fusion frameworks and end-to-end fusion frameworks. The pre-trained fusion framework involves training an auto-encoder on large-scale datasets to perform feature extraction and image reconstruction, which concentrates on designing network architectures and fusion strategies [13, 31]. Li *et al.* proposed the first pre-trained fusion model, termed DenseFuse [13], consisting of three components, *i.e.*, encoder layer, fusion layer, and decoder layer. They utilized the element-wise addition or l_1-norm strategies in the fusion layer and introduced dense connections in the encoder layer to obtain satisfactory fusion results. Subsequently, they introduced the multi-scale architecture and nest connection to extract more comprehensive features [32, 33]. In addition, Zhao *et al.* developed a new encoder for multi-scale decomposition and to extract detailed as well as background features [34]. Similarly, Tang *et al.* incorporated Retinex theory to enhance the fusion results at night by designing multiple encoders to disentangle the illumination and reflection components of visible images [35]. All the above approaches reinforce the feature extraction capability by modifying the network

architecture. However, they employ hand-crafted fusion strategies to integrate deep features, which may limit the performance improvement of the pre-trained fusion frameworks. To tackle this challenge, Xu *et al.* deployed classifiers to perform activate level measurements and assign fusion weight for deep features from the classification perspective [36]. This idea enables fusion strategies to be learnable as well and improves the interpretability of deep models.

The end-to-end fusion framework can effectively eliminate the burden of hand-crafted fusion strategies, which achieves image fusion through elaborately designing loss functions, network architectures, and learning paradigms. Researchers developed many loss functions based on the properties of image fusion to provide sufficient guidance prior to network training. On the basis of the common intensity and gradient loss, Ma *et al.* designed a fusion loss based on the significant target mask to selectively fuse target and background regions. [12]. Moreover, considering the illumination variation, they devised an illumination-aware loss function [37]. The structural similarity (SSIM) loss [38] and perceptual loss [39] are also introduced to constrain the fusion results in order to avoid structural information distortion. In addition, various novel network architectures, such as residual block [12], aggregated residual dense block [38], and gradient residual dense block [3], as well as fusion modules, including cross-modality differential aware fusion module [37], interaction fusion module[40], global spatial attention module [41], and biphasic recurrent fusion module [42], are developed to guarantee the visual appeal of fusion results. Due to the absence of authentic fused images, researchers also introduced some novel learning paradigms, such as generative adversarial mechanisms [14] and contrastive learning. Ma *et al.* first introduced the generative adversarial network into the field of image fusion and proposed FusionGAN [14], which utilizes the discriminator to force the generator to retain more texture details from visible images. The inheritors of FusionGAN, such as DDcGAN [43], AttentionFGAN [44], and SDDGAN [45], also devise dual discriminators to avoid modal imbalance caused by a single discriminator.

In recent years, transformers [23] have demonstrated outstanding potential to outperform CNN in both high-level [46, 47, 48, 49] and low-level [50, 51, 52] vision tasks. Therefore, some Transformer-based fusion models, such as SwinFusion [15], IFT [53], and AFT [54], are developed to fully explore the long-range dependencies in source images. Considering that infrared and visible images usually suffer from varying degrees of misalignment in practice, the latest approaches (*e.g.*, RFNet [55], UMF-CMGR [40], ReCoNet [42], and SuperFusion [41]) incorporate an alignment module before the fusion module to register the misalignment in source images. Furthermore, some methods, including PMGI [56], IFCNN [57], U2Fusion [58], DeFusion [59], and SwinFusion [3] model various image fusion tasks uniformly, as there are commonalities between these tasks. In particular, U2Fusion trains a unified model for multiple fusion tasks, which promotes cross-fertilization between diverse fusion tasks. However, all of the above approaches primarily focus on integrating complementary information in source images and enhancing the visual appeal of fused images. None of them consider the requirements of subsequent high-level vision tasks on the fusion results.

2.1.2. Semantic-driven image fusion

In order to meet the requirements of high-level vision tasks on image fusion, Tang *et al.* proposed the first semantic-driven image fusion algorithm, termed SeAFusion [3]. SeAFusion cascades a semantic segmentation network behind the image fusion network to provide feedback on semantic requirements for the fusion network through gradient back-propagation. Similarly, Liu *et al.* replaced the segmentation network with object detection work and proposed TarDAL [16], which constrains the fusion network to preserve abundant semantic information from the perspective of object detection. It should be noted that the jointly trained model of TarDAL may introduce significant noise in the fusion results, so two separate versions are needed for human perception and machine vision. Additionally, Sun *et al.* trained two object detection models based on infrared and visible images, respectively, and constrained the fusion network using the two detection models jointly [17]. The attention maps generated in the detection network are transferred into the fusion network to achieve sufficient information aggregation. Yet, the domain variation between fused images and source images makes it difficult for the detection models trained on source images to accurately measure semantic information contained in fusion results. Besides, the above solutions employ the specific model to constrain the ultimate fusion results, which may limit the applicability of the fused images to other models. To tackle this challenge, we propose explicitly injecting the semantic features involved in the semantic perception tasks into the fusion network at the feature level.

2.2. Semantic segmentation

2.2.1. Single-modal semantic segmentation

Semantic segmentation is a fundamental task in computer vision. Since Fully Convolutional Network (FCN) [60] treats dense semantic segmentation as an end-to-end per-pixel classification task, CNNs [61, 62] have been dominating this field. Recently, Transformers [48, 48] have gradually taken over the dominance of CNNs due to their distinguished contextual modeling capabilities.

Initially, researchers usually employed the mainstream classification backbone (*e.g.*, ResNet [63] and DenseNet [64]) instead of specific architectures to extract semantic features. However, semantic segmentation is a dense prediction task, which is different from image classification. Thus, some tailored feature extraction networks are developed, such as SETR [47], SegFormer [48], DPT [65], and SegNeXt [66], *etc.* Furthermore, researchers devised some novel decoder networks with various goals, including enlarging receptive fields [67], collecting multi-scale semantics [68], capturing global context [69], and enhancing edge features [70].

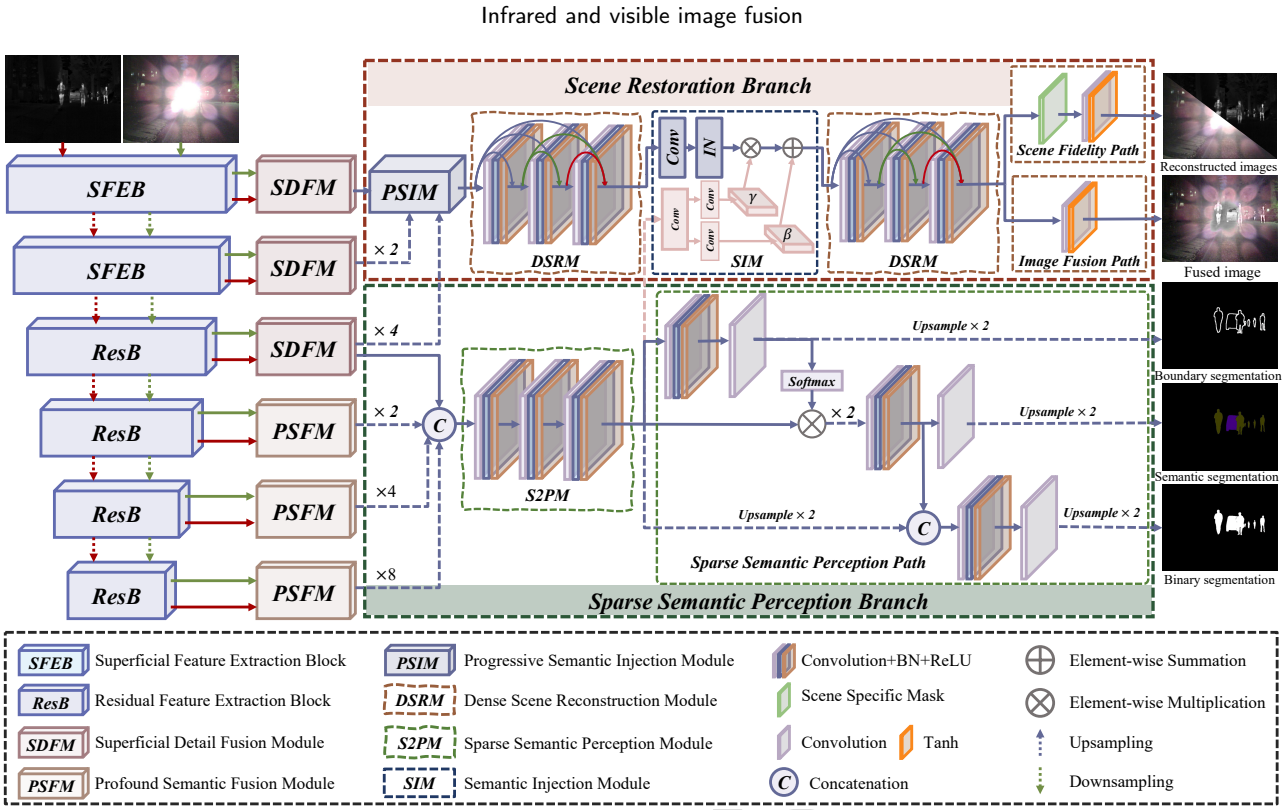


Figure 4: The overall framework of the proposed practical infrared and visible image fusion network based on progressive semantic injection and scene fidelity.

2.2.2. *Multi-modal semantic segmentation*

Although previous works have achieved remarkable segmentation performance on standard RGB-based datasets, in challenging real-world scenarios, using multi-modality sensors is desirable to provide a more comprehensive understanding of the scene [71]. Ha *et al.* released the first multi-spectral semantic segmentation dataset, named MFNet, and implemented multi-modal segmentation using two symmetric and simple encoders along with a mini-inception decoder [72]. However, the simple feature extraction networks fail to extract sufficient semantic features. Therefore, some approaches (*e.g.*, RTFNet [73], FEANet [74], and MDR-Net [22]) introduce the powerful backbones (*e.g.*, ResNet [63]) as encoders to extract more complete features. Some novel feature fusion modules, including complementary activation module [21], multi-modal fusion model [20], channel-wise weighted feature fusion module [22], and deep feature fusion module [18], are devised to achieve sufficient complementary information aggregation. Besides, Zhao *et al.* leveraged the binary [18] and edge prior [20] to provide adequate supervision for multi-modal semantic perception.

Recently, Liu *et al.* applied a more advanced backbone, *i.e.*, SegFormer [48], to accomplish the multi-modal segmentation task and achieved new state-of-the-art performance in this field [75]. It is instructive to note that applying the new backbone to multi-modal segmentation usually involves sophisticated re-designs of the feature fusion modules. The gap between the feature fusion modules and backbones means the improvements in unimodal segmentation may not appear in multi-modal segmentation immediately. Moreover, most approaches adopt two independent encoders (especially large-scale backbones) to extract complementary features from source images, which leads to a significant increase in computation. In this study, we demonstrate the potential of image-level fusion compared to feature-level fusion for the segmentation task, which provides new insights for improving the performance of high-level vision tasks in complex scenes.

3. Methodology

In this section, we present in detail our practical infrared and visible image fusion method, *i.e.*, PSFusion. Firstly, we provide an overview of our proposed method. Subsequently, we introduce the loss functions associated with image fusion and semantic perception.

3.1. Overall framework

Our method proposes for the first time to explicitly inject semantic information at the feature level to accommodate different high-level semantic model. As shown in Figure 4, the proposed method involves a scene restoration branch and a sparse semantic perception branch. More specifically, the scene restoration branch contains a scene fidelity path and an image fusion path, where two paths share the successive progressive semantic injection module (PSIM), dense scene reconstruction module (DSRM), semantic injection

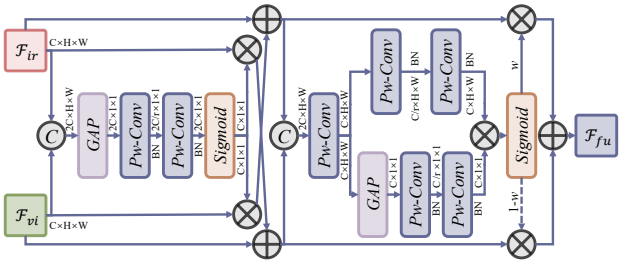


Figure 5: The architecture of the superficial detail fusion module based on the channel-spatial attention mechanism.

