Information Fusion 82 (2022) 28–42

Contents lists available at ScienceDirect

Information Fusion

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

Full length article

Image fusion in the loop of high-level vision tasks: A semantic-aware real-time infrared and visible image fusion network

Linfeng Tang, Jiteng Yuan, Jiayi Ma *

Electronic Information School, Wuhan University, Wuhan 430072, China

ARTICLE INFO

Keywords:
Image fusion
Semantic aware
High-level vision task
Gradient residual dense block
Task-driven evaluation

ABSTRACT

Infrared and visible image fusion aims to synthesize a single fused image that not only contains salient targets and abundant texture details but also facilitates high-level vision tasks. However, the existing fusion algorithms unilaterally focus on the visual quality and statistical metrics of fused images but ignore the demands of high-level vision tasks. To address these challenges, this paper bridges the gap between image fusion and high-level vision tasks and proposes a semantic-aware real-time image fusion network (SeAFusion). On the one hand, we cascade the image fusion module and semantic segmentation module and leverage the semantic loss to guide high-level semantic information to flow back to the image fusion module, which effectively boosts the performance of high-level vision tasks on fused images. On the other hand, we design a gradient residual dense block (GRDB) to enhance the description ability of the fusion network for fine-grained spatial details. Extensive comparative and generalization experiments demonstrate the superiority of our SeAFusion over state-of-the-art alternatives in terms of maintaining pixel intensity distribution and preserving texture detail. More importantly, the performance comparison of various fusion algorithms in task-driven evaluation reveals the natural advantages of our framework in facilitating high-level vision tasks. In addition, the superior running efficiency allows our algorithm to be effortlessly deployed as a real-time pre-processing module for high-level vision tasks. The source code will be released at https://github.com/Linfeng-Tang/SeAFusion.

1. Introduction

Images shot by a single modal sensor fail to effectively and comprehensively describe the imaging scene due to the theoretical and technical limitations [1]. The infrared sensor captures thermal radiation emitted from objects, which could highlight salient targets, but the infrared image neglects texture and is vulnerable to noise. On the contrary, the visible sensor captures reflective light information. The visible image usually contains abundant texture and structure information but is sensitive to the environment, such as illumination and occlusion. The complementary roles encourage us to fuse infrared and visible images to generate a desired image, which highlights prominent targets as well as manifests abundant detail information. Thus, the infrared and visible image fusion has been broadly used as a pre-processing module for high-level vision tasks, *e.g.*, object detection [2], tracking [3], pedestrian re-identification [4] and semantic segmentation [5]. An example in Fig. 1 intuitively shows the contribution of fused images to the segmentation task. From the visible image, the segmentation network could segment cars, bikes and several persons but ignore the pedestrian hidden in the darkness. Although the infrared image helps the segmentation network to exactly split cars and all pedestrians, it neglects bikes. By exploiting the complementary information of infrared and visible images, the fused image enhances the segmentation accuracy of all bikes, cars, and pedestrians.

Due to the practicality of infrared and visible image fusion, it has attracted a great deal of scholarly attention. In the past decades, numerous image fusion techniques have been proposed, including traditional approaches [8–10] and recent deep learning-based methods [1]. The traditional approaches typically fall into five categories, *i.e.*, multi-scale transform (MST)-based methods [11–14], sparse representation (SR)-based methods [15,16], subspace-based methods [17–19], optimization-based methods [20] and hybrid methods [21]. Among the deep learning-based methods, auto-encoder (AE) based framework [6,22,23], convolutional neural network (CNN) based framework [24–26] and generative adversarial network (GAN) based framework [27–30] are the dominant frameworks.

Although the recent deep learning-based image fusion algorithms could generate satisfying fused images, there are still some pressing challenges in the image fusion community. On the one hand, the existing fusion algorithms are inclined to pursue better visual quality and higher evaluation metrics but seldom systematically consider whether

* Corresponding author.
E-mail addresses: linfeng0419@gmail.com (L. Tang), yuanjiteng@whu.edu.cn (J. Yuan), jyma2010@gmail.com (J. Ma).

https://doi.org/10.1016/j.inffus.2021.12.004
Received 25 October 2021; Received in revised form 28 November 2021; Accepted 8 December 2021
Available online 1 January 2022

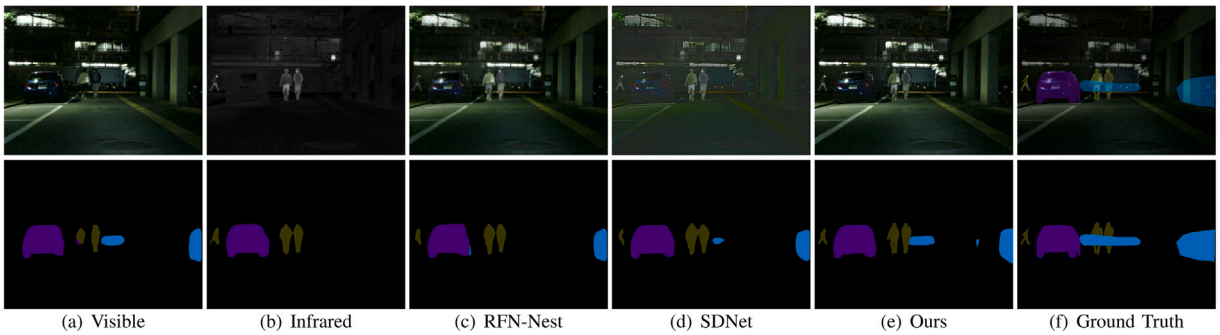

Fig. 1. An example of infrared and visible image fusion and segmentation. The first row is the source images and the fused images. The second row is the segmentation results of the first row. From left to right: visible image, infrared image, fused images of RFN-Nest [6], SDNet [7], and our proposed SeAFusion, and ground truth.

fused images can facilitate high-level vision tasks. Some studies [31–33] demonstrate that only considering the visual quality and quantitative metrics could not help with high-level vision tasks. Although some works introduce perceptual loss to constrain the fused image and source images at the feature level [6,7,24,28,34], the perceptual loss cannot effectively enhance the semantic information in the fused image, as shown in Fig. 1. Moreover, other researchers guide the image fusion process via a segmentation mask [26,35], but the mask only splits some salient targets, which is limited for boosting semantic information. On the other hand, the existing evaluation manners are mainly visual comparison and quantitative evaluation. The visual comparison focuses on the contrast and texture detail of the fused image and the quantitative evaluation relies on some statistical metrics to assess fusion performance. However, neither visual comparison nor quantitative evaluation reflects the facilitation of fused images for high-level vision tasks. In addition, the existing network architectures are not effective in extracting fine-grained detail features. Last but not least, numerous existing fusion algorithms ignore the demand for real-time image fusion while striving for improving visual quality and evaluation metrics.

In this study, a semantic-aware fusion network, known as SeAFusion, is proposed to achieve real-time infrared and visible image fusion. The key of our method is simultaneously obtaining superior performance in both image fusion and high-level vision tasks. Specifically, we introduce a segmentation network to predict the segmentation results on fused images, which is utilized to construct semantic loss. Then, the semantic loss is leveraged to guide the training of the fusion network via back-propagation, forcing fused images to contain more semantic information. Moreover, in order to meet the demand of high-level vision tasks in real-time, we develop a light-weight network based on gradient residual dense block (GRDB). The GRDB could achieve feature reuse via the main dense stream and boost the description ability for fine-grained details by the residual gradient stream.

To sum up, the major contributions of this study are summarized as follows:

- We devise a novel semantic-aware infrared and visible image fusion framework, which effectively achieves superior performance in both image fusion and high-level vision tasks.
- A gradient residual dense block is designed to boost the description ability of the network for fine-grained detail and achieve feature reuse.
- The proposed SeAFusion is a light-weight model that can achieve real-time image fusion. This allows it to be deployed as a pre-processing module for high-level vision tasks.
- We propose a task-driven evaluation manner that evaluates the performance of image fusion from the perspective of high-level vision tasks.

The remainder of this paper is organized as follows. Section 2 briefly describes the related works of image fusion and task-driven algorithms. In Section 3, we introduce our proposed SeAFusion in detail, including the problem analysis, loss function, network architecture and training strategy. Section 4 illustrates the impressive performance of our method in comparison with other alternatives, followed by some concluding remarks in Section 5.

2. Related work

In this section, we first review the existing infrared and visible image fusion algorithms. Then, some task-driven low-level vision algorithms are briefly introduced.

2.1. Image fusion algorithms

2.1.1. Traditional image fusion methods

Since feature reconstruction is usually an inverse process of feature extraction, the key to traditional image fusion algorithms lies in two vital elements, *i.e.*, feature extraction and fusion. Multi-scale decomposition is the most common transformation scheme for feature extraction. In the past decades, numerous multi-scale transforms such as laplacian pyramid (LP), discrete wavelet [11], shearlet [12], nonsubsampled contourlet [13] transform, and latent low-rank representation [15] have been successfully embedded in the multi-scale transform-based image fusion framework. In addition, sparse representation is exploited as a feature extraction technique, which uses the sparse basis in an over-complete dictionary to represent source images [16,36]. Moreover, subspace-based methods have also attracted great attention, which project high-dimensional images into low-dimensional subspace to capture intrinsic structures of source images. Independent component analysis [17], principal component analysis [19], and non-negative matrix factorization [18] are the representative methods in the subspace-based fusion framework.

Beyond the aforementioned methods, the optimization-based methods offer fresh perspectives and prospects for the image fusion community. In particular, Ma et al. defined the infrared and visible image fusion as overall intensity maintenance and texture structure preservation, which lays a solid foundation for CNN-based approaches [26] and GAN-based approaches [20]. Moreover, some researchers combined the advantages of different frameworks and proposed hybrid models to pursue better image fusion performance [21,37]. In particular, Liu et al. developed a general image fusion framework via combining multi-scale transform (MST) and sparse representation (SR) to concurrently conquer the inherent defects of both the MST-based and SR-based fusion approaches.

It is instructive to note that the increasingly complex transformations or representations cannot respond to the demands of real-time image fusion [38]. Furthermore, the hand-crafted activity level measurements and fusion rules activity level measurements fail to integrate semantic information, which will limit the contribution of fused results to high-level vision tasks.

2.1.2. AE-based image fusion methods

With the superior feature learning ability of the neural network, deep learning has become a new favorite for numerous vision tasks. The image fusion community has also actively explored deep learning-based solutions and developed many promising schemes. The auto-encoder (AE)-based framework is a crucial branch, which trains an auto-encoder to achieve feature extraction and reconstruction. Li et al. proposed a simple fusion architecture that consists of three components: encoder layer, fusion layer and decoder layer [22]. The encoder layer contains a convolutional layer and denseblock to high-level features, in which the denseblock is leveraged to get more useful features in the encoding process. The fusion layer leverages element-wise addition strategy or l1-norm strategy to merge high-level features, and the feature reconstruction network contains four convolutional layers to reconstruct the fused image. Furthermore, they also introduced the multi-scale encoder–decoder architectures and nest connection to extract more comprehensive features [6,23]. Nevertheless, the aforementioned approaches applied hand-crafted fusion rules to integrate deep features, restraining the fusion performance severely. To address the limitations of manually designed fusion rules, Xu et al. proposed a classification saliency-based rule for AE-based image fusion framework [39]. The novel fusion rule employs a classifier to measure the saliency of each pixel in feature maps and calculates the fusion weights based on the contribution of each pixel.