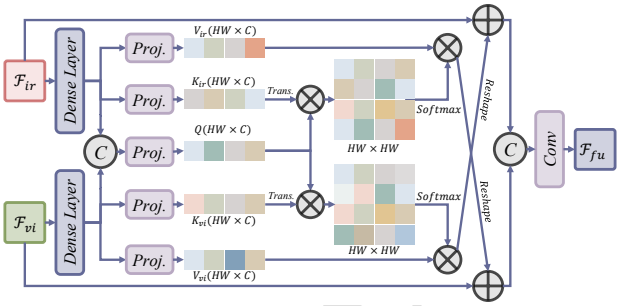


Figure 6: The architecture of the profound semantic fusion module based on the cross-attention mechanism.

module (SIM), and dense scene reconstruction module. The semantic perception branch consists of a sparse semantic perception module (S2PM) and a sparse semantic perception path (S2P2), where the S2P2 is composed of three task-specific headers for perceiving sparse semantics from different perspectives. Given a pair of registered infrared image $I_{ir} \in \mathbb{R}^{H\times W\times 1}$ and visible image $I_{vi} \in \mathbb{R}^{H\times W\times 3}$, the scene restoration branch aims to reconstruct the infrared image $\hat{I}_{ir}$ and visible image $\hat{I}_{vi}$, as well as synthesizing the fused image $I_f \in \mathbb{R}^{H\times W\times 3}$. The sparse semantic perception branch is responsible for predicting boundary segmentation result I_{bd}, semantic segmentation result I_{se}, and binary segmentation result I_{bi}.

To facilitate semantic features of the semantic perception branch being easily injected into the scene restoration branch, we expect to share the feature extract network between these two branches. Unfortunately, maintaining high-resolution features is necessary for preserving fine-grained details in the scene restoration branch, whereas high-level vision tasks require down-sampling to extract sufficient semantic features and capture the overall structure. The contradiction between these two requirements prevents us from using the existing backbone as a feature extraction network. Thus, as shown in Figure 4, we employ ResNet [63] as our basic feature extraction network, and devise two superficial feature extraction blocks (SFEB) to replace the first layer of ResNet. The feature extraction can be formulated as:

$$\left\{\mathcal{F}_{ir}^1, \mathcal{F}_{vi}^1\right\} = \left\{SFEB_{ir}^1(I_{ir}), SFEB_{vi}^1(I_{vi})\right\}, \quad (1)$$

$$\left\{\mathcal{F}_{ir}^2, \mathcal{F}_{vi}^2\right\} = \left\{SFEB_{ir}^2(\mathcal{F}_{ir}^1), SFEB_{vi}^2(\mathcal{F}_{vi}^1)\right\}, \quad (2)$$

$$\left\{\mathcal{F}_{ir}^i, \mathcal{F}_{vi}^i\right\} = \left\{ResB_{ir}^i(\mathcal{F}_{ir}^{i-1}), ResB_{vi}^i(\mathcal{F}_{vi}^{i-1})\right\}, \quad (3)$$

where $i = 3, 4, 5, 6$ and $ResB$ indicates the residual block in ResNet. In particular, when $i = 1, 2, 3$, $\mathcal{F}_{ir}^i$ and $\mathcal{F}_{vi}^i$ represent the superficial infrared and visible features, respectively. If $i = 4, 5, 6$, $\mathcal{F}_{ir}^i$ and $\mathcal{F}_{vi}^i$ represent the profound infrared and visible semantic features, respectively. Correspondingly, we develop the superficial detail fusion module and profound semantic fusion module to aggregate the complementary information in shallow and deep features, respectively.

Considering that the shallow features contain abundant details and structural information, we propose a superficial detail fusion module (SDFM) based on the channel-spatial attention mechanism to integrate the shallow features. The architecture of SDFM is illustrated in Figure 5. In particular, we concatenate the infrared and visible features in the channel dimension, and then feed them into a channel attention module that consists of convolution and pooling operations to generate attention weights. These weights are then applied to weight the original features through element-wise multiplication, and the resulting features are added to the original features from another branch to enhance their representation. The feature reinforcement process can be summarized as follows:

$$\begin{aligned}\hat{\mathcal{F}}_{ir}^i &= \mathcal{F}_{ir}^i \oplus (\mathcal{F}_{vi}^i \otimes \delta(Pw\text{-}Conv^n(GAP(\mathcal{C}(\mathcal{F}_{ir}^i, \mathcal{F}_{vi}^i))))),\\ \hat{\mathcal{F}}_{vi}^i &= \mathcal{F}_{vi}^i \oplus (\mathcal{F}_{ir}^i \otimes \delta(Pw\text{-}Conv^n(GAP(\mathcal{C}(\mathcal{F}_{ir}^i, \mathcal{F}_{vi}^i))))),\end{aligned} \quad (4)$$

where $\oplus$ refers to element-wise summation, $\otimes$ indicates element-wise multiplication, $Pw\text{-}Conv^n(\cdot)$ stands for n cascaded point-wise convolution layers, $\mathcal{C}(\cdot)$ represents the concatenation operation in the channel dimension, $\delta(\cdot)$ and $GAP(\cdot)$ denotes the sigmoid function and global average pooling, respectively. Afterward, the reinforced features are concatenated in the channel dimension and fed into the parallel channel attention and spatial attention modules to generate final fusion weights. The fusion weight generation process can be formulated as:

$$\begin{aligned}\mathcal{A}_C^i &= Pw\text{-}Conv^n(GAP(Pw\text{-}Conv(\mathcal{C}(\hat{\mathcal{F}}_{ir}^i, \hat{\mathcal{F}}_{vi}^i)))),\\ \mathcal{A}_S^i &= Pw\text{-}Conv^n(Pw\text{-}Conv(\mathcal{C}(\hat{\mathcal{F}}_{ir}^i, \hat{\mathcal{F}}_{vi}^i))),\\ \mathcal{W}^i &= \delta(\mathcal{A}_C^i \otimes \mathcal{A}_S^i).\end{aligned} \quad (5)$$

Since the infrared and visible features are complementary, we can use the generated weights for one of the modalities, and the fusion weights of another modality can be expressed as $1 - \mathcal{W}^i$. Therefore, the fusion process of superficial features can be formulated as:

$$\mathcal{F}_{fu}^i = (\mathcal{W}^i \otimes \hat{\mathcal{F}}_{ir}^i) \oplus ((1 - \mathcal{W}^i) \otimes \hat{\mathcal{F}}_{vi}^i). \quad (6)$$

Given that high-level vision tasks often require abundant contextual information for a comprehensive understanding, we develop a profound semantic fusion module (PSFM) based on cross-attention to integrate deep features, as shown in Figure 6. The PSFM first employs the dense layers to enhance the features extracted by the backbone and output the enhanced deep features $\hat{\mathcal{F}}_{ir}^i$ and $\hat{\mathcal{F}}_{vi}^i$. Then, a projection

function, which comprises convolution and reshape operations, is deployed to convert the enhanced features into key and value, as presented in the following:

$$\begin{aligned} K_x^i &= \text{Reshape}(Conv_K^x(\hat{\mathcal{F}}_x^i)), \\ V_x^i &= \text{Reshape}(Conv_V^x(\hat{\mathcal{F}}_x^i)), \end{aligned} \tag{7}$$

where $x \in \{ir, vi\}$ indicates modality, $K_x^i \in \mathbb{R}^{H_i W_i \times C_i}$ stands for key, and $V_x^i \in \mathbb{R}^{H_i W_i \times C_i}$ represents value. $Conv(\cdot)$ and Reshape$(\cdot)$ correspond to the convolutional layer with 3×3 kernel size and the reshape operation, respectively. H_i, W_i, and C_i refer to the high, width, and channel of the input features $\hat{\mathcal{F}}_x^i$, respectively. It should be noted that both infrared and visible features are incorporated to generate our modality-invariant query, as shown in Eq. (8), which allows us to completely exploit complementary properties in the multi-modal features:

$$Q^i = \text{Reshape}(Conv(\mathcal{C}(\hat{\mathcal{F}}_{ir}^i, \hat{\mathcal{F}}_{vi}^i))), \tag{8}$$

where $Q^i \in \mathbb{R}^{H_i W_i \times C_i}$. Then, we calculate the modality-specific attention map $\mathcal{A}_x \in \mathbb{R}^{H_i W_i \times H_i W_i}$ for each modality x according to:

$$\mathcal{A}_x^i = \text{Softmax}(Q^i {K_x^i}^T). \tag{9}$$

Subsequently, the value is multiplied by the attention to obtain the features with the global context. Similar to the SDFM, we add global features to the original features of another branch, and concatenate the resulting features along the channel dimension. Finally, we feed the concatenated features into a convolutional layer to obtain fusion features. This process can be formulated as:

$$\begin{aligned} \mathcal{F}_{fu}^i = Conv(\mathcal{C}(&\mathcal{F}_{vi}^i \oplus \text{Reshape}(\mathcal{A}_{ir}^i V_{ir}^i), \\ &\mathcal{F}_{ir}^i \oplus \text{Reshape}(\mathcal{A}_{vi}^i V_{vi}^i)). \end{aligned} \tag{10}$$

Next, we will first introduce the semantic perception branch and then further present the scene restoration branch, since the scene restoration branch needs to absorb the semantic features from the semantic perception branch. More specifically, the superficial features contain a significant amount of low-level information, *i.e.*, detailed information, which may negatively impact the performance of high-level vision tasks [66]. Therefore, our sparse semantic perception branch only utilizes the deep features and the last shallow features to predict the boundary, semantic, and binary segmentation results. These features first undergo convolution and up-sampling operations before being concatenated in the channel dimension, as expressed in the following:

$$\begin{aligned} \mathcal{F}_{se}^{init} = \mathcal{C}(&Conv(\mathcal{F}_{fu}^3), \uparrow^2(Conv(\mathcal{F}_{fu}^4)), \\ &\uparrow^4(Conv(\mathcal{F}_{fu}^5)), \uparrow^8(Conv(\mathcal{F}_{fu}^6))), \end{aligned} \tag{11}$$

where $\uparrow^n(\cdot)$ denotes up-sampling n times. The concatenated features, as the initial semantic features ($\mathcal{F}_{se}^{init}$), are fed into the sparse semantic perception module (S2PM), which consists of successive convolution blocks. Each convolution block includes a convolution layer with 3×3 kernel size, batch normalization, and ReLU activation function. Next, we devise a sparse semantic perception path inspired by GMNet [18] to predict boundary, semantic, and binary segmentation results. The sparse semantic perception path can be formulated as follows:

$$\begin{aligned} \mathcal{F}_{bd} &= Conv(Conv_{BN}(\mathcal{F}_{se})), \\ I_{bd} &= \uparrow^4(\mathcal{F}_{bd}), \end{aligned} \tag{12}$$

$$\begin{aligned} \hat{\mathcal{F}}_{se} &= Conv_{BN}(\uparrow^2(\text{Softmax}(\mathcal{F}_{bd})) \otimes \mathcal{F}_{se}), \\ I_{se} &= \uparrow^2(Conv(\hat{\mathcal{F}}_{se})), \end{aligned} \tag{13}$$

$$I_{bi} = \uparrow^2(Conv(Conv_{BN}(\mathcal{C}(\hat{\mathcal{F}}_{se}, \uparrow^2(\mathcal{F}_{se}))))), \tag{14}$$

where $\mathcal{F}_{se}$ refers to the output features of the S2FM, and $Conv_{BN}(\cdot)$ indicates a convolution block consisting of a convolution layer with 3×3 kernel size, batch normalization, and ReLU activation function.

It is instructive to note that we expect the scene restoration branch can fully exploit the deep semantic information generated by the semantic perception branch. To achieve this goal, we design the progressive semantic injection module (PSIM) inspired by Zhang *et al.* [76], which consists of multiple semantic injection modules (SIMs), to progressively inject the two latter shallow features into the first superficial features. Specifically, we first inject semantic information from the third shallow features into the second features, and then inject semantic information from the second features into the first superficial features. Given two neighboring features, denoted by $\mathcal{F}_{fu}^i$ and $\mathcal{F}_{fu}^{i+1}$ $(i = 1, 2)$, the SIM can be formulated as:

$$\begin{aligned} \gamma &= Conv_\gamma^n(\uparrow^2(\mathcal{F}_{fu}^{i+1})), \\ \beta &= Conv_\beta^n(\uparrow^2(\mathcal{F}_{fu}^{i+1})), \\ \hat{\mathcal{F}}_{fu}^i &= (\gamma \otimes \text{IN}(Conv_{BN}(\mathcal{F}_{fu}^i))) \oplus \beta, \end{aligned} \tag{15}$$

where IN$(\cdot)$ denotes non-parametric instance normalization.

The first superficial features, which absorb abundant semantic information from other features, are taken as the initial detail features ($\mathcal{F}_{sr}$) of the scene restoration branch. $\mathcal{F}_{sr}$ are fed into a dense scene reconstruction module (DSRM), which consists of convolution blocks and dense connections, to enhance fine-grained details. Subsequently, we inject the semantic features ($\mathcal{F}_{se}$) generated by the S2PM into the scene reconstruction features via the SIM. Another DSRM is deployed to reinforce the fine-grained features and generate the final scene reconstruction features ($\hat{\mathcal{F}}_{sr}$). Finally, we synthesize the fused image I_f using an image fusion path consisting of a convolution layer with 3×3 kernel size and a Tanh activation function. It is worth emphasizing that we also devise a scene fidelity path (SFP) that consists of the modality-specific mask, convolutional layer, and Tanh activation function to reconstruct source images from $\hat{\mathcal{F}}_{sr}$. Thus, the SFP can constrain $\hat{\mathcal{F}}_{sr}$ to contain complete information for reconstructing infrared images $\hat{I}_{ir}$ and visible images $\hat{I}_{vi}$. With the dual constraint of the sparse semantic perception path and scene fidelity

path, fusion results generated by the image fusion path can contain both sufficient semantic information and complete scene information for a comprehensive understanding of the imaging scene.

3.2. Loss function

Our PSFusion not only uses the fusion loss to directly constrain the fusion results, but also leverages the scene fidelity path and sparse semantic perception path to indirectly restrain the feature extraction and aggregation of the fusion network. Next, we describe the fusion loss, scene fidelity loss, and sparse semantic prediction loss in sequence.

3.2.1. Fusion loss

We introduce the intensity loss $\mathcal{L}_{int}$, texture loss $\mathcal{L}_{text}$, and correlation loss $\mathcal{L}_{corr}$ to constrain the visual quality of the fusion results. As shown in Figure 1, SeAFusion [3] fails to fully present the advantages of image fusion in some extreme circumstances due to it only uses the maximum selection strategy to construct the intensity loss. Thus, we rethink the intensity loss in the following two aspects. On the one hand, the high-contrast properties of infrared images characterized by variance are expected to be preserved. We therefore generate a contrast mask by comparing the variance of infrared and visible images to guide the fusion network to adaptively preserve the high-contrast regions in source images. The contrast mask of infrared or visible images can be expressed as:

$$\mathcal{M}_x = \begin{cases} 1, & \text{if } \sigma^2(x) > \sigma^2(y), \\ 0, & \text{otherwise}, \end{cases} \tag{16}$$

where x denotes the infrared or visible modality, and y denotes another modality, $\sigma^2(x) \in \mathbb{R}^{H\times W\times 1}$ represents the variance of image x.

On the other hand, following STDFusionNet [12], we employ the salient target mask $\mathcal{M}_{tar}$ to guide the fusion network to preserve the significant targets in the infrared images. Specifically, it is easy to generate the salient target masks from the semantic segmentation labels. Thus, the intensity loss $\mathcal{L}_{int}$ can be expressed as:

$$\begin{aligned} \mathcal{L}_{int} =& \frac{1}{HW}\left\|\delta(\mathcal{M}_{ir}+\mathcal{M}_{tar})\otimes(I_f - I_{ir})\right\|_1 + \\ & \frac{1}{HW}\left\|\delta(\mathcal{M}_{vi}+(1-\mathcal{M}_{tar}))\otimes(I_f - I_{vi})\right\|_1, \end{aligned} \tag{17}$$

where $\|\cdot\|_1$ stands for the l_1-norm, and $\delta(\cdot)$ indicates the sigmoid function, which is utilized to limit the range of the two combined masks to $[0, 1]$.

In addition, we also deploy the texture loss $\mathcal{L}_{text}$ to force the fused images to contain abundant texture information, which is shown in the following:

$$\mathcal{L}_{text} = \frac{1}{HW}\left\||\nabla I_f| - \max(|\nabla I_{ir}|, |\nabla I_{vi}|)\right\|_1, \tag{18}$$

where ∇ refers to the Sobel gradient operator, $|\cdot|$ denotes the absolute operation, and $\max(\cdot)$ stands for the element-wise maximum selection. Furthermore, we introduce a regularization term $\mathcal{L}_{corr}$ to strengthen the correlation between the fused images and source images, as expressed:

$$\mathcal{L}_{corr} = \frac{1}{corr(I_f, I_{ir}) + corr(I_f, I_{vi})}, \tag{19}$$

in which $corr(\cdot)$ means the calculation of the correlation of two images. Finally, the fusion loss function can be formulated as follows:

$$\mathcal{L}_f = \mathcal{L}_{int} + \mathcal{L}_{text} + \alpha \cdot \mathcal{L}_{corr}, \tag{20}$$

where α is used to strike a balance between different loss functions. With the united constraints of $\mathcal{L}_{int}$, $\mathcal{L}_{text}$, and $\mathcal{L}_{corr}$, our fusion results can provide abundant scene descriptions and favorable visual perception.

3.2.2. Auxiliary loss

We devise a scene fidelity path to constrain the fusion features to contain complete information for reconstructing the source images. Correspondingly, we also design the scene fidelity loss to ensure the scene fidelity path to accomplish this goal. Similar to the fusion loss, the scene fidelity loss also consists of an intensity term and a texture term, as shown in the following:

$$\mathcal{L}_{sf}^{x} = \frac{1}{HW}\left\|I_x - \hat{I}_x\right\|_1 + \frac{1}{HW}\left\||\nabla I_x| - |\nabla \hat{I}_x|\right\|_1. \tag{21}$$

Moreover, we introduce the sparse semantic perception path to ensure the features $\mathcal{F}_{se}$ output by the S2PM contain sufficient semantic information that will be injected into the image fusion network. Following GMNet [18], we design the semantic loss $\mathcal{L}_{se}$, binary segmentation loss $\mathcal{L}_{bi}$, and boundary segmentation loss $\mathcal{L}_{bd}$ to force the semantic features $\mathcal{F}_{se}$ can be efficiently used for perceiving the imaging scene from different perspectives. In particular, the commonly used cross-entropy loss function between predicted boundary result I_{bd} and its ground truth is utilized to construct the boundary segmentation loss. Considering the class imbalance between objects and backgrounds, we utilize the weighted cross-entropy loss to calculate the binary segmentation loss. Moreover, we adopt the OHEMCELoss [77] to calculate the semantic loss, which can alleviate the challenges caused by the hard examples.

Finally, the full objective function of our PSFusion is a weighted sum of the aforementioned loss terms, which is formulated as:

$$\begin{aligned} \mathcal{L}_{\text{PSF}} =& \lambda_f \cdot \mathcal{L}_f + \lambda_{sf} \cdot (\mathcal{L}_{sf}^{vi} + \mathcal{L}_{sf}^{ir}) + \\ & \lambda_{se} \cdot (\mathcal{L}_{se} + \mathcal{L}_{bd} + \mathcal{L}_{bi}), \end{aligned} \tag{22}$$

where λ_f, λ_{sf}, and λ_{se} are the hyper-parameters that control the trade-off of the fusion-related loss, scene fidelity-related losses, and semantic-related losses.

4. Experimental validation

In this section, we first provide some experimental configurations and implementation details. We then compare

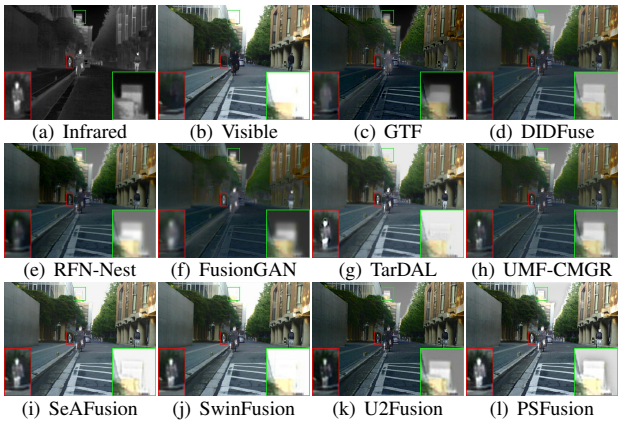


Figure 7: Qualitative comparison of PSFusion with 9 state-of-the-art methods on the 00332D scene from the MSRS dataset.

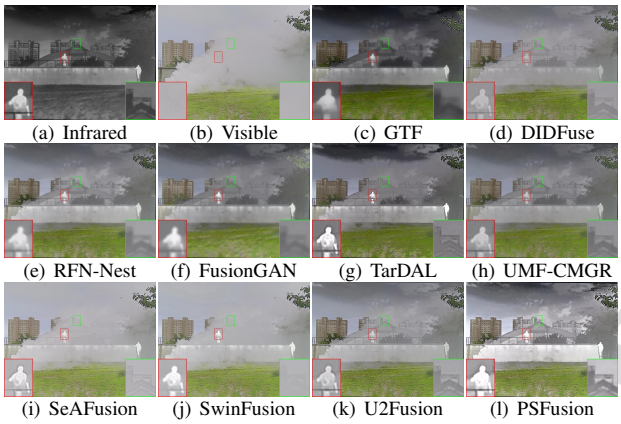


Figure 8: Qualitative comparison of PSFusion with 9 state-of-the-art methods on the 00922 scene from the M^{3}FD dataset.

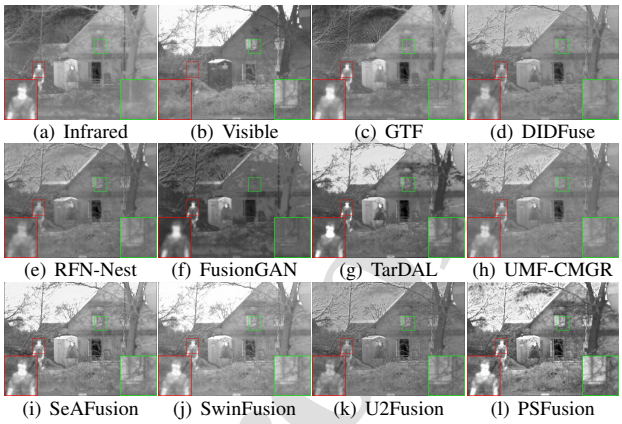


Figure 9: Qualitative comparison of PSFusion with 9 state-of-the-art methods on the meeting scene from the TNO dataset.