2.1.3. CNN-based image fusion methods

CNN-based fusion frameworks either achieve implicit feature extraction, aggregation and image reconstruction guided by the elaborate loss function, or employ convolutional neural network (CNN) as part of the overall fusion framework to implement activity level measurements and feature integration. LP-CNN is a pioneer in applying CNN to the image fusion field, which combines the LP with classification-type CNN to achieve medical image fusion [40]. In addition, Zhang et al. developed a general image fusion framework via the generic network structure, *i.e.*, feature extraction layer, fusion layer and image reconstruction layer [24]. It is instructive to note that their fusion layer is embedded in the training process. Hence, IFCNN could mitigate the limitation imposed by the manually designed fusion rules (element-wise max, element-wise min, or element-wise mean).

Moreover, researchers have also explored an alternative solution, *i.e.*, end-to-end CNN-based image fusion framework, to avoid the shortcomings of hand-crafted rules. The CNN-based methods inherit the core concept of traditional optimization-based approaches, which defines the objective function of image fusion as overall intensity fidelity and texture structure preservation [20]. Zhang et al. modeled the unified image fusion as proportional maintenance of gradient and intensity and designed a general loss function for different image fusion missions [41]. Based on the gradient and intensity paths, they also devised a squeeze-and-decomposition network to improve the fidelity of fused images [7]. In addition, an adaptive decision block was introduced to allocate weights for gradient loss terms according to the texture richness of source images. Considering the cross-fertilization between different image fusion missions, Xu et al. trained a unified model for multi-fusion tasks [42]. In order to reinforce the semantic information in fused images, Ma et al. [26] leveraged a salient mask to construct the desired information for infrared and visible image fusion. Although the proposed network can detect salient targets, the simple salient target mask only enhances the semantic information of the salient target region. In addition, for the image fusion task it is difficult to provide the ground truth to construct loss functions, which means the CNN-based fusion network cannot release its full potential performance.

2.1.4. GAN-based image fusion methods

Since the adversarial loss constructs networks from the perspective of probability distributions, the generative adversarial network is ideal for unsupervised tasks, *e.g.*, image-to-image translation [43,44] and image fusion [27,45]. Ma et al. creatively introduced the GAN into the image fusion community, which leverages a discriminator to force the generator to synthesize fused images with abundant textures [27]. In order to improve the quality of detail information and sharpen the edge of thermal targets, they also introduced the detail loss and edge-enhancement loss [28]. However, a single discriminator may cause mode collapse, *i.e.*, the fused image is biased towards visible or infrared images. Therefore, Ma et al. further proposed a dual-discriminator conditional generative adversarial network to improve the robustness of the GAN-based framework and maintain the balance between infrared and visible images [29]. Subsequently, Li et al. integrated a multi-scale attention mechanism, prompting the generator and discriminator to pay more attention to the typical regions, into the GAN-based fusion framework [46]. In addition, Ma et al. transformed image fusion into a multi-distribution simultaneous estimation problem and achieved the balance between infrared and visible images from the perspective of classifiers [47].

However, both traditional approaches and deep learning-based methods emphasize the improvements of fused image quality and evaluation metrics while ignoring the demands of high-level vision tasks. In practice, a fused image with excellent image quality may be suitable for human visual perception, but may not facilitate high-level vision tasks. A robust image fusion algorithm should enhance the semantic information of fused images while fully integrating the complementary information in the source images. Some deep learning-based algorithms tried to enhance the semantic information in the fused images via introducing the perceptual loss or salient target masks. But the perceptual loss has limited benefit for semantic information enhancement. In addition, the salient target mask only reinforces the semantic information in the salient target regions. To this end, it is imperative to develop a semantic-aware image fusion algorithm.

2.2. Task-driven low-level vision algorithms

In fact, some practical solutions, combining the low-level algorithms with the demands of high-level vision tasks, have been proposed. Li et al. designed a light-weight dehazing network, called AOD-Net, which can easily be embedded into other deep models, *e.g.*, Faster R-CNN, for improving high-level vision tasks on hazy images [48]. AOD-Net is the first work that explores the correlation between the dehazing algorithms and the high-level vision task performance. Subsequently, Haris et al. explored how image super-resolution can contribute to object detection in low-resolution images [31]. They developed a novel super-resolution framework where the super-resolution sub-network explicitly incorporates a detection loss in its training objective. Moreover, Liu et al. proposed a task-driven image denoising scheme, which cascaded two modules for image denoising and various high-level vision tasks, and used a joint loss for updating only the parameters of denoising model [49,50]. Their solution could overcome the performance degradation of high-level vision tasks and generate more visually appealing results with the guidance of high-level vision information. Recently, considering the practical demands of automatic driving, Guo et al. devised a novel deraining framework that contains a semantic refinement residual network and a two-stage segmentation aware joint training strategy [51].

In this work, we devise a semantic-aware real-time infrared and visible image fusion framework to reinforce the semantic information in the fused images. More specifically, we first introduce the semantic loss to integrate more semantic information into fused images. Then, we tailor a joint low-level and high-level adaptive training strategy to maintain the balance of low-level and high-level vision tasks. Finally, our fusion model can generate more visually appealing fused images while achieving excellent performance for high-level vision tasks with the guidance of a semantic loss.

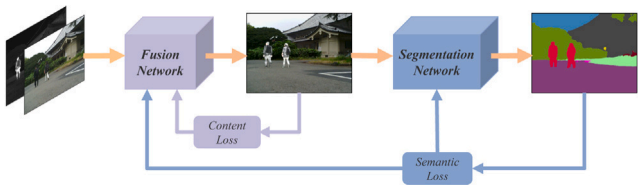

Fig. 2. The overall framework of the proposed semantic-aware infrared and visible image fusion algorithm.

3. Methodology

We comprehensively describe the semantic-aware real-time infrared and visible image fusion framework in this section. Firstly, we provide the problem formulation of our SeAFusion. Then, the content loss and semantic loss are presented in detail. Afterwards, the architecture of our gradient residual dense block-based fusion network is presented. Finally, we introduce the training strategy in detail.

3.1. Problem formulation

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 image fusion is achieved via feature extraction, aggregation and reconstruction with the guidance of a tailored loss function. Therefore, the quality of the fused image $I_f \in \mathbb{R}^{H\times W\times 3}$ depends greatly on the loss function. In order to improve fusion performance, we devise a joint loss, consisting of content loss and semantic loss, to constrain the fusion network. The overall framework of our semantic-aware infrared and visible image fusion algorithm is presented in Fig. 2.

First of all, a light-weight fusion network based on gradient residual dense block (GRDB) is devised to fully integrate the complementary information in source images. More specifically, we apply a feature extraction module E_F to extract deep features with abundant fine-grained detail information from infrared and visible images, which can be represented as:

$$\left\{F_{ir}, F_{vi}\right\} = \left\{E_F(I_{ir}), E_F(I_{vi})\right\}, \tag{1}$$

where F_{ir} and F_{vi} mean infrared features and visible features, respectively. Moreover, GRDBs embedded in the feature extraction module, are deployed to boost the description ability for fine-grained details while extracting high-level semantic features (we will discuss its network architecture in Section 3.3). Given the input F^i of GRDB, its output F^{i+1} can be denoted as:

$$\begin{aligned} F^{i+1} &= GRDB(F^i) \\ &= Conv^n(F^i) \oplus Conv(\nabla F^i), \end{aligned} \tag{2}$$

where $Conv(\cdot)$ indicates the convolutional layer and $Conv^n(\cdot)$ stands for n cascaded convolution layers. ∇ refers to the gradient operator, *i.e.*, a special convolutional operation whose convolutional kernel is manually devised. The gradient operator convolves the input features with the high-frequency convolution kernel to extract the fine-grained detail information. In this study, the well-known Sobel operator is exploited to compute gradient magnitude. Moreover, $\oplus$ denotes element-wise summation. GRDB aggregates the learnable convolutional features with gradient magnitude information.

Then, the fused image is reconstructed by a feature integration and image reconstruction module. The concatenation fusion strategy is leveraged to integrate the deep infrared and visible features, containing abundant fine-grained spatial details. The fusion process is expressed as follows:

$$F_f = C(F_{ir}, F_{vi}), \tag{3}$$

where $C(\cdot)$ refers to concatenation in the channel dimension. Eventually, the fused image I_f is recovered from the fused features F_f via the image reconstructor $\mathcal{R}_I$, which is presented as:

$$I_f = \mathcal{R}_I(F_f). \tag{4}$$

In addition, taking full account of the demands of high-level vision tasks on fused images, we adopt a semantic loss to measure the semantic information contained in the fused image. More specifically, we introduce a segmentation model N_s to perform segmentation on the fused image $I_f \in \mathbb{R}^{H\times W\times 3}$ [52]. The gap between the segmentation result $I_s \in \mathbb{R}^{H\times W\times C}$ and semantic label $L_s \in (1, C)^{H\times W}$ can reflect the richness of the semantic information contained in the fused image, where H and W are the height and width of an image, respectively, and C denotes the number of object categories. Given a fused image I_f, the semantic-aware process is denoted as:

$$I_s = N_s(I_f). \tag{5}$$

The gap between the segmentation result and semantic label is denoted as $\mathcal{L}_{semantic}$ and defined as:

$$\mathcal{L}_{semantic} = \mathcal{E}(I_s, L_s), \tag{6}$$

where $\mathcal{E}(\cdot)$ indicates the error function. More details about the error function will be presented in Section 3.2.

3.2. Loss function

Our SeAFusion aims to reinforce the semantic information in the fused images while boosting the visual quality and evaluation metrics. In order to achieve these goals, we devise our loss functions from two perspectives. On the one hand, SeAFusion needs to fully integrate complementary information in source images such as the prominent targets in the infrared image and texture details in the visible image. To this end, the content loss is designed to ensure the visual fidelity of fused images. On the other hand, the fused image should effectively facilitate high-level vision tasks. For this purpose, we construct a semantic loss to reflect the degree to that fused images contribute to high-level vision tasks.

3.2.1. Content loss

In order to promote our fusion model integrate more meaningful information and improve the visual quality and quantitative metrics, we devise a content loss. The content loss consists of two components, *i.e.*, intensity loss $\mathcal{L}_{int}$ and texture loss $\mathcal{L}_{texture}$. The exact definition of content loss is expressed as follows:

$$\mathcal{L}_{content} = \mathcal{L}_{int} + \alpha\mathcal{L}_{texture}, \tag{7}$$

where $\mathcal{L}_{int}$ constrains the overall apparent intensity of fused images, and $\mathcal{L}_{texture}$ forces fused images to contain more fine-grained texture details. Here, α is used to strike a balance between the intensity loss $\mathcal{L}_{int}$ and texture loss $\mathcal{L}_{texture}$.

The intensity loss measures the difference between fused images and source images at the pixel level. Therefore, we define the intensity loss of infrared and visible images as:

$$\mathcal{L}_{int} = \frac{1}{HW}\left\|I_f - \max(I_{ir}, I_{vi})\right\|_1, \tag{8}$$

where H and W are the height and width of an image, respectively, $\|\cdot\|_1$ stands for the l_1-norm and $\max(\cdot)$ denotes the element-wise maximum selection. We integrate the pixel intensity distribution of infrared and visible images via a maximum selection strategy. Then, the integrated distribution is leveraged to constrain the pixel intensity distribution of the fused image.