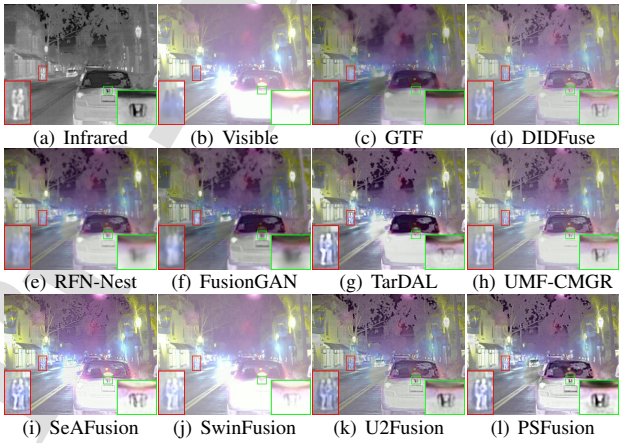


Figure 10: Qualitative comparison of PSFusion with 9 state-of-the-art methods on the 080 scene from the RoadScene dataset.

the fusion performance of various algorithms on several datasets from both qualitative and quantitative perspectives. Furthermore, we deploy different segmentation models to verify the advantages and generalization ability of our PSFusion for high-level vision tasks. After that, we fully examine the potential of image-level fusion in comparison to feature-level fusion for high-level vision tasks. Finally, we conduct ablation studies to demonstrate the effectiveness of the specific designs.

4.1. Configurations and implementation details

We train our model on the MSRS dataset [3] and completely validate the fusion performance of PSFusion on the MSRS [3], M^{3}FD [16], TNO [78], and RoadScene [58] datasets. The comparison fusion algorithms include a traditional approach, *i.e.*, GTF [29], two pre-trained approaches, namely DIDFusion [34] and RFN-Nest [33], and six end-to-end approaches. The latter involves two GAN-based methods, *i.e.*, FusionGAN [14] and TarDAL [16], three CNN-based methods, *i.e.*, UMF-CMGR [40], SeAFusion [3], and U2Fusion [58], and one Transformer-based method, *i.e.*, SwinFusion [15]. SeAFusion and TarDAL are semantic-driven methods, while SwinFusion and U2Fusion are general fusion algorithms. Furthermore, we select three semantic segmentation models, including BANet [79], SegFormer [48], and SegNeXt [66], to measure semantic information contained in the fusion results on the MSRS dataset. Additionally, we compare the performance of our method and the feature-level fusion algorithm for high-level vision tasks on a more challenging dataset, *i.e.*, the MFNet dataset [72]. Feature-level fusion-based multi-modal segmentation algorithms involve MFNet [72], RTFNet [73], GMNet [18], FEANet [74], EGFNet [20], LASNet [21], MDRNet [22], and CMX [75].

Six statistical evaluation metrics are utilized to quantitatively evaluate fusion performance, including entropy (EN) [80], standard deviation (SD) [81], average gradient (AG) [82], spatial frequency (SF) [83], sum of correlation differences (SCD) [84], visual information fidelity (VIF) [85]. A higher value of these metrics indicates better fusion performance. In addition, the pixel intersection over union (IoU) is used to quantify the segmentation performance. Both the MSRS and MFNet datasets involve nine categories of objects, *i.e.*, car, person, bike, curve, car stop, guardrail, color tone, and background.

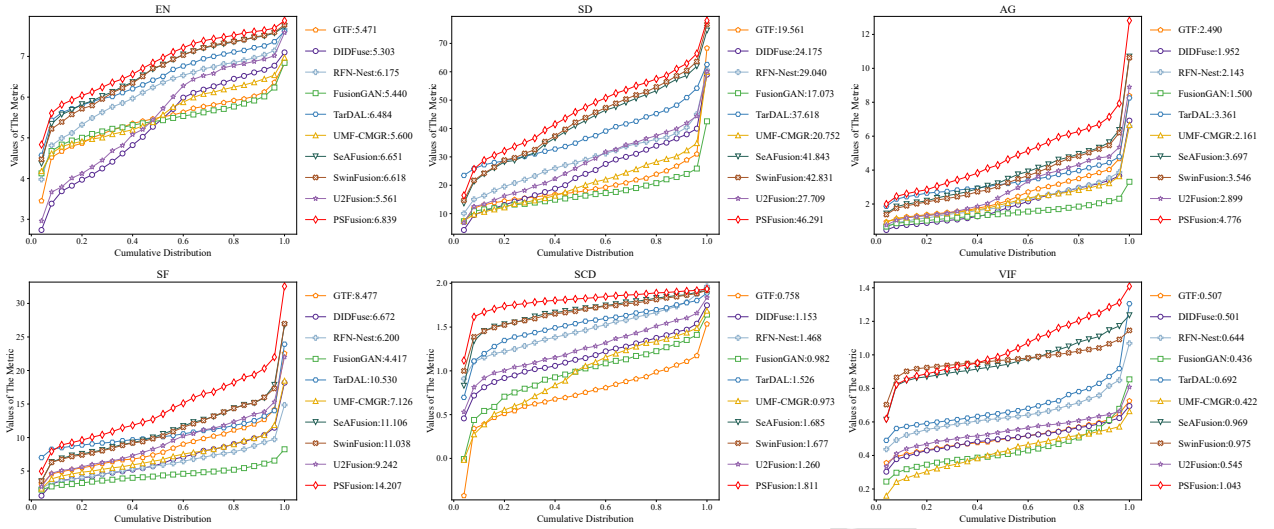


Figure 11: Quantitative comparisons of the six metrics on 361 image pairs from the MSRS dataset. A point (x, y) on the curve denotes that there are $(100 * x)\%$ percent of image pairs which have metric values no more than y.

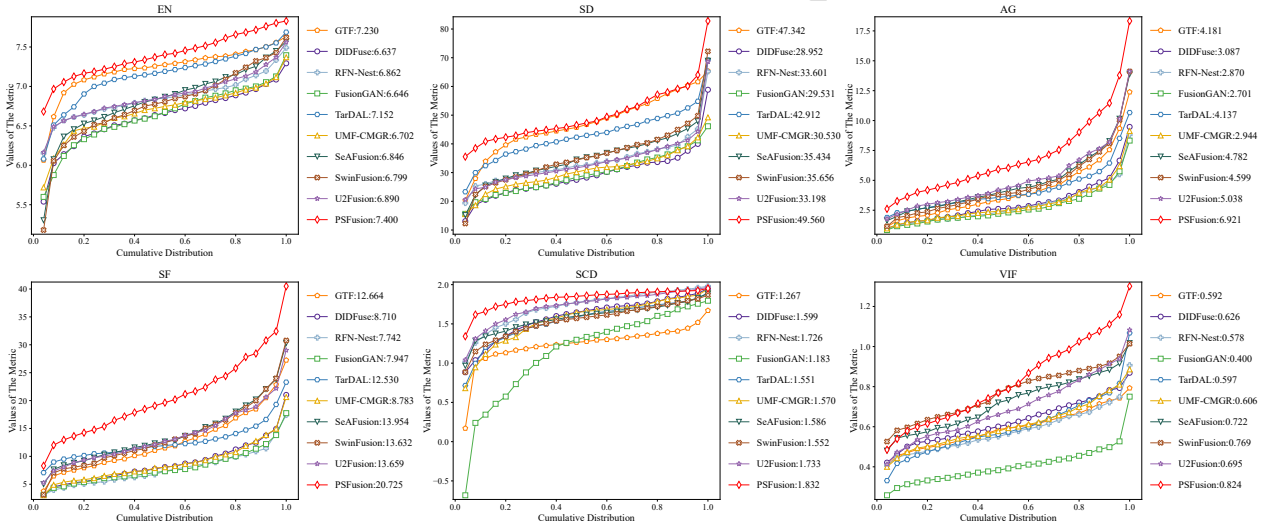


Figure 12: Quantitative comparisons of PSFusion with 9 methods on 300 image pairs from the M^3FD dataset.

We jointly train the scene restoration branch and semantic perception branch in a single framework. The hyper-parameters controlling the trade-off of various loss terms are empirically set as $\lambda_f = 10, \lambda_{sf} = 5, \lambda_{se} = 10$, and $\alpha = 0.1$. Stochastic Gradient Descent (SGD) is applied to train our model, and the batch size is set to 16. We set the initial learning rate as 0.001 and employ the poly-learning rate decay policy. We train our model for $2,500$ epochs, which allows us to completely excavate semantic information. All images are normalized to $[0, 1]$ before being fed into networks. As suggested by SeAFusion [3], we process color information in the YCbCr color space. Our PSFusion is implemented on the PyTorch platform [86]. All comparison algorithms and the segmentation model are implemented following their original papers. All experiments are conducted on the NVIDIA GeForce RTX 3090 and 2.90GHz Intel(R) Xeon(R) Platinum 8375C CPU.

4.2. Fusion comparison and analysis

4.2.1. Qualitative comparison and analysis

The visual results of various fusion algorithms on the MSRS, M^3FD, TNO, and Roadscene datasets are presented in Figures 7, 8, 9, and 10, respectively. From Figure 7, it can be observed that TarDAL, SeAFusion, and SwinFusion fail to accurately present the rooftops on the street due to overexposure in the visible image, which results in insufficient information being captured. Although other methods can present the rooftops, they inevitably weaken salient targets (*e.g.*, the pedestrian in the red box), especially GTF, RFN-Nest, and FusionGAN. In contrast, our method is

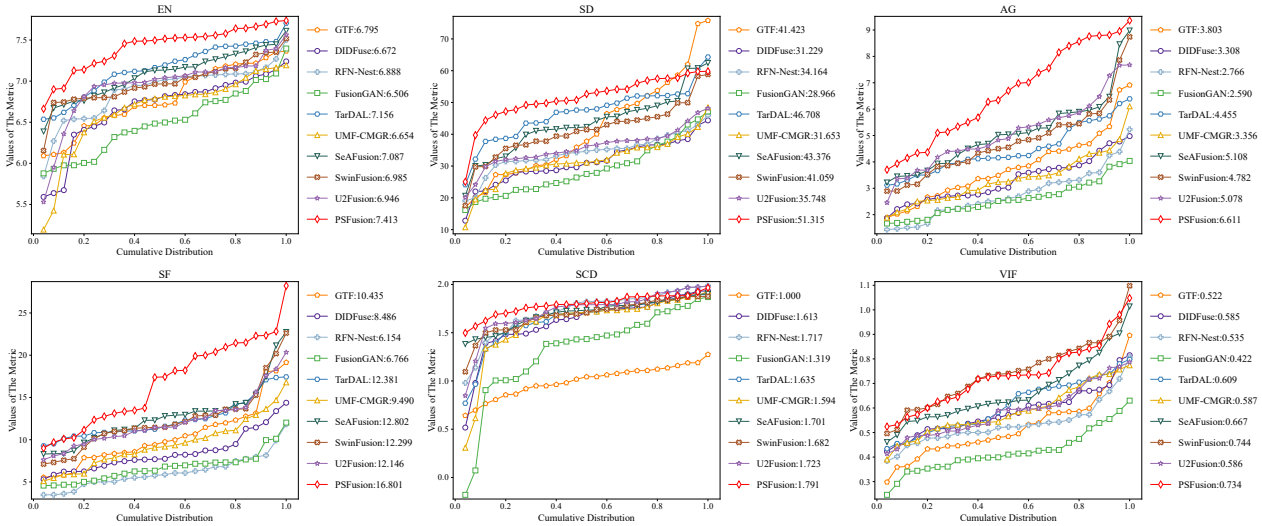


Figure 13: Quantitative comparisons on 25 image pairs from the TNO dataset.

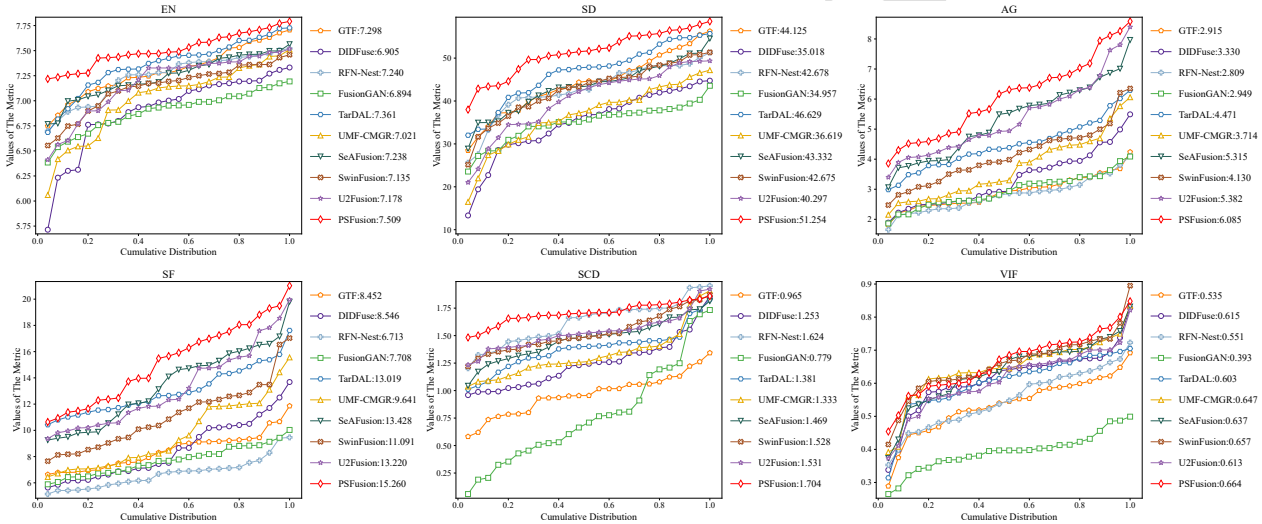


Figure 14: Quantitative comparisons on 25 image pairs from the RoadScene dataset.

able to fully integrate complementary information from the source images to provide a more comprehensive description of the imaging scene. Similarly, in the smoke-filled surrounding, most fusion algorithms cannot clearly depict the building hidden in the smoke, although they can preserve the prominent targets in the infrared image. As shown in Figure 8, the fusion results synthesized by our PSFusion not only highlight the prominent pedestrians, but also effectively display the details of the buildings. This advantage can be attributed to two aspects. On the one hand, we design more refined loss functions based on the contrast mask and significant target mask to maintain the visual appeal of fused images. On the other hand, we introduce a scene fidelity path to indirectly constrain the fusion results to contain as much complementary information as possible in source images.

Additionally, the visual results on the TNO and RoadScene demonstrate the superiority of our method in terms of fusion performance. From Figure 9, one can notice that our PSFusion can maintain both the prominent soldiers in the infrared image and the texture details of the visible images, such as the tree branches by the window. As presented in Figure 10, only our methods can clearly render both the pedestrians on the sides of the road and the marking of vehicles concurrently in the street environments. Extensive qualitative comparisons and analysis sufficiently demonstrate the superior fusion performance of our PSFusion from the perspective of visual perception. Particularly, the proposed method can effectively cope with harsh environments, such as nighttime, fog, occlusion, and overexposure.

4.2.2. Quantitative comparison and analysis

The quantitative comparison results are illustrated in Figures 11-14. For the datasets with separate training and test sets, *i.e.*, the MSRS and M^3FD datasets, we evaluate the

Table 1
Per-class segmentation results on the MSRS dataset. UMF., DIDF., RFN., SwinF., U2F., and SeAF. denote UMF-CMGR, DIDFuse, RFN-Nest, SwinFusion, U2Fusion, and SeAFusion, respectively. The best result is highlighted in **Red**, whereas the second best one is highlighted in *Blue*.

		Infrared	Visible	GTF	FusionGAN	UMF.	DIDF.	RFN.	TarDAL	SwinF.	U2F.	SeAF.	Ours
Background	BANet	98.24	98.26	98.44	98.49	98.48	98.29	98.50	98.50	98.49	98.49	*98.60*	**98.69**
	SegFormer	98.35	98.41	98.51	98.57	98.59	98.44	98.65	98.55	98.64	98.63	*98.66*	**98.72**
	SegNeXt	98.44	98.47	98.56	98.67	98.66	98.51	98.67	98.65	98.66	98.67	*98.70*	**98.73**
Car	BANet	87.33	89.03	89.12	89.86	90.24	88.83	89.95	90.04	90.20	89.78	*90.38*	**91.66**
	SegFormer	88.85	90.76	89.37	89.36	90.16	89.16	90.83	89.10	91.05	*91.08*	91.01	**91.55**
	SegNeXt	89.97	91.18	90.74	91.57	91.58	90.65	91.63	91.69	**91.89**	*91.81*	**91.89**	*91.81*
Person	BANet	70.46	59.94	71.76	72.83	72.20	69.92	72.03	73.17	72.24	72.93	*74.56*	**76.81**
	SegFormer	72.28	65.54	73.33	*75.37*	74.54	73.04	74.96	75.31	75.11	75.22	75.14	**76.43**
	SegNeXt	72.47	65.76	73.76	75.66	75.19	73.14	75.25	75.55	75.32	75.33	*75.93*	**77.66**
Bike	BANet	69.23	70.00	72.04	71.95	71.20	67.82	71.39	71.13	70.31	70.99	*72.09*	**72.93**
	SegFormer	70.28	71.48	72.20	72.23	72.44	69.57	72.31	71.62	72.30	72.28	*72.47*	**73.07**
	SegNeXt	70.75	72.33	73.24	73.32	73.16	71.03	*73.50*	72.80	72.95	72.68	72.12	**73.64**
Curve	BANet	58.74	60.69	64.47	63.45	62.71	56.85	62.04	61.92	60.53	62.84	**65.26**	*64.48*
	SegFormer	59.15	59.70	62.72	63.93	64.17	57.26	**64.88**	63.84	62.32	64.12	*64.50*	63.76
	SegNeXt	60.49	61.42	62.17	**64.93**	63.25	59.23	63.80	*64.59*	62.55	64.39	64.48	62.76
Car Stop	BANet	68.85	71.43	70.59	71.68	70.53	66.11	**74.92**	72.94	70.85	72.13	*74.38*	74.32
	SegFormer	70.02	77.49	76.67	78.83	77.82	75.37	78.76	*79.24*	77.56	78.34	79.12	**80.81**
	SegNeXt	74.93	78.24	77.67	79.27	79.30	76.91	78.65	78.89	78.23	78.01	*79.91*	**80.15**
Cuardrail	BANet	65.57	77.90	68.95	79.45	77.96	82.30	74.87	73.08	**86.06**	79.25	*84.34*	80.60
	SegFormer	65.51	83.52	**85.68**	73.06	84.31	80.67	80.80	80.97	78.06	84.04	*85.04*	83.82
	SegNeXt	76.48	81.26	81.76	83.13	*86.52*	80.34	79.98	78.88	83.40	80.28	81.48	**88.92**
Color cone	BANet	56.93	63.42	63.71	64.35	64.65	59.54	63.41	64.77	63.88	63.59	*66.21*	**68.47**
	SegFormer	56.27	65.27	57.36	63.56	62.27	59.10	63.82	64.24	63.14	65.28	*67.63*	**68.21**
	SegNeXt	59.71	62.40	60.40	64.03	66.00	61.29	64.18	67.02	65.74	67.42	*67.75*	**70.17**
Bump	BANet	72.72	75.31	74.40	75.36	76.74	76.11	79.53	75.67	*80.36*	77.12	78.35	**81.01**
	SegFormer	74.70	78.38	74.88	78.54	76.69	74.37	79.99	77.19	**81.12**	76.39	*80.54*	80.28
	SegNeXt	74.26	80.00	74.09	76.47	78.38	75.55	77.07	75.17	76.72	*80.39*	**81.39**	79.58
mIoU	BANet	72.01	74.00	74.83	76.38	76.08	73.97	76.29	75.69	76.99	76.35	*78.24*	**78.77**
	SegFormer	72.82	76.73	76.75	77.05	77.89	75.22	78.33	77.78	77.70	78.38	*79.34*	**79.63**
	SegNeXt	75.28	76.79	76.93	78.56	79.12	76.30	78.08	78.14	78.38	78.78	*79.29*	**80.38**

metrics of different algorithms on the test set. For datasets without a predefined split, such as the TNO and RoadScene datasets, we follow the configurations of SeAFusion [3] and randomly select 25 pairs of images to report the quantitative comparison results. From the results, we can observe that our PSFusion exhibits significant advantages in all metrics on the four datasets. Only on the TNO dataset, our method lags slightly behind SwinFusion in the VIF metric.

Our method obtains the best EN implying that the fusion results generated by PSFusion contain sufficient and abundant information. The best result in the SCD metric means that our fused images have the highest correlation with the source images. This advantage is attributed to our scene fidelity path, which constrains the fusion network to retain complete information for reconstructing the source images. Furthermore, our PSFusion ranks first in the AG and SF metrics, indicating that our fusion results contain abundant texture details, which is consistent with the qualitative analysis. Additionally, our method achieves the highest scores in the SD metric, implying that our fusion results have the best contrast. This advantage is attributed to the fact that we introduce the contrast mask to construct the intensity loss. Finally, the best VIF indicates that our fused images are more compatible with the human visual system. In summary, both qualitative and quantitative comparisons demonstrate the excellent fusion performance of our proposed method.

4.3. Segmentation comparison and analysis

4.3.1. Quantitative comparison and analysis

Following SeAFusion [3], we evaluate the semantic performance of different approaches on the MSRS dataset [37] from the perspective of semantic segmentation. In addition to BANet [79], adopted by SeAFusion, we introduce more sophisticated segmentation models, *i.e.*, SegFormer [48] and SegNeXt [66], to reflect the semantic performance of diverse fusion algorithms. Table 1 shows the quantitative segmentation results, where we select the SegFormer-B2 and SegNeXt-Base models for evaluation.

Regardless of which segmentation is adopted, our fusion results achieve the highest IoU for the significant categories, such as car, person, and bike. Besides, our method also obtains impressive scores in other categories. As a result, all three segmentation models achieve the highest mean IoU (mIoU) on our fused images. We attribute this advan-