We expect the fused image to maintain the best intensity distribution and preserve abundant texture details from source images simultaneously. However, the intensity loss only provides a coarse-grained distribution constrain for model learning. Therefore, a texture

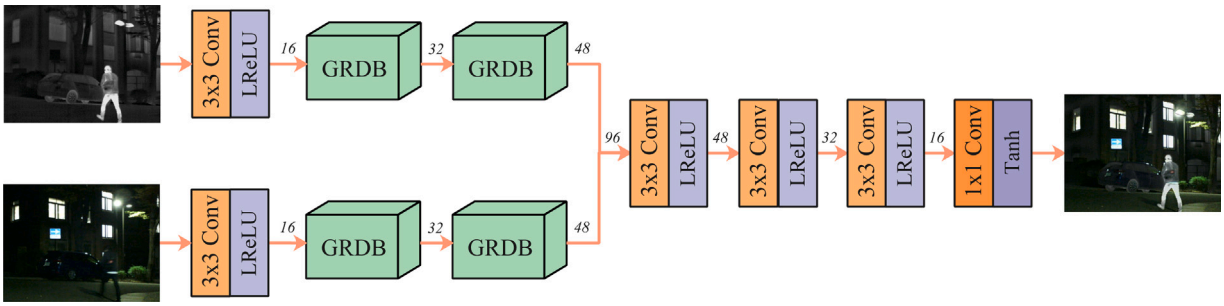

Fig. 3. The architecture of the real-time infrared and visible image fusion network based on gradient residual dense block.

loss is introduced to force the fused image to contain more fine-grained texture information. The texture loss is defined as:

$$\mathcal{L}_{texture} = \frac{1}{HW}\left\|\left|\nabla I_f\right| - \max(\left|\nabla I_{ir}\right|, \left|\nabla I_{vi}\right|)\right\|_1, \tag{9}$$

where ∇ indicates the Sobel gradient operator, which measures the fine-grained texture information of an image. $|\cdot|$ refers to the absolute operation. We assume that the optimal texture of the fused image is the maximum aggregate of infrared and visible image textures.

In conclusion, our fusion network based on gradient residual dense block can simultaneously achieve the optimal intensity distribution and preserve abundant detail information with the guidance of the content loss. In other words, the content loss could effectively guarantee that our model achieves the first goal, *i.e.*, improving the visual quality and statistical evaluation metrics of fused images.

3.2.2. Semantic loss

Adequately boosting the semantic information of fused images is the greatest innovation in our algorithm. We creatively devise a semantic loss to fulfill this goal. More specifically, we introduce a real-time semantic segmentation model [52] to split the fused images, and the segmentation network outputs segmentation result $I_s \in \mathbb{R}^{H\times W\times C}$ and auxiliary segmentation result $I_{sa} \in \mathbb{R}^{H\times W\times C}$. The semantic loss consists of two elements, *i.e.*, main semantic loss and auxiliary semantic loss. The primary semantic loss and auxiliary semantic loss are defined as follows:

$$\mathcal{L}_{main} = \frac{-1}{H\times W}\sum_{h=1}^{H}\sum_{w=1}^{W}\sum_{c=1}^{C} L_{so}^{(h,w,c)} \log(I_s^{(h,w,c)}), \tag{10}$$

$$\mathcal{L}_{aux} = \frac{-1}{H\times W}\sum_{h=1}^{H}\sum_{w=1}^{W}\sum_{c=1}^{C} L_{so}^{(h,w,c)} \log(I_{sa}^{(h,w,c)}), \tag{11}$$

where $L_{so} \in \mathbb{R}^{H\times W\times C}$ denotes a one-hot vector transformed from the segmentation label $L_s \in (1, C)^{H\times W}$. The main semantic loss and auxiliary loss reflect the semantic information contained in the fused images from different perspectives. Ultimately, the semantic loss is expressed as follows:

$$\mathcal{L}_{semantic} = \mathcal{L}_{main} + \lambda\mathcal{L}_{aux}, \tag{12}$$

where λ is a constant for balancing the main semantic loss and auxiliary semantic loss, which is set to 0.1 referring to the original paper [52]. It is worth remarking that beyond constraining the fusion network, semantic loss is also deployed to train the segmentation model. More detailed descriptions of loss functions and network architecture for the segmentation model refer to [52].

Finally, a joint loss is constructed to guide the training of fusion model, which is defined as:

$$\mathcal{L}_{joint} = \mathcal{L}_{content} + \beta\mathcal{L}_{semantic}, \tag{13}$$

where β is a hyper-parameter characterizing the importance of semantic loss $\mathcal{L}_{semantic}$. It is important to emphasize that β increases progressively according to the joint low-level and high-level adaptive training strategy since the segmentation network becomes adaptive with the fusion model as training progresses.

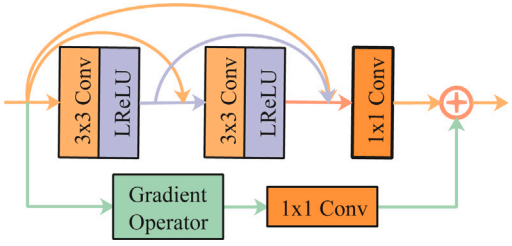

Fig. 4. The specific devise of the gradient residual dense block. The Sobel operator is selected as the Gradient Operator to extract fine-grained detail information of feature maps.

3.3. Network architecture

To achieve real-time image fusion, we propose a light-weight infrared and visible image fusion network based on GRDB, which is illustrated in Fig. 3. Our fusion network consists of a feature extractor and image reconstructor, where the feature extractor contains two GRDBs to extract fine-grained features.

As shown in Fig. 3, the feature extractor involves two parallel infrared and visible feature extraction streams, and each containing a common convolutional layer and two GRDBs. The common convolutional layer, whose kernel size is 3×3 and activate function is Leaky Rectified Linear Unit (LReLU), is deployed to extract shallow features. Immediately following are two GRDBs for extracting fine-grained features from the shallow features. The specific design of GRDB is illustrated in Fig. 4. The gradient residual dense block is a variant of resblock [26], where the main stream employs the dense connection and the residual stream integrates the gradient operation. We can observe from Fig. 4 that the main stream deploys two 3×3 convolutional layers with LReLU and one common convolutional layer whose kernel size is 1×1. It should be emphasized that we introduce the dense connection into the main stream to make full use of features extracted by various convolutional layers. The residual stream employs a gradient operation to calculate the gradient magnitude of features and a 1×1 regular convolutional layer to eliminate channel dimensional differences. Then, adding the outputs of the main dense stream and residual gradient stream via an element-wise addition to integrate deep features and fined-grained detail features.

Subsequently, the fine-grained features of infrared and visible images are integrated via the concatenation strategy, and the results are fed into the image reconstructor to achieve feature aggregation and image reconstruction. The image reconstructor consists of three tandem 3×3 convolutional layers and one 1×1 convolutional layer. All 3×3 convolutional layers employ LReLU as the activation function, while the activation function of the 1×1 convolutional layer is Tanh.

It is well-known that information loss is a catastrophic issue in the image fusion task. Therefore, the padding in our fusion network is set as same, and stride is set to 1 except for 1×1 convolutional layers. As a result, our network does not introduce any down-sampling, and the size of fused images is consistent with source images.

Algorithm 1: Joint low-level and high-level adaptive training strategy

Input: Infrared images I_{ir} and visible images I_{vi}
Output: Fused images I_f

for $m \leqslant$ *Max iterations* M **do**
  for p *steps* **do**
    Select b infrared images $\{I_{ir}^1, I_{ir}^2, \ldots, I_{ir}^b\}$;
    Select b visible images $\{I_{vi}^1, I_{vi}^2, \ldots, I_{vi}^b\}$;
    Update the weight of semantic loss β according to Eq. (14);
    Update the parameters of the fusion network N_F by AdamOptimizer: $\bigtriangledown_{N_F}(\mathcal{L}_{joint}(N_F))$;
  end
  Generate fused images from infrared and visible images in the training set;
  for q *steps* **do**
    Select b fused images $\{I_f^1, I_f^2, \ldots, I_f^b\}$;
    Update the parameters of the segmentation network N_S by SGD Optimizer: $\bigtriangledown_{N_S}(\mathcal{L}_{semantic}(N_S))$;
  end
end

3.4. Joint low-level and high-level adaptive training strategy

The existing task-driven low-level vision methods either adopt a pre-trained high-level model to guide the training of low-level vision task models or joint train low-level and high-level vision task models in one stage. However, in the image fusion field it is difficult to provide the ground truth of fused images for training a high-level vision task model. In addition, one stage joint training strategy may lead to difficulties in maintaining the balance of performance between low-level and high-level vision tasks. To this end, we devise a joint low-level and high-level training strategy to train our fusion network. More specifically, we iteratively train the fusion network and segmentation network, and set the iterations to M. Firstly, all parameters in the fusion network are optimized by the Adam optimizer with the guidance of the joint loss. Moreover, the hyper-parameter β of joint loss is dynamically adjusted with the iteration, which is expressed as follows:

$$\beta = \gamma \times (m - 1), \tag{14}$$

where m denotes the mth iteration. β increases gradually as training progresses due to the fact that the segmentation network fits the fusion model better as iterations increase, and the semantic loss can guide the fusion network training more exactly. γ is a constant for balancing the semantic loss and content loss. Then, given the current fused results, the parameters of the segmentation model are updated via optimizing the semantic loss. In each iteration, the training steps of the fusion model and segmentation model are p and q, respectively. The joint low-level and high-level adaptive training strategy is summarized in Algorithm 1.

4. Experimental validation

In this section, we first provide the experimental configurations and implementation details. Then, we present some comparative experiments and generalization experiments to reveal the superiority of our proposed SeAFusion. In addition, some task-driven evaluation experiments are performed to evaluate different fusion methods from the perspective of high-level vision tasks. Next, we compare the running efficiency of different approaches to verify the superiority of our light-weight network for real-time image fusion. Finally, some ablation studies are performed to demonstrate the effectiveness of our specific designs, including the semantic loss, gradient residual dense block and joint low-level and high-level adaptive training strategy.

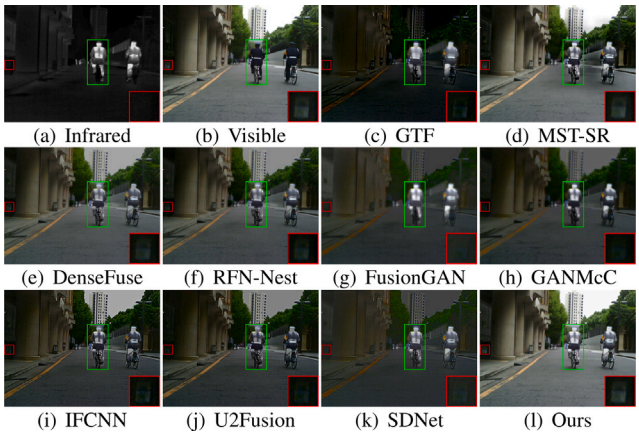

Fig. 5. Qualitative comparison of SeAFusion with 9 state-of-the-art methods on *00537D* image from the MFNet dataset. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

4.1. Experimental configurations

To comprehensively evaluate the proposed algorithm, we perform extensively qualitative and quantitative experiments on the MFNet [5], RoadScene [42] and TNO [53] datasets. We compare our method with nine state-of-the-art approaches, including two traditional approaches, *i.e.*, GTF [20] and MST-SR [37], two AE-based approaches, *i.e.*, DenseFuse [22], and RFN-Nest [6], two GAN-based approaches, *i.e.*, FusionGAN [27] and GANMcC [47], and three CNN-based methods, *i.e.*, IFCNN [24], U2Fusion [42] and SDNet [47]. The implementations of all these nine methods are publicly available, and we set the parameters as reported in the original papers. It is worth noting that we adopt Laplace pyramid (LP) as the multi-scale transformation (MST) in MST-SR. The element-wise addition, element-wise maximum fusion strategy, and residual fusion network (RFN) are deployed to integrate the deep features for DenseFuse, IFCNN, and RFN-Nest, respectively.