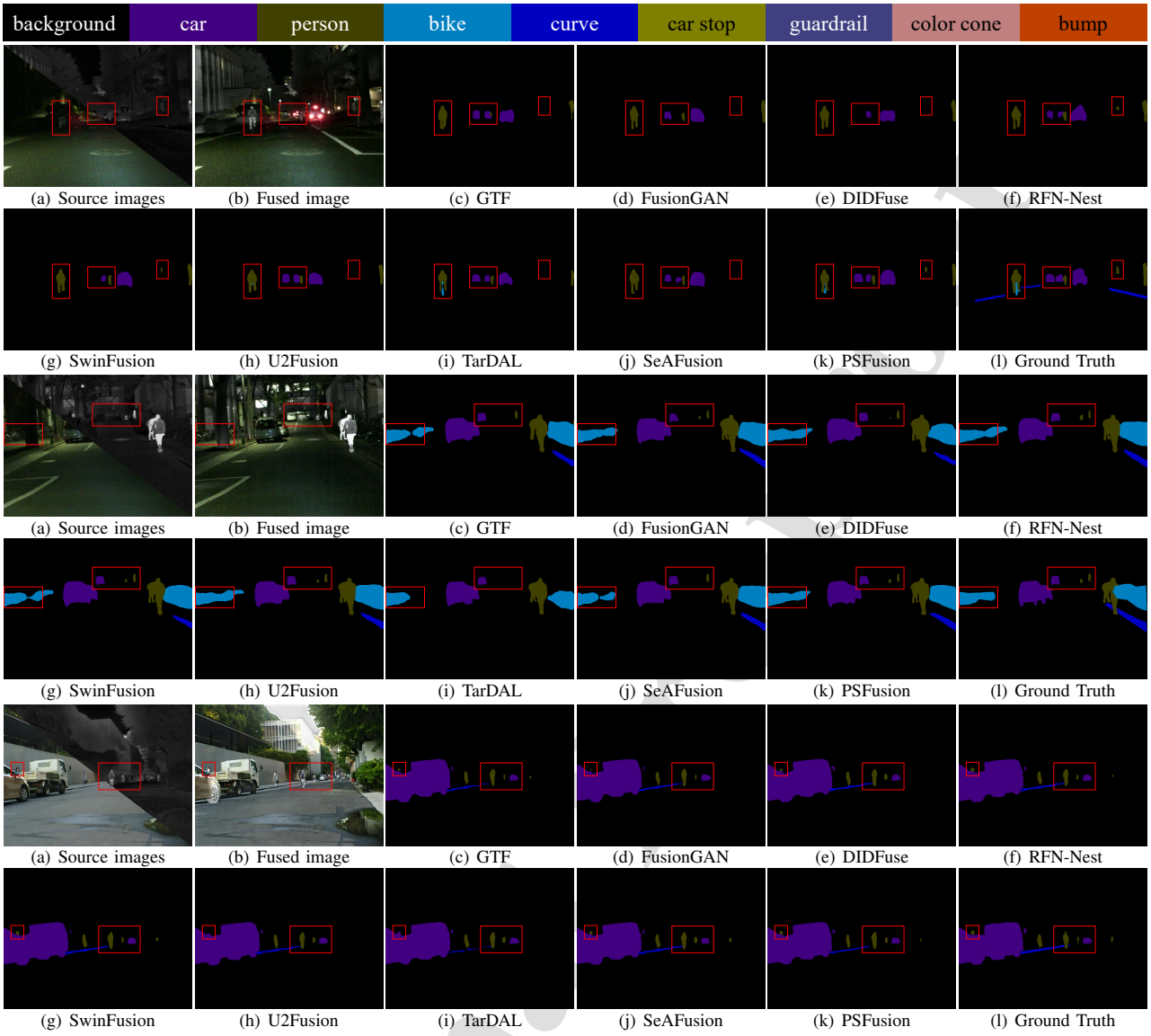


Figure 15: Segmentation results of various fusion algorithms on the MSRS dataset. Every two rows represent a scene, and from top to bottom, the segmentation models utilized are BANet, SegFormer, and SegNeXt, respectively.

tage to two factors. On the one hand, we deploy multiple semantic perception tasks, including boundary segmentation, semantic segmentation, and binary segmentation, to collaboratively constrain the extraction of semantic features. We also design the semantic injection module to explicitly inject semantic information into the fusion network. Thus, the fusion results contain abundant semantic information while presenting impressive visual appeal. On the other hand, we propose for the first time to inject semantic information into the fusion network at the feature level, which makes our fusion results more compatible with most high-level semantic models. In addition, SeAFusion achieves suboptimal scores on most categories and mIoU, indicating that the image-level semantic constraints may not be sufficient to retain enough semantic information while failing to generalize to other models effectively. Another semantics-driven approach, *i.e.*, TarDAL, does not show superior performance on the semantic segmentation as it uses the object detection task to guide image fusion.

4.3.2. Qualitative comparison and analysis

The visualization segmentation results are presented in Figure 15. In the first scene, BANet only recognizes the bike on the fused images generated by our method and TarDAL. Unfortunately, TarDAL does not provide enough information about the pedestrian hidden in the bushes. Additionally, the segmentation model fails to correctly segment distant vehicles and pedestrians from the fusion results synthesized by most algorithms, except for RFN-Nest, U2Fusion, TarDAL, SeAFusion, and our PSFusion. In the second

Table 2
Per-class segmentation results of image-level fusion and feature-level fusion on the MFNet dataset.

	Backbone	Background	Car	Person	Bike	Curve	Car Stop	Cuardrail	Color cone	Bump	mIoU
Infrared	SegFormer-B2	97.69	83.43	70.51	53.70	38.95	16.58	5.61	31.80	53.56	50.20
	SegFormer-B4	97.76	84.84	70.67	55.05	41.79	18.48	6.87	42.03	50.03	51.95
	SegNeXt-Base	97.79	84.89	70.73	56.29	41.94	24.15	7.60	35.91	48.64	51.99
	SegNeXt-Large	97.80	86.31	70.91	58.59	40.54	24.13	7.91	42.00	48.99	53.02
Visible	SegFormer-B2	97.84	87.19	61.48	62.83	31.53	28.42	4.40	50.51	45.37	52.17
	SegFormer-B4	97.98	88.33	63.36	61.41	33.82	34.69	9.36	51.10	53.92	54.89
	SegNeXt-Base	97.93	88.29	62.42	63.67	35.34	36.95	5.77	51.20	47.74	54.37
	SegNeXt-Large	97.94	87.93	62.80	65.15	34.72	41.39	10.51	53.58	46.42	55.60
MFNet$_{17}$	Customize	96.26	60.95	53.44	43.14	22.94	9.44	0.00	18.80	23.47	36.49
GMNet$_{21}$	ResNet-50	98.00	86.46	73.05	61.72	43.96	42.25	*14.52*	48.70	47.42	57.34
MDRNet$_{23}$	ResNet-50	97.90	87.07	69.81	60.87	**47.80**	34.18	8.21	50.18	54.98	56.78
RTFNet$_{19}$	ResNet-152	98.00	87.41	70.30	62.74	45.33	29.80	0.00	29.07	55.71	53.15
FEANet$_{21}$	ResNet-152	98.00	87.41	70.30	62.74	45.33	29.80	0.00	29.07	48.95	55.28
EGFNet$_{22}$	ResNet-152	98.01	87.84	71.12	61.08	46.48	22.10	6.64	**55.35**	47.12	54.76
LASNet$_{23}$	ResNet-152	97.67	87.57	69.80	58.80	42.76	33.76	7.02	48.32	40.10	54.91
CMX-B2$_{23}$	SegFormer-B2	97.39	84.23	67.12	56.93	41.11	39.56	**18.94**	48.84	54.42	58.31
CMX-B4$_{23}$	SegFormer-B4	**98.32**	**89.79**	**75.42**	65.57	46.51	33.12	10.16	51.52	**60.60**	*59.70*
SeAFusion	SegFormer-B2	98.12	87.81	71.97	63.48	41.54	30.40	6.92	51.17	46.92	55.37
	SegFormer-B4	98.10	88.05	72.90	63.31	40.77	36.27	10.30	50.59	44.23	56.06
	SegNeXt-Base	98.20	88.76	72.40	*65.77*	44.52	37.29	6.08	*55.22*	51.92	57.80
	SegNeXt-Large	98.23	*89.18*	73.87	65.72	43.57	**44.90**	12.57	53.49	50.64	59.13
PSFusion	SegFormer-B2	98.08	86.85	73.91	63.70	41.07	37.50	0.13	50.32	52.60	56.02
	SegFormer-B4	98.14	87.65	74.14	63.79	44.45	35.23	8.76	53.85	51.54	57.51
	SegNeXt-Base	98.20	88.63	73.92	**66.23**	45.93	41.14	3.75	53.83	53.62	58.36
	SegNeXt-Large	*98.26*	88.62	*75.00*	65.53	*47.34*	*42.59*	8.58	52.96	*58.68*	**59.73**

scenario, only FusionGAN, RFN-Nest, and our approach provide promising fused images that enable SegFormer to correctly segment roadside bicycles and distant cars as well as pedestrians. Moreover, we can notice that in the third scene, SegNeXt is only able to segment the people hidden behind vehicles from the fused images generated by RFN-Nest, SwinFusion, SeAFusion, and our PSFusion. Both quantitative and qualitative comparisons fully validate the superior potential of the proposed approach over state-of-the-art image fusion algorithms for high-level vision tasks.

4.4. Exploring the advantages and potential of image-level fusion for high-level vision tasks

4.4.1. Quantitative analysis

In fact, another mainstream scheme that exploits complementary properties in multi-modal images to improve the performance of high-level vision tasks is the feature-level fusion-based scheme. However, this scheme, limited by sophisticated feature fusion, usually utilizes the venerable backbone as the feature extraction network. Notably, image-level fusion can synthesize a single fused image containing complementary information, which can be fed directly into state-of-the-art segmentation models without any redesign. Table 2 exhibits the quantitative comparison results of image-level fusion and feature-level fusion on the semantic segmentation task.

It is evident that the majority of feature-level fusion-based schemes employ ResNet [63] as the feature extraction network. The mean intersection over union (mIoU) of most of these schemes is around 55%, and only MFNet reaches 57.34% in mIoU. Even the most recent algorithms, such as LASNet and MDRNet can only achieve scores of 54.91% and 56.78% in mIoU, respectively. It is worth noting that CMX achieves a new state-of-the-art (SOTA) in the field of multi-modal segmentation by utilizing a more advanced backbone (*i.e.*, SegFormer [48]) as the feature extraction network. Even with the light-weight configuration (*i.e.*, SegFormer-B2), CMX still outperforms other algorithms. However, feature-level fusion usually requires two individual feature extraction networks to extract semantic features, which will significantly increase the number of parameters, especially when using a large-scale backbone.

The success of CMX inspired us to directly feed the fusion results into the existing large-scale segmentation models to improve segmentation performance. As shown in Table 2, although our PSFusion only achieves sub-optimal IoU in most categories, it performs relatively balanced across categories. In particular, PSFusion achieves segmentation performance (*i.e.*, mIoU) comparable to CMX-B4 when we perform semantic segmentation on the fusion results with SegNeXt-Large [66]. Even when using SegFormer as the segmentation model, our scheme is more competitive than the dominant multi-modal segmentation approaches based on feature-level fusion. Furthermore, another image-level fusion-based solution, *i.e.*, SeAFusion, also shows remarkable performance. The above situations indicate

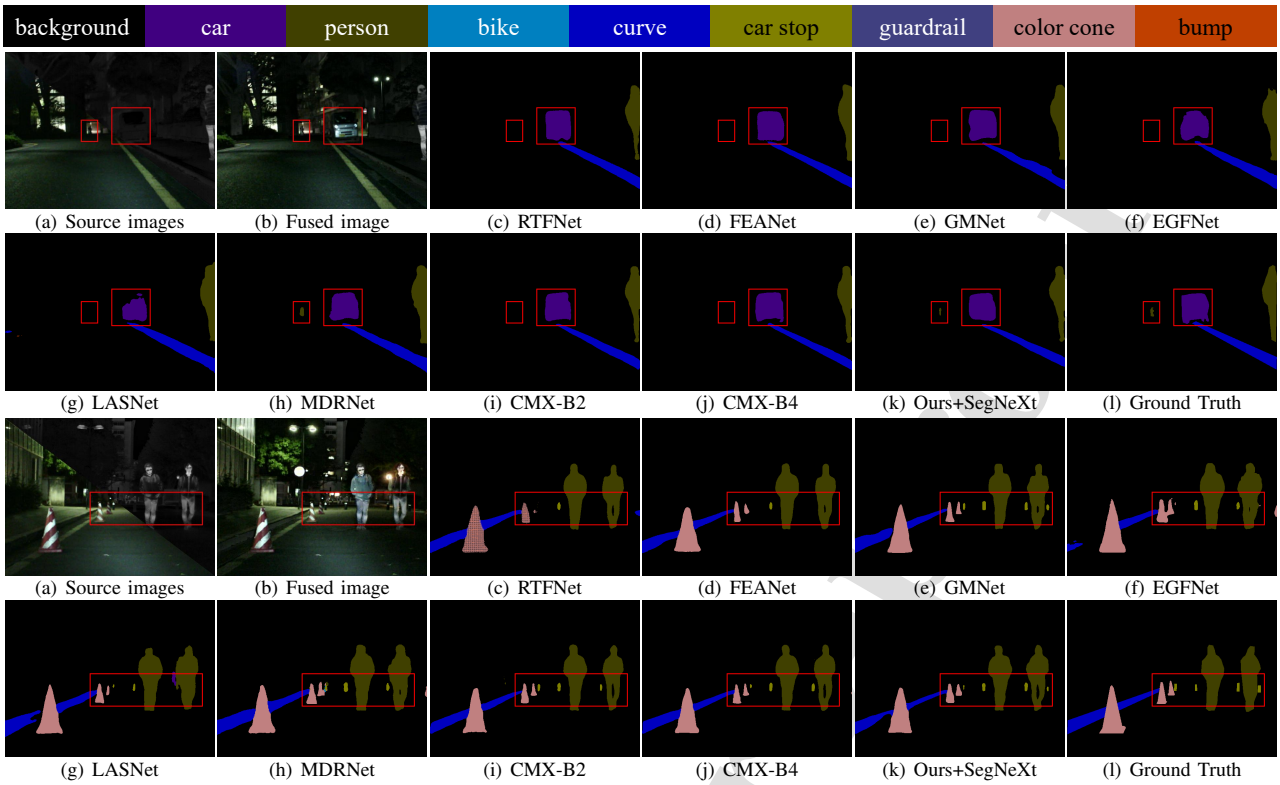


Figure 16: Segmentation results of feature-level fusion-based multi-modal segmentation algorithms and our image-level fusion-based solution on the MFNet dataset.

Figure 17: Visualization results of ablation studies.

that image-level fusion schemes can efficiently integrate the advantages of multi-modal inputs and novel unimodal semantic segmentation techniques and yield superior segmentation performance. Particularly in the era of large-scale models, unimodal segmentation techniques are developing rapidly. Image-level fusion can promptly combine the state-of-the-art unimodal segmentation models with multi-modal properties without any re-design, thus better coping with complex real-world circumstances.

4.4.2. Qualitative analysis

The visualization results of different schemes are presented in Figure 16. Although most methods can accurately segment objects in scenes, they exhibit differences in segmenting small objects. For instance, in the first scene, most methods, except for MDRNet and our solution, fail to correctly identify pedestrians in the distance. In the second scenario, our solution and GMNet can precisely segment four car stops from the scene, while other approaches can recognize only three or even fewer car stops. This advantage of our solution stems from the effective integration of complementary information in multi-modal images and the elimination of irrelevant and redundant information through image-level fusion. Both quantitative and qualitative analyses demonstrate that the combination of semantic-driven image-level fusion with state-of-the-art unimodal segmentation models is comparable to multi-modal segmentation schemes based on feature-level fusion. Particularly, with the rapid advancements of unimodal semantic segmentation techniques, image-level fusion-based schemes are more attractive and promising.

4.5. Ablation studies and discussion

4.5.1. Mask-based fusion loss analysis

We construct a specific intensity loss to maintain the visual appeal of fused images based on the contrast mask and salient target mask. Therefore, an ablation study that uses a rudimentary intensity loss to replace the mask-based intensity loss is performed to observe the particular function of our mask-based loss. The rudimentary intensity loss is

Table 3
The fusion performance of ablation studies. The best result is highlighted in Red.

	EN	SD	AG	SF	SCD	VIF
w/o Mask	5.213 ± 0.500	14.612 ± 4.689	1.603 ± 0.439	4.153 ± 1.047	0.984 ± 0.241	0.606 ± 0.066
w/o SFP	5.640 ± 0.553	18.704 ± 5.763	1.899 ± 0.557	5.067 ± 1.255	1.278 ± 0.169	0.654 ± 0.098
PSFusion	**6.838 ± 0.696**	**46.279 ± 12.595**	**4.762 ± 1.774**	**14.189 ± 4.470**	**1.811 ± 0.098**	**1.050 ± 0.166**

Table 4
The segmentation performance of ablation studies. SegNeXt-Base is employed to measure semantic information contained in the fused images. The best result is highlighted in Red.

	Background	Car	Person	Bike	Curve	Car Stop	Cuardrail	Color cone	Bump	mIoU
w/o S2P2	98.69	**92.08**	75.7	72.69	64.5	78.82	80.5	68.06	79.16	78.91
w/o PSFM	98.71	91.83	75.76	72.78	**65.46**	79.5	83.14	67.81	**80.65**	79.52
PSFusion	**98.73**	91.81	**77.66**	**73.64**	62.76	**80.15**	**88.92**	**70.17**	79.58	**80.38**

formulated as:

$$\mathcal{L}_{int} = \frac{1}{HW}\left\|I_f - I_{ir}\right\|_1 + \frac{1}{HW}\left\|I_f - I_{vi}\right\|_1. \quad (23)$$

As shown in Figure 17 (b), without the constraint of the mask-based loss, the fusion results behave in a more prosaic manner. Specifically, the fusion results can incorporate complementary information, whereas the significant targets are weakened, and the texture details suffer from redundant thermal radiation information. In addition, the quantitative results in Table 3 also show that all metrics reflecting fusion performance degrade significantly after removing the mask-based loss. This dilemma is blamed on the fact that without the guidance of specific masks, the fusion network fails to remove harmful information and preserve significant information purposefully.

4.5.2. Scene fidelity path analysis

The scene fidelity path is introduced to constrain the fusion features to retain as much complete information about source images as possible. As illustrated in Figure 17 (c), the visual effect of fused images is also slightly affected after removing the scene fidelity path. The fusion results fail to maintain the intensity of prominent objects without distortion. In contrast, our PSFusion can preserve significant information and deal with harsh environments under the synergistic constraints of the mask-based fusion loss and the scene fidelity path. Moreover, the quantitative results in Table 3 also demonstrate that our model could not maintain the excellent fusion performance if the SFP is removed.

4.5.3. Sparse semantic perception path analysis

Our method relies on the sparse semantic perception path (S2P2) to extract sufficient semantic features and inject them into the fusion network via the semantic injection module. As shown in Table 4, after removing the sparse semantic perception path, the segmentation performance of the segmentation model on the fused images degrades significantly. We believe this is due to the fact that the fusion network cannot adequately retain semantic information without the sparse semantic perception path to constrain the extraction of semantic features.

4.5.4. Profound semantic fusion module analysis

The profound semantic fusion module (PSFM) based on the cross-attention mechanism is developed to aggregate semantic features extracted from source images. We replace the PSFM with a superficial detail fusion module to verify its effectiveness in aggregation semantic features. As presented in Table 4, the performance of the segmentation model on the fused images is degraded without deploying the PSFM to merge semantic features of different domains. This indicates that integrating contextual semantic features from a global perspective can completely exploit and fuse semantic information.

4.5.5. Segmentation performance discussion

Considering Figure 4, our sparse semantic perception path is capable of outputting the semantic segmentation results. However, the design related to semantic segmentation is relatively simplified, as our primary goal is to generate fused images containing rich semantic information. Semantic segmentation serves as an auxiliary task in this context. As presented in Table 5, the segmentation results directly outputted by our model have significant room for improvement compared to the segmentation results obtained by feeding the fused image into SegNeXt. There are several reasons behind this issue. Firstly, we employ a simple backbone, *i.e.*, ResNet-34 as the basic feature extraction network and introduce two superficial feature extraction blocks to alleviate the conflicting demands between the high- and low-level vision tasks, which inevitably affects the performance of high-level vision tasks. Secondly, our S2P2 just utilizes simple CNNs to predict the segmentation results without involving more novel decoders in order to avoid increasing the computational load. Finally, we position semantic segmentation as an auxiliary task without designing specific modules to exploit the priors of image fusion to facilitate the improvement of the semantic segmentation

Table 5
Semantic segmentation results in the discussion section. Direct segmentation refers to the segmentation results directly outputted by our PSFusion model. PSFusion+SegNeXt indicates the segmentation results obtained by feeding the fused images of PSFusion into SegNeXt.

	Background	Car	Person	Bike	Curve	Car Stop	Cuardrail	Color cone	Bump	mIoU
Direct segmentation	98.44	89.38	74.24	71.25	58.36	66.9	69.43	62.51	74.05	73.84
PSFusion+SegNeXt	**98.73**	**91.81**	**77.66**	**73.64**	**62.76**	**80.15**	**88.92**	**70.17**	**79.58**	**80.38**

task. To address these issues, a potential direction for improvement is to introduce multi-task learning to fully leverage the mutual benefits of image fusion and high-level vision tasks, and integrate more advanced backbones into the network structure.

5. Conclusion

In this study, we proposed a practical image fusion network called PSFusion, which is based on progressive semantic injection and scene fidelity constraints. On the one hand, a sparse semantic perception branch that includes boundary segmentation, semantic segmentation, and binary segmentation is devised to extract sufficient semantic features. Then, we developed a semantic injection module to progressively integrate these semantic features into the fusion network. On the other hand, we introduce a scene fidelity path in the scene restoration branch that is responsible for reconstructing the source images to maintain the scene fidelity of the fusion features. Furthermore, we constructed a specific fusion loss based on the contrast mask and salient target mask to guarantee the visual appeal of fused images. The synergy of the scene restoration branch and semantic perception branch enables our fusion results to be informative for human visual observation and beneficial for machine vision perception. Extensive experiments demonstrated the superiority of our PSFusion over the existing image fusion algorithms in terms of visual quality and high-level semantics. Furthermore, quantitative and qualitative analyses demonstrate the potential of image-level fusion compared to feature-level fusion for high-level vision tasks, especially in the era of large-scale models. In particular, with the rapid advancements of unimodal semantic segmentation techniques, semantic-driven image-level fusion can fully integrate the superiority of multi-modal data and SOTA unimodal segmentation techniques without any re-design, thus better coping with complex scenes.

Acknowledgments

This research was sponsored by the National Natural Science Foundation of China (62276192).

References

[1] 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.

[2] Linfeng Tang, Hao Zhang, Han Xu, and Jiayi Ma. Deep learning-based image fusion: A survey. *Journal of Image and Graphics*, 28(1):3–36, 2023.

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

[4] Nirmala Paramanandham and Kishore Rajendiran. Infrared and visible image fusion using discrete cosine transform and swarm intelligence for surveillance applications. *Infrared Physics & Technology*, 88:13–22, 2018.

[5] Deepak Kumar Jain, Xudong Zhao, Germán González-Almagro, Chenquan Gan, and Ketan Kotecha. Multimodal pedestrian detection using metaheuristics with deep convolutional neural network in crowded scenes. *Information Fusion*, 95:401–414, 2023.

[6] Pengyu Zhang, Dong Wang, Huchuan Lu, and Xiaoyun Yang. Learning adaptive attribute-driven representation for real-time rgb-t tracking. *International Journal of Computer Vision*, 129:2714–2729, 2021.

[7] Qiang Zhang, Shenlu Zhao, Yongjiang Luo, Dingwen Zhang, Nianchang Huang, and Jungong Han. Abmdrnet: Adaptive-weighted bi-directional modality difference reduction network for rgb-t semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2633–2642, 2021.

[8] Jun Chen, Xuejiao Li, Linbo Luo, Xiaoguang Mei, and Jiayi Ma. Infrared and visible image fusion based on target-enhanced multiscale transform decomposition. *Information Sciences*, 508:64–78, 2020.

[9] Zhizhong Fu, Xue Wang, Jin Xu, Ning Zhou, and Yufei Zhao. Infrared and visible images fusion based on rpca and nsct. *Infrared Physics & Technology*, 77:114–123, 2016.

[10] Hui Li, Xiao-Jun Wu, and Josef Kitler. Mdlatlrr: A novel decomposition method for infrared and visible image fusion. *IEEE Transactions on Image Processing*, 29:4733–4746, 2020.

[11] Jinlei Ma, Zhiqiang Zhou, Bo Wang, and Hua Zong. Infrared and visible image fusion based on visual saliency map and weighted least square optimization. *Infrared Physics & Technology*, 82:8–17, 2017.

[12] Jiayi Ma, Linfeng Tang, Meilong Xu, Hao Zhang, and Guobao Xiao. Stdfusionnet: An infrared and visible image fusion network based on salient target detection. *IEEE Transactions on Instrumentation and Measurement*, 70:5009513, 2021.

[13] Hui Li and Xiao-Jun Wu. Densefuse: A fusion approach to infrared and visible images. *IEEE Transactions on Image Processing*, 28(5):2614–2623, 2018.

[14] Jiayi Ma, Wei Yu, Pengwei Liang, Chang Li, and Junjun Jiang. Fusiongan: A generative adversarial network for infrared and visible image fusion. *Information Fusion*, 48:11–26, 2019.