Six statistical evaluation metrics are selected to quantify the evaluation, including entropy (EN) [54], mutual information (MI) [55], visual information fidelity (VIF) [56], spatial frequency (SF) [57], standard deviation (SD) and Q_{abf}. EN measures the amount of information contained in the fused image, and MI takes stock of the amount of information transferred from the source images to the fused image. Both EN and MI evaluate the fusion performance from the information theory perspective. VIF evaluates the information fidelity of the fused image from the perspective of the human visual system. SF measures the spatial frequency information contained in the fused images. SD reflects the distribution and contrast of fused images from the statistical perspective. Q_{abf} takes stock of the amount of edge information transferred from source images to the fused image. EN, SF and SD are reference-free metrics. Moreover, a fusion algorithm with larger EN, MI, VIF, SF, SD and Q_{abf} indicates better fusion performance.

4.2. Implementation details

We train our semantic-aware fusion network on the MFNet dataset. The training set contains 1083 pairs of infrared and visible images and the testing set consists of 361 image pairs. The MFNet dataset provides semantic labels for nine objects, *i.e.*, car, person, bike, curve, car stop, guardrail, color Tone and background. Moreover, all images are normalized to $[0, 1]$ before being fed into networks.

We iteratively train the fusion network and the segmentation network according to the joint low-level and high-level adaptive training strategy. All parameters in our joint adaptive training strategy are set as follows: $M = 4$, $p = 2,700$, $q = 20{,}000$ and $\gamma = 1$. In addition, the hyper-parameter of the content loss is set as $\alpha = 10$. We leverage

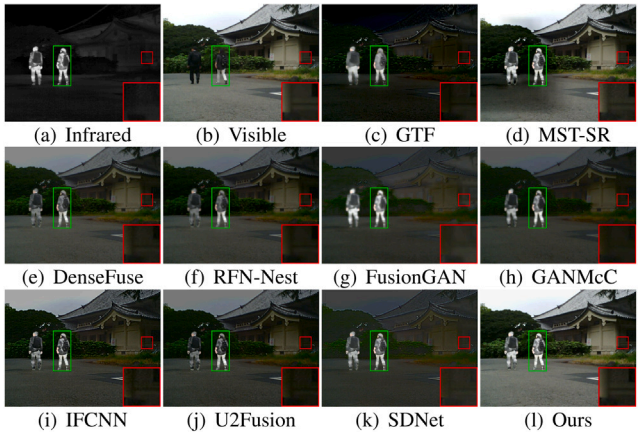

Fig. 6. Qualitative comparison of SeAFusion with 9 state-of-the-art methods on *00633D* image from the MFNet dataset. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

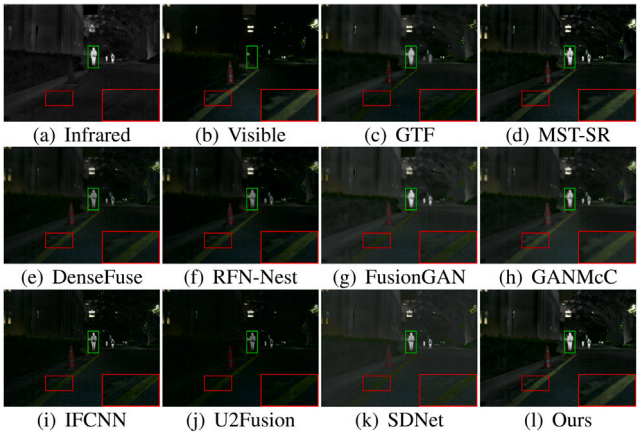

Fig. 7. Qualitative comparison of SeAFusion with 9 state-of-the-art methods on *00858N* image from the MFNet dataset. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Adam optimizer with a batch size of 8, $\beta 1$ of 0.9, $\beta 2$ of 0.99, epsilon of $1e^{-8}$, weight decay of 0.0002, the initial learning rate of 0.001 to optimize our fusion model with the guidance of joint loss. Furthermore, we utilize mini-batch stochastic gradient descent with a batch size of 16, momentum of 0.9, and weight decay of 0.0005 to optimize the segmentation network. The initial learning rate is set as 0.01 and the learning rate is updated by the initial learning rate multiplied by $(1-\frac{iter}{max_{iter}})^{power}$, where power is set as 0.9 [52]. The proposed method is implemented on the PyTorch platform [58]. All experiments are conducted on the NVIDIA TITAN RTX GPU and 3.50 GHz Intel(R) Core(TM) i9-9920X CPU.

We use a special strategy [59] to process color information since the MFNet and RoadScene datasets contain color visible images. More specifically, we first convert visible images to the YCbCr color space. Then, different fusion algorithms are leveraged to fuse the Y channel of visible images and infrared images. Finally, the fused image can be converted into the RGB color space with Cb and Cr channels of visible images. Moreover, we feed the RGB fused image into the segmentation model directly.

4.3. Comparative experiment

In order to sufficiently evaluate the fusion performance of our method, we first compare the proposed SeAFusion with other nine algorithms on the MFNet dataset.

4.3.1. Qualitative results

The MFNet dataset contains two typical scenes, *i.e.*, the daytime scene and nighttime scene. To intuitively exhibit the superiority of our fusion framework in integrating complementary information and improving the visual quality of fused images, we select two daytime scenes and two nighttime scenarios for subjective evaluation. The visualized results are presented in Figs. 5–8. In daytime scenes, thermal radiation information of infrared images can be leveraged as supplementary information for visible images. Therefore, a fused image with pleasing visual quality should contain abundant texture detail of the visible image and enhance prominent targets in the infrared image. As presented in Figs. 5 and 6, GTF and FusionGAN cannot preserve texture detail of visible images, and FusionGAN fails to sharpen the edge of prominent targets. Although DenseFuse, RFN-Nest, GANMcC, U2Fusion and SDNet integrate the detail information of visible images with the thermal radiation information of infrared images, both types of information are inevitably interfered with by useless information during the fusion process. We zoom in on a region with the red box to illustrate the phenomenon that texture details suffer from varying degrees of spectral contamination. In addition, a salient area is highlighted by a green box to reveal the issue of useless information weakening prominent targets. Only our SeAFusion and MST-SR could preserve abundant texture details while highlighting prominent targets. Unfortunately, MST-SR is vulnerable to contamination by thermal radiation information in some background regions (*e.g.*, the ground in Figs. 5 and 6). Thus, only our method can effectively integrate the complementary information from source images and guarantee the visual quality of the fused image simultaneously.

In the nighttime scenarios, both infrared and visible images only provide limited scene information. Thus, it is a challenge to integrate meaningful information from infrared and visible images adaptively. As shown in Figs. 7 and 8, we can observe that all algorithms merge the complementary information in infrared and visible images to some extent, but there are still some slight variations in fused results of different algorithms. In particular, GTF and FusionGAN blur the contours of thermal radiation targets, and the texture region of GTF suffers from severe spectral contamination. Except for our SeAFusion, other methods introduce some useless information into the fused images, which are reflected in the contamination of detailed textures and the weakening of salient targets. We zoom in on a textured area (*i.e.*, the red box) and highlight a target (*i.e*, the green box) to display the aforementioned issues. It is worth emphasizing that our SeAFusion purposely integrates the meaningful information in source images with guidance of semantic loss and generates fused images containing abundant semantic information. Furthermore, the fused images synthesized by our fusion model contain abundant texture details, benefiting from the powerful description capability of the gradient residual dense block for fine-grained details.

4.3.2. Quantitative results

The quantitative results of six statistical metrics on 361 image pairs are displayed in Fig. 9. One can notice that our method exhibits significant superiority in four metrics, *i.e.*, EN, MI, VIF and Q_{abf}. The best EN indicates fused images generated by SeAFusion contain the most information, and the highest MI means our method transfers the most information from source images to fused images. Moreover, our SeAFusion presents the best VIF, which indicates that our fused images are more consistent with the human visual system. Furthermore, the proposed approach achieves the best Q_{abf}, which implies that more edge information is preserved in the fused results, benefiting from the powerful fine-grained feature extraction ability of GRDB. In addition, our SeAFusion exhibits the best SD, meaning our fused images have the highest contrast. Our method only follows IFCNN and MST-SR by a narrow margin in the SF metric.

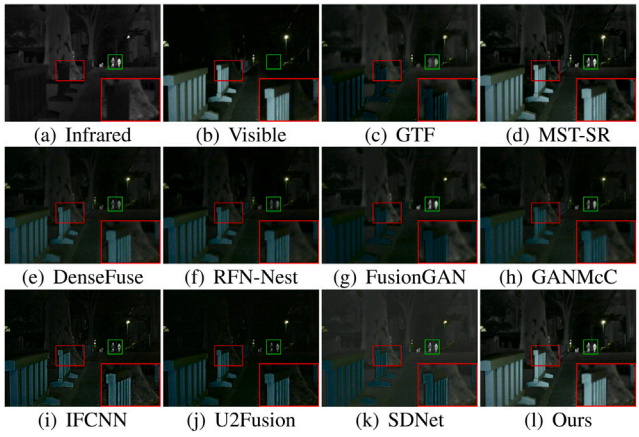

Fig. 8. Qualitative comparison of SeAFusion with 9 state-of-the-art methods on *01024N* image from the MFNet dataset. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

4.4. Generalization experiment

It is well known that generalization performance is an essential aspect in evaluating deep learning-based methods. Therefore, we provide generalization experiments on the RoadScene and TNO datasets to demonstrate the generalizability of the proposed SeAFusion. Noteworthy, our fusion model is trained on the MFNet dataset and tested directly on RoadScene and TNO datasets.

4.4.1. Qualitative results

The qualitative comparisons of different algorithms on the RoadScene dataset are shown in Figs. 10 and 11. Almost all methods introduce meaningless information in the fusion process, which is manifested by texture areas suffering from thermal radiation contamination and weakened salient targets. In order to intuitively exhibit the effect of meaningless information for fused images, we zoom in on an area with rich texture details in a red box and highlight a prominent target in the green box. We can observe that texture details in the background regions are disturbed by thermal radiation information. GTF, DenseFuse, RFN-Nest, FusionGAN, GANMcC and SDNet are particularly evident. Moreover, the intensity information of salient targets is weakened to varying degrees. GTF and FusionGAN cannot retain the sharpened edge of targets. It is worth mentioning that MST-SR, IFCNN, U2Fusion and SeAFusion suffer from only minor interference of useless information. In particular, our fused results are similar to visible images in background areas, and the pixel intensities of salient targets are consistent with infrared images.

The visualized results of different methods on the TNO dataset are shown in Figs. 12 and 13. As can be seen from the green box, MST-SR, DenseFuse, RFN-Nest and U2Fusion severely weaken the salient targets. Moreover, FusionGAN and GANMcC blur the edge of thermal targets. In addition, the fused images generated by other methods suffer from serious spectral contamination in the background regions, *e.g.*, the bush in Fig. 13. Only our method successfully preserves the texture details of visible images and maintains the intensity of salient targets.

4.4.2. Quantitative results

We also select 25 image pairs from the RoadScene and TNO datasets for quantitative evaluation, respectively. The comparative results of different methods on the six metrics are displayed in Figs. 14 and 15. From Fig. 14, we can notice that SeAFusion presents remarkable superiority in EN, MI, VIF, SF and SD on the RoadScene dataset. Such phenomena mean our fused images not only contain abundant information and texture details but also have the highest contrast and the best visual quality. Moreover, SeAFusion ranks second in Q_{abf}, which implies that our method transfers sufficient edge information from source images to fused images.

As shown in Fig. 15, SeAFusion ranks first in EN, MI, VIF and Q_{abf} metrics by a significant margin on the TNO dataset. For the SD metric, our method still ranks first, although the advantage is not particularly pronounced. Finally, the proposed method only follows IFCNN by a narrow margin in the SF metric.

In conclusion, extensive qualitative and quantitative results on various datasets demonstrate the superiority of our method in terms of prominent target maintenance and texture preservation. We attribute the advantage to the following aspects. On the one hand, we define a content loss consisting of the intensity loss and texture loss to constrain the fusion network to effectively integrate meaningful information from the perspective of pixel intensity distribution and high-order texture detail. On the other hand, our fusion network can adaptively merge complementary features in conjunction with semantic information with the guidance of semantic loss. Finally, the elaborate gradient residual dense block is deployed to enhance the description ability of the network for fine-grained details.

4.5. Task-driven evaluation

The fused images are utilized not only for visual observations but also for high-level vision tasks. However, the existing evaluation manners only concentrate on the visual quality of fused images and statistical metrics. In this section, we break away from the restraints of existing assessment manners and propose a task-driven evaluation criterion. More specifically, we perform semantic segmentation and object detection on fused images and compute the segmentation or detection performance of different fusion methods.

4.5.1. Segmentation performance

For fair comparisons, we re-train the segmentation network [52] for different fusion algorithms on the original MFNet dataset. The configuration of the training and testing sets is consistent with that used for training our SeAFusion. Firstly, we generate the fused images using each fusion method. Then, eleven segmentation models are trained with 80,000 steps on the infrared, visible and nine fused image training sets, respectively. Our SeAFusion employs the joint low-level and high-level adaptive training strategy to jointly train the fusion model and segmentation model, and the segmentation network is also trained with 80,000 steps. The segmentation performance is measured by pixel intersection-over-union (IoU). The segmentation-driven evaluation results are reported in Table 1. One can observe that our algorithm achieves the highest IoU in almost all categories and ranks first in mIoU. We attribute the advantage to two points. On the one hand, the complementary information of infrared and visible images is completely merged by our fusion network. The complementary information helps segmentation models to understand imaging scenes comprehensively, which is an important reason why fusion can improve segmentation performance. On the other hand, our SeAFusion adaptively integrates the meaningful/semantic information with the guidance of semantic loss. Therefore, our fused images contain abundant semantic information, allowing the segmentation network to describe the imaging scenes more accurately. We believe boosting the semantic information in the fused images is the core factor that elevates our method ahead of other algorithms in segmentation performance.

In addition, we also provide some visualized examples to show the segmentation results on infrared, visible and different fused images. We only present the segmentation results of five representative fusion algorithms, *i.e.*, GTF, FusionGAN, DenseFuse, IFCNN and SDNet, as shown in Fig. 16. From the results, we can find that infrared images provide more information about prominent targets such as pedestrians, while visible images could offer a better description for backgrounds (*e.g.*, cars, bikes and color cones). The excellent fusion algorithms can integrate complementary information from source images and achieve a

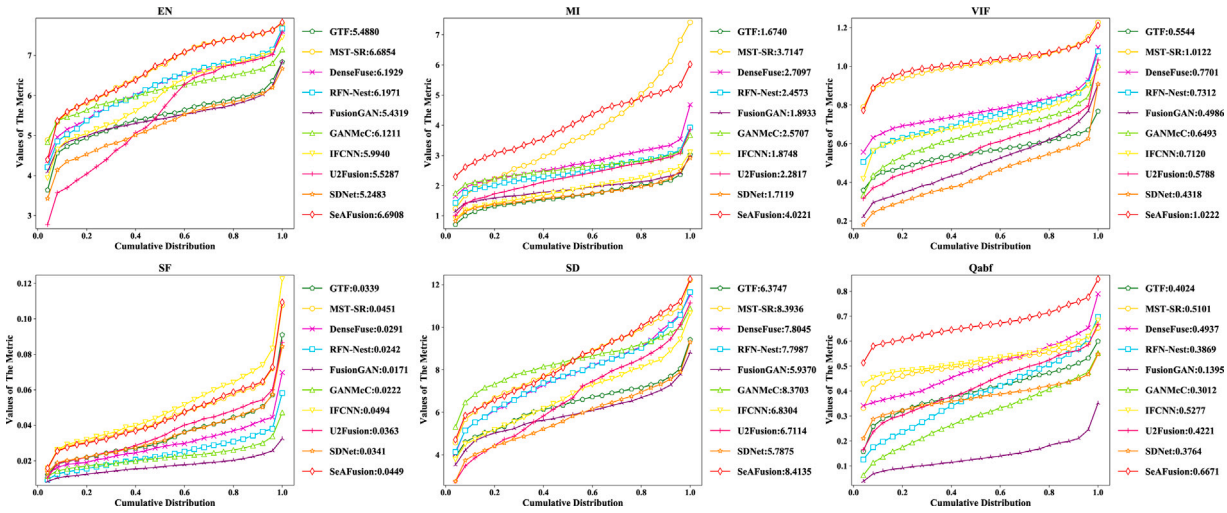

Fig. 9. Quantitative comparisons of the six metrics, *i.e.*, EN, MI, VIF, SF, SD and Q_{abf}, on 361 image pairs from the MFNet dataset. A point (x, y) on the curve denotes that there are $100 * x$ percent of image pairs which have metric values no more than y.

Table 1
Segmentation performance (mIoU) of visible, infrared and fused images on the MFNet dataset. RED indicates the best result and *BLUE* represents the second best result.

	Background	Car	Person	Bike	Curve	Car Stop	Cuardrail	Color Tone	Bump	mIoU
Visible	98.26	89.03	59.94	70.00	60.69	71.43	77.90	63.42	75.31	74.00
Infrared	98.24	87.33	70.46	69.23	58.74	68.85	65.57	56.93	72.72	72.01
GTF	98.44	89.12	71.76	*72.04*	64.47	70.59	68.95	63.71	74.40	74.83
MST-SR	98.50	*90.03*	72.21	71.45	62.75	70.21	75.78	*65.44*	77.84	76.02
DenseFuse	98.50	89.30	72.77	71.72	63.43	72.16	74.43	64.88	*80.10*	76.36
RFN-Nest	98.50	89.95	72.03	71.39	62.04	74.92	74.87	63.41	79.53	76.29
IFCNN	98.48	90.00	72.30	71.41	62.43	70.55	73.26	63.25	77.32	75.44
FusionGAN	98.49	89.86	72.83	71.95	63.45	71.68	*79.45*	64.35	75.36	*76.38*
GANMcC	98.47	89.26	72.11	71.74	62.71	72.94	74.05	63.26	77.42	75.77
U2Fusion	98.49	89.78	72.93	70.99	62.84	72.13	79.25	63.59	77.12	76.35
SDNet	*98.52*	89.58	*73.41*	71.61	*63.68*	**75.59**	75.31	61.82	75.43	76.11
Ours	**98.61**	**90.43**	**74.30**	**72.18**	**65.01**	*74.08*	**85.25**	**66.50**	**81.41**	**78.64**

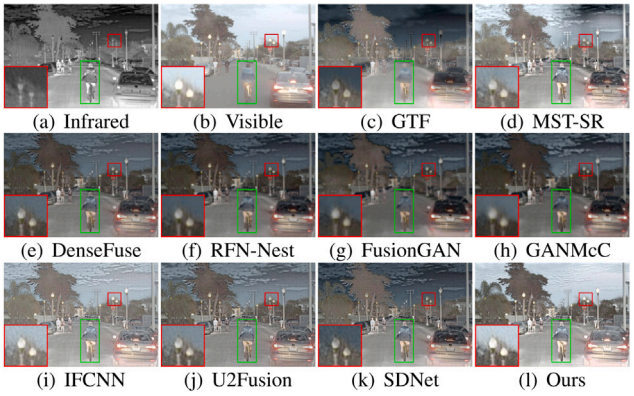

Fig. 10. Qualitative comparison of SeAFusion with 9 state-of-the-art methods on *FLIR_06832* from the RoadScene dataset. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

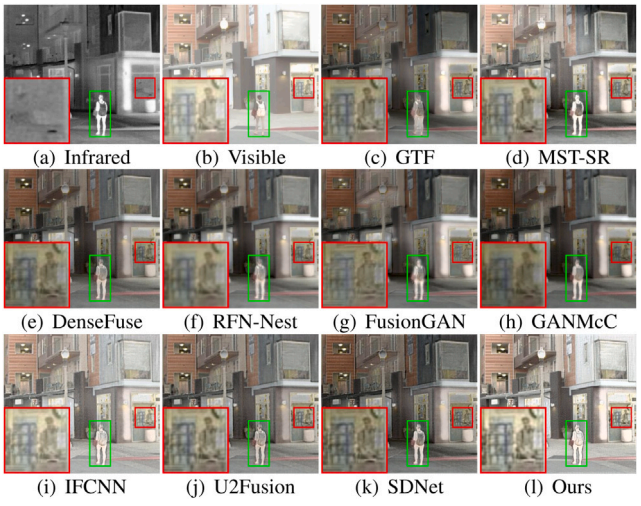

Fig. 11. Qualitative comparison of SeAFusion with 9 state-of-the-art methods on *FLIR_08835* from the RoadScene dataset. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

more comprehensive description for imaging scenes. Thus, the segmentation model could obtain better segmentation results on fused images. It is worth mentioning that our fusion method fully merges the semantic information of source images during the fusion process. Therefore, the segmentation model can generate more accurate segmentation results on our fused images, *e.g.*, the car stops in *00127D* scenario, the color cone and bikes in *00504D* scene and the bicycles in *01066N* scenario.

Besides the re-trained segmentation models, Deeplab-V3+ [60], trained on the Cityscapes dataset [61], is also leveraged to measure the contribution of various fusion algorithms to high-level vision tasks. Visible images, infrared images, and fused results of different fusion methods are directly input into Deeplabv3+, respectively. The visualized results are reported in Fig. 17. Since DeeplabV3+ is trained on a

Table 2
Object detection performance (mAP) of visible, infrared and fused images on the MFNet dataset. The best result is indicated by RED and the second best result is represented by *BLUE*.

	AP@0.5			AP@0.7			AP@0.9			mAP@[0.5:0.95]		
	Person	Car	All	Person	Car	All	Person	Car	All	Person	Car	All
Infrared	**0.963**	0.707	0.835	*0.888*	0.629	0.759	**0.260**	0.228	0.244	**0.706**	0.501	0.604
Visible	0.671	0.941	0.806	0.407	**0.889**	0.648	0.017	0.239	0.128	0.346	0.720	0.533
GTF	0.801	0.916	0.859	0.625	0.849	0.737	0.080	0.407	0.243	0.499	0.720	0.609
MST-SR	0.934	**0.964**	*0.949*	0.846	0.851	0.849	0.100	0.485	0.292	0.629	0.753	0.691
DenseFuse	0.946	0.937	0.941	0.823	*0.887*	0.855	0.089	0.534	0.312	0.618	0.754	0.686
RFN-Nest	0.870	0.933	0.902	0.687	0.836	0.761	0.062	0.401	0.232	0.536	0.717	0.627
IFCNN	0.961	0.919	0.940	0.835	0.843	0.839	0.103	0.513	0.308	0.615	0.749	0.682
FusionGAN	0.879	0.916	0.898	0.755	0.836	0.796	0.143	0.464	0.304	0.594	0.722	0.658
GANMcC	0.948	*0.952*	**0.95**	0.839	0.886	*0.863*	0.112	0.500	0.306	0.625	0.753	0.689
U2Fusion	0.941	0.927	0.934	0.790	0.855	0.822	0.092	*0.596*	*0.344*	0.597	0.753	0.675
SDNet	0.961	0.927	0.944	0.830	0.886	0.858	0.124	0.526	0.325	0.639	0.753	*0.696*
Ours	*0.962*	0.928	0.945	**0.897**	0.885	**0.891**	*0.222*	**0.645**	**0.434**	*0.692*	**0.779**	**0.736**

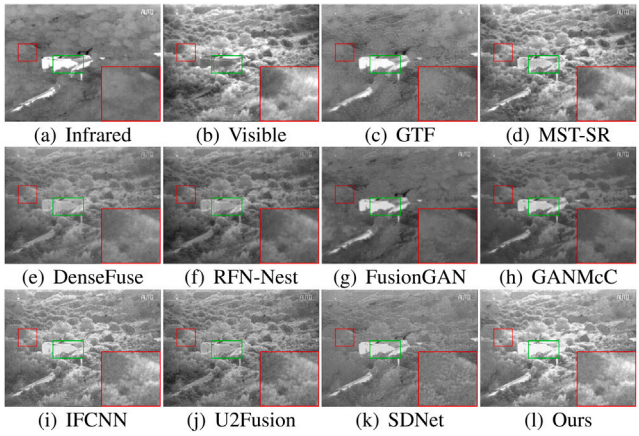

Fig. 12. Visualized results of SeAFusion compared with 9 state-of-the-art algorithms on *Kaptein_1123* from the TNO dataset. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

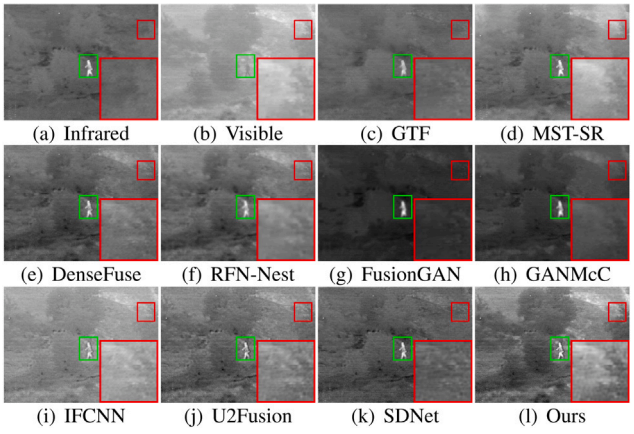

Fig. 13. Visualized results of SeAFusion compared with 9 state-of-the-art algorithms on *Tree_4915* from the TNO dataset. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

visible image dataset, a fusion algorithm that can withstand thermal radiation interference will achieve better segmentation performance. One can focus on the sky of the *00275D* scene to find this phenomenon. Moreover, a practical fusion method could integrate the semantic information from source images to improve the ability of the segmentation model for understanding scenarios. The advantage of our method in enhancing semantic information is exemplified by the traffic signs in the *00119D* scenario.

4.5.2. Detection performance

The performance of object detection, a general high-level computer vision task, also reflects well the semantic information integrated into the fused images. Therefore, a state-of-the-art detector, *i.e.*, YOLOv5 [62] is employed to evaluate object detection performance on the fused images. We randomly select 80 images from the MFNet dataset as the test set, which almost describe all urban scenarios. We manually annotate these images with two critical categories, *i.e.*, person and car. Infrared images, visible images and fused results of various fusion methods are directly input into the YOLOv5 detector, respectively. The mean average precision (mAP) is utilized to measure the detection performance. The detection-driven evaluation results are presented in Table 2, where mAP@0.5, mAP@0.7 and mAP@0.9 indicate the mAP values at IoU thresholds of 0.5, 0.7 and 0.9, respectively, and mAP@[0.5:0.95] denotes the average of all mAP values at different IoU thresholds (from 0.5 to 0.95 in steps of 0.05).

From the results, we can find that infrared images have the best mAP values on the person at almost all IoU thresholds, which means infrared images could offer the detector sufficient semantic information about salient targets (*e.g.*, person). However, the detection results of infrared images on the car are disappointing. Fortunately, visible images can provide the detector with a great deal of semantic information about the car. Different fusion algorithms can integrate the complementary information of infrared and visible images, so that the detection performance of fused images on the car is satisfactory. However, the irrelevant information in the fusion process will weaken the salient target, so the fused images give degraded detection results on the person compared to the infrared images. It is worth emphasizing that our semantic-aware fusion framework can achieve intensity maintenance and texture preservation guided by the content loss and defense against interference from meaningless information with the guidance of the semantic loss. Thus, our detection performance on the person only lags behind infrared images by a narrow margin. And our fused results lead the pack in terms of car detection and average detection accuracy.

Moreover, we also provide some visualized examples in Fig. 18 to illustrate the advantages of our fusion algorithm in facilitating object detection. In *00479D* scene, the detector fails to detect the pedestrians from the visible image due to illumination factors. While GTF and FusionGAN cannot maintain the sharpened edges of pedestrian, DenseFuse and RFN-Nest weaken the salient targets due to the interference of negative information. Therefore, the detector is also unable to detect the person from the fused images generated by the above methods. On the contrary, our approach can fully integrate the semantic information in source images while maintaining the optimal intensity distribution and preserving abundant texture details. Hence, the detector detects all objects from our fused image and improves the confidence compared to the source images. A similar phenomenon occurs in *00689N* scenario. It is worth noting that our fused results significantly improve the confidence of detection results, which indicates that our fused images could provide more semantic information for the detector.

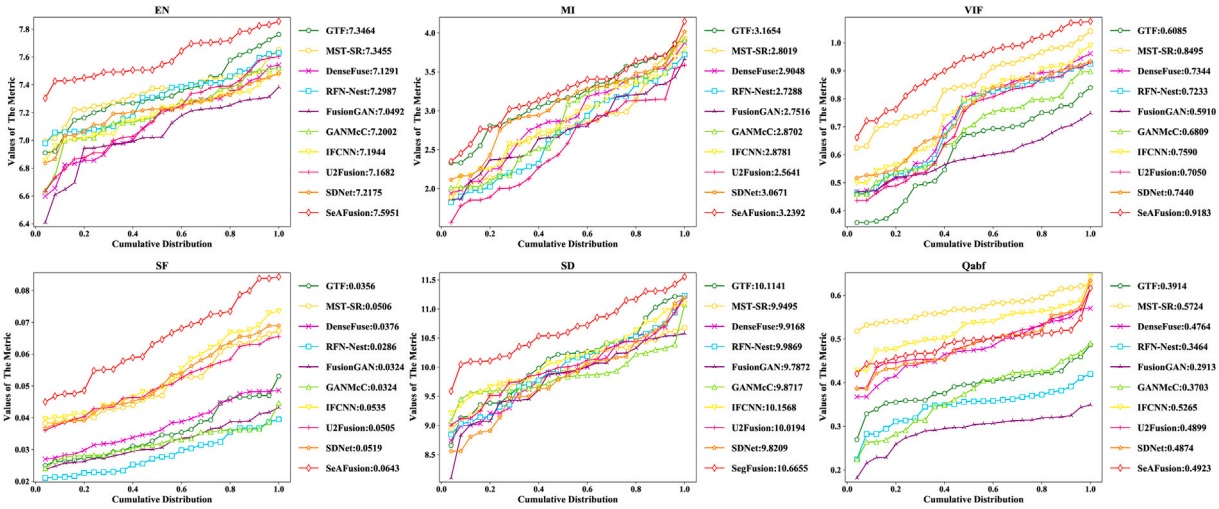

Fig. 14. Quantitative comparisons of the six metrics, *i.e.*, EN, MI, VIF, SF, SD and Q_{abf}, on 25 image pairs from the RoadScene dataset. A point (x, y) on the curve denotes that there are $100 * x$ percent of image pairs which have metric values no more than y.

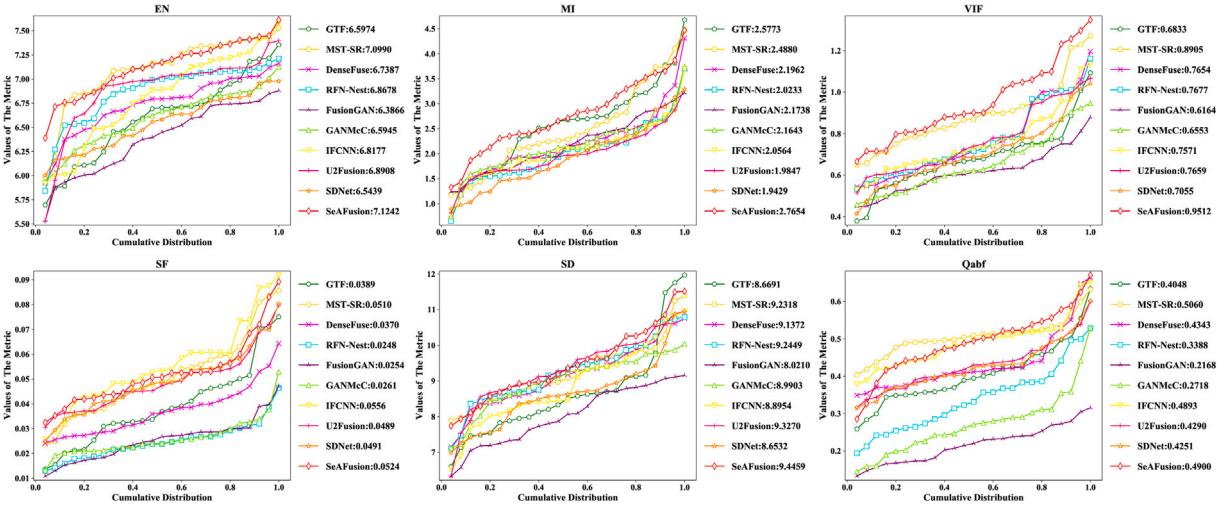

Fig. 15. Quantitative comparisons of the six metrics, *i.e.*, EN, MI, VIF, SF, SD and Q_{abf}, on 25 image pairs from the TNO dataset. A point (x, y) on the curve denotes that there are $100 * x$ percent of image pairs which have metric values no more than y.

Table 3
Mean and standard deviation of the running times of all methods on the MFNet, RoadScene and TNO datasets (unit: second, RED indicates the best result and *BLUE* represents the second best result).

	MFNet	RoadScene	TNO
GTF	5.4498 ± 1.1383	3.7036 ± 2.6401	3.3130 ± 2.0961
MST-SR	0.6269 ± 0.0271	0.3147 ± 0.0786	0.6206 ± 0.3617
DenseFuse	0.2829 ± 0.1532	0.6065 ± 0.0804	0.6791 ± 0.2956
RFN-Nest	0.1924 ± 0.0901	0.1147 ± 0.0224	0.1951 ± 0.0979
FusionGAN	0.0681 ± 0.1090	0.4251 ± 0.0449	0.3267 ± 0.3286
GANMcC	0.1333 ± 0.1985	0.7171 ± 0.1932	0.6339 ± 0.6393
IFCNN	0.0160 ± 0.0781	*0.0080 ± 0.0014*	*0.0123 ± 0.0059*
U2Fusion	0.1352 ± 0.1350	0.7483 ± 0.0929	0.5507 ± 0.5186
SDNet	*0.0154 ± 0.1105*	0.1795 ± 0.0538	0.1465 ± 0.1628
SeAFusion	**0.0115 ± 0.1081**	**0.0060 ± 0.0025**	**0.0049 ± 0.0017**

4.6. Efficiency comparison

As mentioned above, our fusion model is a light-weight network, which can achieve real-time image infusion. To this end, we provide the average running times of different algorithms in Table 3 to demonstrate our efficiency advantage. One can find that all deep learning-based algorithms present noticeable running efficiency, benefiting from the GPU acceleration. In addition, our SeAFusion is the fastest method on all three datasets. We attribute this superiority to two factors: the light-weight network design and the PyTorch-based algorithm implementation. Therefore, our algorithm could be easily deployed as a pre-processing module for high-level vision tasks.

4.7. Ablation studies

4.7.1. Semantic loss analysis

Our light-weight fusion network could boost the semantic information in the fused images with the guidance of semantic loss. In order to verify the particular role of semantic loss, we devise an ablation study on semantic loss. More specifically, we train a fusion model only guided by the content loss. Some typical examples are displayed in Fig. 19. We can notice that the fusion network cannot purposefully preserve the meaningful information of source images without the guidance of semantic loss. This is specifically manifested by smoothed texture details in background regions and the weakened salient targets. We also provide segmentation results in Table 4, where **Without Semantic Loss** indicates that we only utilize the content loss to train the fusion network, to demonstrate the vital role of semantic loss for facilitating

Fig. 16. Segmentation results for infrared, visible and fused images from the MFNet dataset. The segmentation models are re-trained on infrared, visible and fused image sets. Each two rows represent a scene, and from top to bottom is: *00127D*, *00504D* and *01066N*.

high-level vision tasks. We only present IoU for the person, car and bike and mIoU for all categories. One can find that the segmentation performance on fused images degrades significantly without the guidance of semantic loss. In contrast, our SeAFusion achieves salient target intensity maintenance and texture preservation while effectively improving the segmentation performance on fused images.

4.7.2. Gradient residual dense block analysis

Another critical component in our fusion network is the GRDB, which reinforces the description capability of the network for fine-grained details. To this end, we also implement an ablation study on GRDB and present the visualized results in Fig. 19. In the ablation experiment, we remove the residual gradient stream from GRDB. From the visualized examples, we can find that the fused images could maintain a proper intensity distribution, but fail to effectively preserve texture details in the background. On the contrary, the proposed SeAFusion simultaneously maintains the intensity distribution of prominent targets and enhances the description of texture details.

In conclusion, our SeAFusion purposely achieves intensity distribution retention and texture detail preservation while promoting the segmentation performance on fused images, which is benefited from our special designs, *i.e.*, the semantic loss and gradient residual dense block.

Table 4
The segmentation performance of ablation studies and different training strategies. RED indicates the best result and *BLUE* represents the second best result.

	Person	Car	Bike	**mIoU**
Without Semantic Loss	71.31	89.35	70.78	76.19
One-stage Joint Training	*71.66*	*89.69*	*71.23*	*77.41*
Pre-trained on Visible	60.20	88.99	69.91	74.01
Pre-trained on Infrared	70.16	86.83	68.42	70.90
SeAFusion	**74.3**	**90.43**	**72.18**	**78.64**

4.8. Training strategy comparison

We propose a joint low-level and high-level adaptive training strategy to improve the fusion and segmentation performance across the board. In this section, we design training strategy comparison experiments to demonstrate the effectiveness of the proposed joint low-level and high-level adaptive training strategy. More specifically, besides the proposed training strategy, the one-stage joint training strategy is utilized for training the fusion network. Moreover, the segmentation models, trained respectively on infrared and visible image sets, are deployed to guide the training of the fusion network. Hence, three fusion models trained with different training strategies are leveraged as

Fig. 17. Segmentation results for infrared, visible and fused images from the MFNet dataset. The segmentation model is Deeplabv3+, pre-trained on the Cityscapes dataset. Each two rows represent a scene, and from top to bottom is: *00275D* and *00119D*.

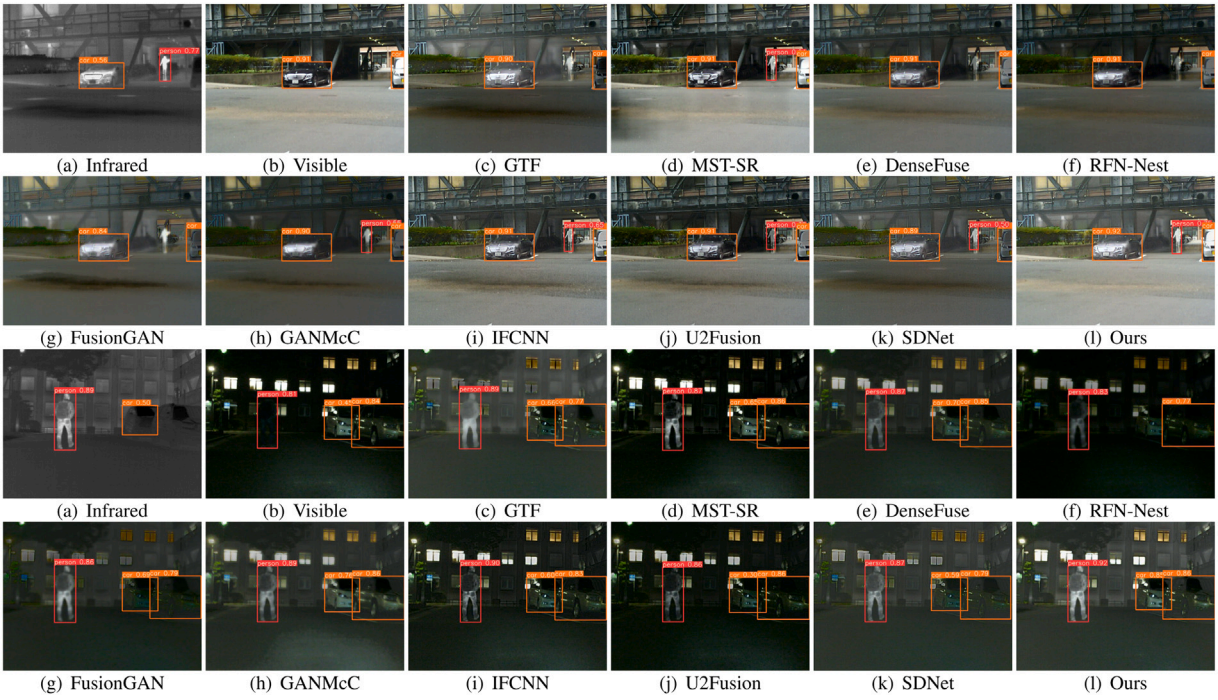

Fig. 18. Object detection results for infrared, visible and fused images from the MFNet dataset. The YOLOv5 detector, pre-trained on the Coco dataset is deployed to achieve object detection. Each two rows represent a scene, and from top to bottom is: *00479D* and *00689N*.

alternatives. The fusion results and segmentation results are presented in Fig. 19 and Table 4, where **Without Semantic Loss** indicates that we only utilize the content loss to train the fusion network, **Pre-trained on Infrared** and **pre-trained on Visible** represent that we leverage the segmentation models trained on infrared and visible image datasets to guide the training of fusion model, respectively, **One-stage Joint Training** means one-stage joint training strategy is employed to jointly train the fusion network and segmentation model. One can notice that the model trained with the one-stage joint training strategy improves the segmentation performance but degrades the visual quality of fused images. In addition, the fusion model guided by the segmentation model, which is trained on the visible image set, neglects the prominent targets. Similarly, using a segmentation model trained on infrared images to guide the training of the fusion network will result in losing texture detail information.

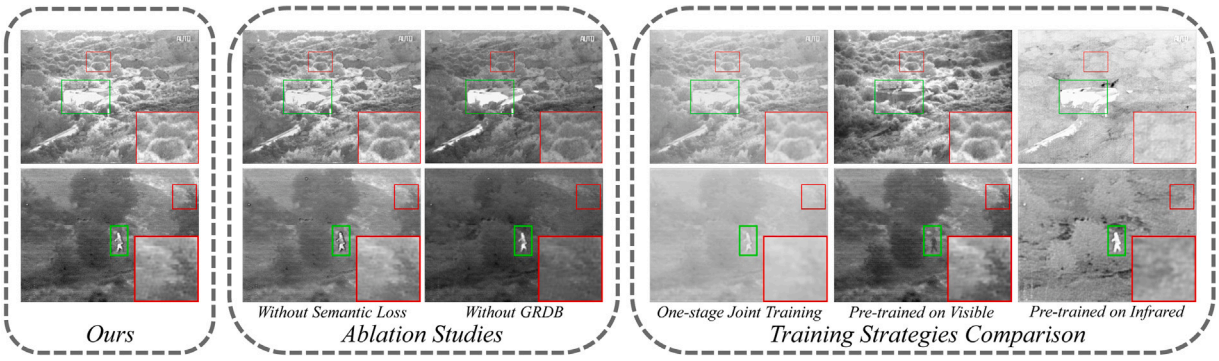

Fig. 19. Visualized results of ablation studies and different training strategies.

5. Conclusion

In this study, a semantic-aware image fusion framework termed SeAFusion is proposed to achieve real-time infrared and visible image fusion. On the one hand, a gradient residual dense block is devised to boost the description ability of the fusion network for fine-grained details. Combined with the elaborate content loss, our fusion network effectively achieves salient target intensity maintenance and texture detail preservation. On the other hand, we introduced a semantic loss to improve the facilitation of fused results for high-level vision tasks. More specifically, the semantic loss allows high-level semantic information to flow back to the image fusion module, which benefits high-level vision tasks in achieving superior performance on the fused results. Moreover, we proposed a joint low-level and high-level adaptive training strategy to achieve simultaneously impressive performance in both image fusion and various high-level vision tasks. Extensive comparative and generalization experiments demonstrate the superiority of our SeAFusion over state-of-the-art competitors in both subjective effect and quantitative metrics. In addition, abundant task-driven evaluation experiments reveal the natural strengths of our framework in facilitating high-level vision tasks. Furthermore, the remarkable advantage in running efficiency allows our algorithm to be easily deployed as a pre-processing module for high-level vision tasks.

CRediT authorship contribution statement

Linfeng Tang: Conceptualization, Methodology, Experiment, Writing. **Jiteng Yuan:** Experiment. **Jiayi Ma:** Conceptualization, Methodology, Revised the manuscript.

Declaration of competing interest

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

Acknowledgment

This research was sponsored by the National Natural Science Foundation of China (61773295).

References

[1] H. Zhang, H. Xu, X. Tian, J. Jiang, J. Ma, Image fusion meets deep learning: A survey and perspective, Inf. Fusion 76 (2021) 323–336.
[2] Y. Cao, D. Guan, W. Huang, J. Yang, Y. Cao, Y. Qiao, Pedestrian detection with unsupervised multispectral feature learning using deep neural networks, Inf. Fusion 46 (2019) 206–217.
[3] C. Li, C. Zhu, Y. Huang, J. Tang, L. Wang, Cross-modal ranking with soft consistency and noisy labels for robust RGB-T tracking, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 808–823.
[4] Y. Lu, Y. Wu, B. Liu, T. Zhang, B. Li, Q. Chu, N. Yu, Cross-modality person re-identification with shared-specific feature transfer, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020, pp. 13379–13389.
[5] Q. Ha, K. Watanabe, T. Karasawa, Y. Ushiku, T. 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, 2017, pp. 5108–5115.
[6] H. Li, X.-J. Wu, J. Kittler, RFN-nest: An end-to-end residual fusion network for infrared and visible images, Inf. Fusion 73 (2021) 720–786.
[7] H. Zhang, J. Ma, SDNet: A versatile squeeze-and-decomposition network for real-time image fusion, Int. J. Comput. Vis. 129 (10) (2021) 2761–2785.
[8] Z. Zhou, B. Wang, S. Li, M. Dong, Perceptual fusion of infrared and visible images through a hybrid multi-scale decomposition with Gaussian and bilateral filters, Inf. Fusion 30 (2016) 15–26.
[9] H. Li, X. Qi, W. Xie, Fast infrared and visible image fusion with structural decomposition, Knowl.-Based Syst. 204 (2020) 106182.
[10] J. Ma, Y. Zhou, Infrared and visible image fusion via gradientlet filter, Comput. Vis. Image Underst. 197–198 (2020) 103016.
[11] Y. Liu, J. Jin, Q. Wang, Y. Shen, X. Dong, Region level based multi-focus image fusion using quaternion wavelet and normalized cut, Signal Process. 97 (2014) 9–30.
[12] X. Liu, W. Mei, H. Du, Structure tensor and nonsubsampled shearlet transform based algorithm for CT and MRI image fusion, Neurocomputing 235 (2017) 131–139.
[13] Q. Zhang, X. Maldague, An adaptive fusion approach for infrared and visible images based on NSCT and compressed sensing, Infrared Phys. Technol. 74 (2016) 11–20.
[14] J. Chen, X. Li, L. Luo, X. Mei, J. Ma, Infrared and visible image fusion based on target-enhanced multiscale transform decomposition, Inform. Sci. 508 (2020) 64–78.
[15] H. Li, X.-J. Wu, J. Kitler, MDLatLRR: A novel decomposition method for infrared and visible image fusion, IEEE Trans. Image Process. 29 (2020) 4733–4746.
[16] Y. Liu, X. Chen, R.K. Ward, Z.J. Wang, Image fusion with convolutional sparse representation, IEEE Signal Process. Lett. 23 (12) (2016) 1882–1886.
[17] N. Cvejic, D. Bull, N. Canagarajah, Region-based multimodal image fusion using ICA bases, IEEE Sens. J. 7 (5) (2007) 743–751.
[18] J. Mou, W. Gao, Z. Song, Image fusion based on non-negative matrix factorization and infrared feature extraction, in: Proceedings of the International Congress on Image and Signal Processing, 2013, pp. 1046–1050.
[19] Z. Fu, X. Wang, J. Xu, N. Zhou, Y. Zhao, Infrared and visible images fusion based on RPCA and NSCT, Infrared Phys. Technol. 77 (2016) 114–123.
[20] J. Ma, C. Chen, C. Li, J. Huang, Infrared and visible image fusion via gradient transfer and total variation minimization, Inf. Fusion 31 (2016) 100–109.
[21] J. Ma, Z. Zhou, B. Wang, H. Zong, Infrared and visible image fusion based on visual saliency map and weighted least square optimization, Infrared Phys. Technol. 82 (2017) 8–17.
[22] H. Li, X.-J. Wu, Densefuse: A fusion approach to infrared and visible images, IEEE Trans. Image Process. 28 (5) (2018) 2614–2623.
[23] H. Li, X.-J. Wu, T. Durrani, Nestfuse: An infrared and visible image fusion architecture based on nest connection and spatial/channel attention models, IEEE Trans. Instrum. Meas. 69 (12) (2020) 9645–9656.
[24] Y. Zhang, Y. Liu, P. Sun, H. Yan, X. Zhao, L. Zhang, IFCNN: A general image fusion framework based on convolutional neural network, Inf. Fusion 54 (2020) 99–118.
[25] H. Xu, J. Ma, Z. Le, J. Jiang, X. Guo, Fusiondn: A unified densely connected network for image fusion, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2020, pp. 12484–12491.
[26] J. Ma, L. Tang, M. Xu, H. Zhang, G. Xiao, STDFusionNet: An infrared and visible image fusion network based on salient target detection, IEEE Trans. Instrum. Meas. 70 (2021) 5009513.

[27] J. Ma, W. Yu, P. Liang, C. Li, J. Jiang, FusionGAN: A generative adversarial network for infrared and visible image fusion, Inf. Fusion 48 (2019) 11–26.
[28] J. Ma, P. Liang, W. Yu, C. Chen, X. Guo, J. Wu, J. Jiang, Infrared and visible image fusion via detail preserving adversarial learning, Inf. Fusion 54 (2020) 85–98.
[29] J. Ma, H. Xu, J. Jiang, X. Mei, X.-P. Zhang, DDcGAN: A dual-discriminator conditional generative adversarial network for multi-resolution image fusion, IEEE Trans. Image Process. 29 (2020) 4980–4995.
[30] H. Zhang, J. Yuan, X. Tian, J. Ma, GAN-FM: Infrared and visible image fusion using GAN with full-scale skip connection and dual Markovian discriminators, IEEE Trans. Comput. Imaging 7 (2021) 1134–1147.
[31] M. Haris, G. Shakhnarovich, N. Ukita, Task-driven super resolution: Object detection in low-resolution images, 2018, arXiv preprint arXiv:1803.11316.
[32] Y. Pei, Y. Huang, Q. Zou, Y. Lu, S. Wang, Does haze removal help cnn-based image classification?, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 682–697.
[33] S. Li, I.B. Araujo, W. Ren, Z. Wang, E.K. Tokuda, R.H. Junior, R. Cesar-Junior, J. Zhang, X. Guo, X. Cao, Single image deraining: A comprehensive benchmark analysis, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 3838–3847.
[34] Y. Long, H. Jia, Y. Zhong, Y. Jiang, Y. Jia, RXDNFuse: A aggregated residual dense network for infrared and visible image fusion, Inf. Fusion 69 (2021) 128–141.
[35] W. Wu, D. Zhang, J. Hou, Y. Wang, T. LU, H. Zhou, Semantic guided infrared and visible image fusion, IEICE Trans. Fundam. Electron. Commun. Comput. Sci. (2021).
[36] M. Wu, Y. Ma, F. Fan, X. Mei, J. Huang, Infrared and visible image fusion via joint convolutional sparse representation, J. Opt. Soc. Amer. A 37 (7) (2020) 1105–1115.
[37] Y. Liu, S. Liu, Z. Wang, A general framework for image fusion based on multi-scale transform and sparse representation, Inf. Fusion 24 (2015) 147–164.
[38] S. Li, X. Kang, L. Fang, J. Hu, H. Yin, Pixel-level image fusion: A survey of the state of the art, Inf. Fusion 33 (2017) 100–112.
[39] H. Xu, H. Zhang, J. Ma, Classification saliency-based rule for visible and infrared image fusion, IEEE Trans. Comput. Imaging 7 (2021) 824–836.
[40] Y. Liu, X. Chen, J. Cheng, H. Peng, A medical image fusion method based on convolutional neural networks, in: Proceedings of the International Conference on Information Fusion, 2017, pp. 1–7.
[41] H. Zhang, H. Xu, Y. Xiao, X. Guo, J. 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, 2020, pp. 12797–12804.
[42] H. Xu, J. Ma, J. Jiang, X. Guo, H. Ling, U2fusion: A unified unsupervised image fusion network, IEEE Trans. Pattern Anal. Mach. Intell. 44 (1) (2022) 502–518.
[43] J.Y. Zhu, T. Park, P. Isola, A.A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2223–2232.
[44] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, J. Choo, Stargan: Unified generative adversarial networks for multi-domain image-to-image translation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8789–8797.
[45] H. Xu, P. Liang, W. Yu, J. Jiang, J. Ma, Learning a generative model for fusing infrared and visible images via conditional generative adversarial network with dual discriminators, in: Proceedings of the International Joint Conference on Artificial Intelligence, 2019, pp. 3954–3960.
[46] J. Li, H. Huo, C. Li, R. Wang, Q. Feng, AttentionFGAN: Infrared and visible image fusion using attention-based generative adversarial networks, IEEE Trans. Multimed. 23 (2020) 1383–1396.
[47] J. Ma, H. Zhang, Z. Shao, P. Liang, H. Xu, GANMcC: A generative adversarial network with multiclassification constraints for infrared and visible image fusion, IEEE Trans. Instrum. Meas. 70 (2021) 5005014.
[48] B. Li, X. Peng, Z. Wang, J. Xu, D. Feng, Aod-net: All-in-one dehazing network, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 4770–4778.
[49] D. Liu, B. Wen, X. Liu, Z. Wang, T.S. Huang, When image denoising meets high-level vision tasks: A deep learning approach, in: 27th International Joint Conference on Artificial Intelligence, International Joint Conferences on Artificial Intelligence, 2018, pp. 842–848.
[50] D. Liu, B. Wen, J. Jiao, X. Liu, Z. Wang, T.S. Huang, Connecting image denoising and high-level vision tasks via deep learning, IEEE Trans. Image Process. 29 (2020) 3695–3706.
[51] M. Guo, M. Chen, C. Ma, Y. Li, X. Li, X. Xie, High-level task-driven single image deraining: Segmentation in rainy days, in: Proceedings of the International Conference on Neural Information Processing, 2020, pp. 350–362.
[52] C. Peng, T. Tian, C. Chen, X. Guo, J. Ma, Bilateral attention decoder: A lightweight decoder for real-time semantic segmentation, Neural Netw. 137 (2021) 188–199.
[53] A. Toet, TNO image fusion dataset, 2014, http://dx.doi.org/10.6084/m9.figshare.1008029.v1, URL https://figshare.com/articles/dataset/TNO_Image_Fusion_Dataset/1008029.
[54] J.W. Roberts, J.A. Van Aardt, F.B. Ahmed, Assessment of image fusion procedures using entropy, image quality, and multispectral classification, J. Appl. Remote Sens. 2 (1) (2008) 023522.
[55] G. Qu, D. Zhang, P. Yan, Information measure for performance of image fusion, Electron. Lett. 38 (7) (2002) 313–315.
[56] Y. Han, Y. Cai, Y. Cao, X. Xu, A new image fusion performance metric based on visual information fidelity, Inf. Fusion 14 (2) (2013) 127–135.
[57] A.M. Eskicioglu, P.S. Fisher, Image quality measures and their performance, IEEE Trans. Commun. 43 (12) (1995) 2959–2965.
[58] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al., Pytorch: An imperative style, high-performance deep learning library, in: Advances in Neural Information Processing Systems, 2019, pp. 8026–8037.
[59] K. Ram Prabhakar, V. Sai Srikar, R. Venkatesh Babu, DeepFuse: A deep unsupervised approach for exposure fusion with extreme exposure image pairs, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 4714–4722.
[60] L.C. Chen, Y. Zhu, G. Papandreou, F. Schroff, H. Adam, Encoder-decoder with atrous separable convolution for semantic image segmentation, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 801–818.
[61] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, B. Schiele, The cityscapes dataset for semantic urban scene understanding, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 3213–3223.
[62] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: Unified, real-time object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 779–788.