[15] 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.

[16] 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 Conference on Computer Vision and Pattern Recognition*, pages 5802–5811, 2022.

[17] Yiming Sun, Bing Cao, Pengfei Zhu, and Qinghua Hu. Detfusion: A detection-driven infrared and visible image fusion network. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4003–4011, 2022.

[18] Wujie Zhou, Jinfu Liu, Jingsheng Lei, Lu Yu, and Jenq-Neng Hwang. Gmnet: graded-feature multilabel-learning network for rgb-thermal urban scene semantic segmentation. *IEEE Transactions on Image Processing*, 30:7790–7802, 2021.

[19] Yanpeng Cao, Xing Luo, Jiangxin Yang, Yanlong Cao, and Michael Ying Yang. Locality guided cross-modal feature aggregation and pixel-level fusion for multispectral pedestrian detection. *Information Fusion*, 88:1–11, 2022.

[20] 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.

[21] 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.

[22] 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.

[23] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

[24] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11976–11986, 2022.

[25] Yipeng Liu, Jing Jin, Qiang Wang, Yi Shen, and Xiaoqiu Dong. Region level based multi-focus image fusion using quaternion wavelet and normalized cut. *Signal Processing*, 97:9–30, 2014.

[26] Qiong Zhang and Xavier Maldague. An adaptive fusion approach for infrared and visible images based on nsct and compressed sensing. *Infrared Physics & Technology*, 74:11–20, 2016.

[27] Yu Liu, Xun Chen, Rabab K Ward, and Z Jane Wang. Image fusion with convolutional sparse representation. *IEEE Signal Processing Letters*, 23(12):1882–1886, 2016.

[28] Jiao Mou, Wei Gao, and Zongxi Song. Image fusion based on non-negative matrix factorization and infrared feature extraction. In *Proceedings of the International Congress on Image and Signal Processing*, pages 1046–1050, 2013.

[29] Jiayi Ma, Chen Chen, Chang Li, and Jun Huang. Infrared and visible image fusion via gradient transfer and total variation minimization. *Information Fusion*, 31:100–109, 2016.

[30] Yu Liu, Shuping Liu, and Zengfu Wang. A general framework for image fusion based on multi-scale transform and sparse representation. *Information Fusion*, 24:147–164, 2015.

[31] Meilong Xu, Linfeng Tang, Hao Zhang, and Jiayi Ma. Infrared and visible image fusion via parallel scene and texture learning. *Pattern Recognition*, 132:108929, 2022.

[32] Hui Li, Xiao-Jun Wu, and Tariq Durrani. Nestfuse: An infrared and visible image fusion architecture based on nest connection and spatial/channel attention models. *IEEE Transactions on Instrumentation and Measurement*, 69(12):9645–9656, 2020.

[33] Hui Li, Xiao-Jun Wu, and Josef Kittler. Rfn-nest: An end-to-end residual fusion network for infrared and visible images. *Information Fusion*, 73:720–86, 2021.

[34] Zixiang Zhao, Shuang Xu, Chunxia Zhang, Junmin Liu, Jiangshe Zhang, and Pengfei Li. Didfuse: deep image decomposition for infrared and visible image fusion. In *International Joint Conference on Artificial Intelligence*, pages 970–976, 2020.

[35] Linfeng Tang, Xinyu Xiang, Hao Zhang, Meiqi Gong, and Jiayi Ma. Divfusion: Darkness-free infrared and visible image fusion. *Information Fusion*, 91:477–493, 2023.

[36] Han Xu, Hao Zhang, and Jiayi Ma. Classification saliency-based rule for visible and infrared image fusion. *IEEE Transactions on Computational Imaging*, 7:824–836, 2021.

[37] Linfeng Tang, Jiteng Yuan, Hao Zhang, Xingyu Jiang, and Jiayi Ma. Piafusion: A progressive infrared and visible image fusion network based on illumination aware. *Information Fusion*, 83:79–92, 2022.

[38] Yongzhi Long, Haitao Jia, Yida Zhong, Yadong Jiang, and Yuming Jia. Rxdnfuse: a aggregated residual dense network for infrared and visible image fusion. *Information Fusion*, 69:128–141, 2021.

[39] Jiayi Ma, Pengwei Liang, Wei Yu, Chen Chen, Xiaojie Guo, Jia Wu, and Junjun Jiang. Infrared and visible image fusion via detail preserving adversarial learning. *Information Fusion*, 54:85–98, 2020.

[40] Di Wang, Jinyuan Liu, Xin Fan, and Risheng Liu. Unsupervised misaligned infrared and visible image fusion via cross-modality image generation and registration. In *International Joint Conference on Artificial Intelligence*, pages 3508–3515, 7 2022.

[41] Linfeng Tang, Yuxin Deng, Yong Ma, Jun Huang, and Jiayi Ma. Superfusion: A versatile image registration and fusion network with semantic awareness. *IEEE/CAA Journal of Automatica Sinica*, 9(12):2121–2137, 2022.

[42] Zhanbo Huang, Jinyuan Liu, Xin Fan, Risheng Liu, Wei Zhong, and Zhongxuan Luo. Reconet: Recurrent correction network for fast and efficient multi-modality image fusion. In *Proceedings of the European Conference on Computer Vision*, pages 539–555, 2022.

[43] Jiayi Ma, Han Xu, Junjun Jiang, Xiaoguang Mei, and Xiao-Ping Zhang. Ddcgan: A dual-discriminator conditional generative adversarial network for multi-resolution image fusion. *IEEE Transactions on Image Processing*, 29:4980–4995, 2020.

[44] Jing Li, Hongtao Huo, Chang Li, Renhua Wang, and Qi Feng. Attentionfgan: Infrared and visible image fusion using attention-based generative adversarial networks. *IEEE Transactions on Multimedia*, 23:1383–1396, 2020.

[45] Huabing Zhou, Wei Wu, Yanduo Zhang, Jiayi Ma, and Haibin Ling. Semantic-supervised infrared and visible image fusion via a dual-discriminator generative adversarial network. *IEEE Transactions on Multimedia*, 25:635–648, 2023.

[46] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the International Conference on Learning Representations*, 2020.

[47] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6881–6890, 2021.

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

[49] Liting Lin, Heng Fan, Zhipeng Zhang, Yong Xu, and Haibin Ling. Swintrack: A simple and strong baseline for transformer tracking. In *Advances in Neural Information Processing Systems*, 2022.

[50] Fuzhi Yang, Huan Yang, Jianlong Fu, Hongtao Lu, and Baining Guo. Learning texture transformer network for image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5791–5800, 2020.

[51] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12299–12310, 2021.

[52] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1833–1844, 2021.

[53] Vibashan Vs, Jeya Maria Jose Valanarasu, Poojan Oza, and Vishal M Patel. Image fusion transformer. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 3566–3570. IEEE, 2022.

[54] Zhihao Chang, Zhixi Feng, Shuyuan Yang, and Quanwei Gao. Aft: Adaptive fusion transformer for visible and infrared images. *IEEE Transactions on Image Processing*, 32:2077–2092, 2023.

[55] Han Xu, Jiayi Ma, Jiteng Yuan, Zhuliang Le, and Wei Liu. Rfnet: Unsupervised network for mutually reinforcing multi-modal image registration and fusion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 19679–19688, 2022.

[56] 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.

[57] Yu Zhang, Yu Liu, Peng Sun, Han Yan, Xiaolin Zhao, and Li Zhang. Ifcnn: A general image fusion framework based on convolutional neural network. *Information Fusion*, 54:99–118, 2020.

[58] 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.

[59] Pengwei Liang, Junjun Jiang, Xianming Liu, and Jiayi Ma. Fusion from decomposition: A self-supervised decomposition approach for image fusion. In *Proceedings of the European Conference on Computer Vision*, pages 719–735, 2022.

[60] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015.

[61] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12):2481–2495, 2017.

[62] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2881–2890, 2017.

[63] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[64] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4700–4708, 2017.

[65] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12179–12188, 2021.

[66] Meng-Hao Guo, Cheng-Ze Lu, Qibin Hou, Zheng-Ning Liu, Ming-Ming Cheng, and Shi-min Hu. Segnext: Rethinking convolutional attention design for semantic segmentation. In *Advances in Neural Information Processing Systems*, 2022.

[67] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834–848, 2017.

[68] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision*, pages 801–818, 2018.

[69] Hang Zhang, Kristin Dana, Jianping Shi, Zhongyue Zhang, Xiaogang Wang, Ambrish Tyagi, and Amit Agrawal. Context encoding for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7151–7160, 2018.

[70] Xiangtai Li, Xia Li, Li Zhang, Guangliang Cheng, Jianping Shi, Zhouchen Lin, Shaohua Tan, and Yunhai Tong. Improving semantic segmentation via decoupled body and edge supervision. In *Proceedings of the European Conference on Computer Vision*, pages 435–452, 2020.

[71] Yifei Zhang, Désiré Sidibé, Olivier Morel, and Fabrice Mériaudeau. Deep multimodal fusion for semantic image segmentation: A survey. *Image and Vision Computing*, 105:104042, 2021.

[72] 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 International Conference on Intelligent Robots and Systems*, pages 5108–5115, 2017.

[73] Yuxiang Sun, Weixun Zuo, and Ming Liu. Rtfnet: Rgb-thermal fusion network for semantic segmentation of urban scenes. *IEEE Robotics and Automation Letters*, 4(3):2576–2583, 2019.

[74] 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 *IEEE International Conference on Intelligent Robots and Systems*, pages 4467–4473. IEEE, 2021.

[75] Huayao Liu, Jiaming Zhang, Kailun Yang, Xinxin Hu, and Rainer Stiefelhagen. Cmx: Cross-modal fusion for rgb-x semantic segmentation with transformers. *arXiv preprint arXiv:2203.04838*, 2023.

[76] Wendong Zhang, Junwei Zhu, Ying Tai, Yunbo Wang, Wenqing Chu, Bingbing Ni, Chengjie Wang, and Xiaokang Yang. Context-aware image inpainting with learned semantic priors. In *International Joint Conference on Artificial Intelligence*, pages 1323–1329, 2021.

[77] Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. Training region-based object detectors with online hard example mining. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 761–769, 2016.

[78] Alexander Toet. The tno multiband image data collection. *Data in Brief*, 15:249–251, 2017.

[79] Chengli Peng, Tian Tian, Chen Chen, Xiaojie Guo, and Jiayi Ma. Bilateral attention decoder: A lightweight decoder for real-time semantic segmentation. *Neural Networks*, 137:188–199, 2021.

[80] 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.

[81] Yun-Jiang Rao. In-fibre bragg grating sensors. *Measurement Science and Technology*, 8(4):355, 1997.

[82] Guangmang Cui, Huajun Feng, Zhihai Xu, Qi Li, and Yueting Chen. Detail preserved fusion of visible and infrared images using regional saliency extraction and multi-scale image decomposition. *Optics Communications*, 341:199–209, 2015.

[83] Ahmet M Eskicioglu and Paul S Fisher. Image quality measures and their performance. *IEEE Transactions on Communications*, 43(12):2959–2965, 1995.

[84] 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.

[85] Yu Han, Yunze Cai, Yin Cao, and Xiaoming Xu. A new image fusion performance metric based on visual information fidelity. *Information Fusion*, 14(2):127–135, 2013.

[86] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pages 8026–8037, 2019.

1. We demonstrate the advantages of image-level fusion over feature-level fusion.
2. We propose PSFusion, a sematic-driven IR/VIS fusion model based on PSIM and SFP.
3. We use SIM to inject semantics at the feature level to accommodate various models.
4. SFP is developed to maintain complete information of the fusion features.
5. Abundant results prove the superiority of PSFusion in visual appeal and semantic.

Linfeng Tang: Methodology, Experiment, Writing – original draft. **Hao Zhang**: Methodology. **Han Xu:** Methodology. **Jiayi Ma**: Conceptualization of this study, Methodology, Writing – review & editing.

Declaration of interests

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

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: