Information Fusion 83-84 (2022) 79–92

Contents lists available at ScienceDirect

Information Fusion

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

Full length article

PIAFusion: A progressive infrared and visible image fusion network based on illumination aware

Linfeng Tang[1], Jiteng Yuan[1], Hao Zhang, Xingyu Jiang, Jiayi Ma *

Electronic Information School, Wuhan University, Wuhan 430072, China

ARTICLE INFO

Keywords:
Image fusion
Illumination aware
Cross-modality differential aware fusion
Deep learning

ABSTRACT

Infrared and visible image fusion aims to synthesize a single fused image containing salient targets and abundant texture details even under extreme illumination conditions. However, existing image fusion algorithms fail to take the illumination factor into account in the modeling process. In this paper, we propose a progressive image fusion network based on illumination-aware, termed as PIAFusion, which adaptively maintains the intensity distribution of salient targets and preserves texture information in the background. Specifically, we design an illumination-aware sub-network to estimate the illumination distribution and calculate the illumination probability. Moreover, we utilize the illumination probability to construct an illumination-aware loss to guide the training of the fusion network. The cross-modality differential aware fusion module and halfway fusion strategy completely integrate common and complementary information under the constraint of illumination-aware loss. In addition, a new benchmark dataset for infrared and visible image fusion, *i.e.*, Multi-Spectral Road Scenarios (available at https://github.com/Linfeng-Tang/MSRS), is released to support network training and comprehensive evaluation. Extensive experiments demonstrate the superiority of our method over state-of-the-art alternatives in terms of target maintenance and texture preservation. Particularly, our progressive fusion framework could round-the-clock integrate meaningful information from source images according to illumination conditions. Furthermore, the application to semantic segmentation demonstrates the potential of our PIAFusion for high-level vision tasks. Our codes will be available at https://github.com/Linfeng-Tang/PIAFusion.

1. Introduction

Information captured by a single modal sensor or under a single shooting setting cannot effectively and comprehensively describe the imaging scene since the theoretical and technical limitations of hardware devices [1]. Therefore, image fusion techniques are emerging, which aim to combine the complementary information shot by multi-modal sensors or under different shooting settings. According to the existence or absence of modal difference, image fusion divides into multi-modal image fusion and digital photographic image fusion. In multi-modal image fusion missions, the infrared and visible image fusion has been broadly used in military actions, object detection [2], tracking [3], pedestrian re-identification [4] and semantic segmentation [5] since the sufficiently complementary nature of source images. A typical application of infrared and visible image fusion is presented in Fig. 1. One can notice that the infrared image can effectively highlight thermal targets (*e.g.*, pedestrians) but neglect other objects (*e.g.*, static vehicles and bikes) since the infrared image captures thermal radiation emitted by objects. Hence, the detector could not detect bikes from the infrared image and misidentifies the bicycle as a pedestrian. On the contrary, the visible image captures reflected information, hence bicycles can be detected but the salient targets hidden in the dark or smoke are ignored. It is worth noting that the detector could detect all pedestrians and more bikes in the fused image since the fused image adequately integrates the complementary information of source images, compared with the single modal image.

In the past decades, numerous image fusion techniques have been developed, including traditional approaches [8–10] and data-driven methods [7,11,12]. The traditional methods exploit mathematical transformation to transform source images into the transform domain and perform activity level measurements and design fusion rules in the transform domain to achieve image fusion. Traditional image fusion techniques include multi-scale decomposition-based methods [13,14],

* Corresponding author.
E-mail addresses: linfeng0419@gmail.com (L. Tang), yuanjiteng@whu.edu.cn (J. Yuan), zhpersonalbox@gmail.com (H. Zhang), jiangx.y@whu.edu.cn (X. Jiang), jyma2010@gmail.com (J. Ma).
[1] The first two authors Linfeng Tang and Jiteng Yuan contributed equally to the work.

https://doi.org/10.1016/j.inffus.2022.03.007
Received 17 August 2021; Received in revised form 26 November 2021; Accepted 18 March 2022
Available online 29 March 2022

Fig. 1. Application of infrared and visible image fusion to object detection. From left to right: detecting results of infrared image, visible image and fused image.

subspace clustering-based methods [15], sparse representation-based methods [16], optimization-based methods [17], and hybrid methods [18]. Images synthesized by the above methods could satisfy the demands of subsequent tasks in specific scenarios. However, the development of traditional methods is currently experiencing a bottleneck. On the one hand, the transformations or representations employed by traditional approaches have become increasingly complex in order to achieve more impressive fusion performance, which cannot respond to the requirements of real-time computer applications [19]. On the other hand, the hand-crafted activity level measurement and fusion rules fail to accommodate sophisticated scenarios.

In recent years, the prosperity of deep learning motivates the image fusion community to explore data-driven image fusion schemes. According to the utilized baselines, the mainstream data-driven methods can be roughly divided into three categories, *i.e.*, auto-encoder (AE) based methods [6,20,21], convolutional neural network (CNN) based methods [22–24] and generative adversarial network (GAN) based methods [7,25,26]. The AE-based methods first train an auto-encoder on the large nature image datasets as the feature extractor and image reconstructor. Then, the feature extractor is employed to extract complementary information from multi-modal images, and the specific fusion rules, *e.g.* concatenation [24], element-wise addition [27], element-wise weight summation [6] and element-wise maximum [28] are utilized to merge those features. Finally, the image reconstructor undertakes the burden of reconstructing the fused image from fused features. However, the AE-based fusion framework is not fully learnable since the hand-crafted fusion rules are exploited to merge the deep features. Therefore, other researchers focus on exploring the end-to-end CNN-based image fusion networks, which rely on superior network structures and elaborate loss functions to ensure fusion performance. Given that the image fusion tasks lack ground truth, some works attempt to define image fusion as a game between the generator and discriminator. More specifically, they force the fused image with abundant texture details by constraining the probability distribution between the fused image and source images. It should be emphasized that too strong constraints may introduce artificial textures into the fused image.

Although data-driven image fusion methods could synthesize relatively satisfying fused results, there are still some obstacles that deserve to be emphasized. First of all, the infrared and visible image fusion community currently lacks a large benchmark dataset for training robust fusion networks. The mainstream datasets, *i.e.*, the TNO dataset [29] and the RoadScene dataset [30] include a few image pairs with simple scenes, in particular the TNO dataset. A fusion network trained on those datasets is prone to overfitting and fails to cope with more complex scenarios. Therefore, developing a new benchmark dataset with a great number of image pairs for infrared and visible image fusion is promising. In addition, a crucial issue exists, *i.e.*, the illumination imbalance has never been investigated. Illumination imbalance means the difference of lighting conditions between the daytime and nighttime scenes [31]. Intuitively, we provide an example to present the illumination imbalance in Fig. 2, one can observe that the visible image has clearer texture detail than the infrared image, but the infrared image could highlight the distinct pedestrians in the daytime. Comparatively, the infrared image provides more salient targets and has more abundant textures than the visible image at nighttime. Unfortunately, the existing methods [7,17,26] generally assume that textures only exist in visible images, which is reasonable in daytime scenarios. However, such assumptions may cause fused images to be suffered from losing texture details at nighttime. Moreover, AE-based approaches [6] utilize inappropriate fusion rules to merge the deep features, which may weaken the texture details and salient targets.

Finally, the existing techniques focus more on how to fusion images/features but ignore when to fusion images/features. More specifically, researchers work on designing more elaborate fusion rules and loss functions to improve fusion performance. But what stage to merge images/features is rarely investigated. Currently, fusion occurs in two phases, *i.e.*, input fusion and halfway fusion. Input fusion means cascading the multi-modal inputs together as the input of fusion networks [7,26], which may lead to the network failing to merge the semantic information of source images. Halfway fusion implies using a specific fusion rule to merge deep features extracted by the feature extractor [6,20,22,24]. In this case, the selected fusion rules are critical since an inappropriate rule cannot integrate complementary information adequately.

To address the above issues, we propose a progressive infrared and visible image fusion framework based on illumination-aware, known as PIAFusion and release a large benchmark dataset for infrared and visible image fusion. First, we collect numerous infrared and visible image pairs from the existing MFNet dataset [5] which is shot for multi-modal image semantic segmentation. The MFNet dataset contains a great number of infrared images with low contrast and low signal-to-noise ratio and some misaligned image pairs. Therefore, we sharpen and enhance the infrared images and remove the unaligned image pairs. Subsequently, we design an illumination-aware sub-network to evaluate the illumination conditions. More specifically, the pre-trained illumination-aware sub-network calculates the probability that the current scenario is day or night. Then, these probabilities are utilized to construct the illumination-aware loss for guiding the training of the fusion network. Finally, a progressive feature extractor containing cross-modality differential aware fusion (CMDAF) modules is employed to extract and merge complementary information in multi-modal images adequately. Afterwards, the complementary and common features are fused via a halfway fusion manner, and an image reconstructor transforms the fused features to the fused image domain. With the guidance of illumination-aware loss, our fusion framework could adaptively integrate vital information such as distinctive targets and texture details according to illumination conditions. Moreover, the CMDAF module and the halfway fusion strategy allow our network to merge the common and complementary information at various stages. As a result, our fused images could fully maintain textures while emphasizing prominent targets regardless of illumination conditions, which can be observed from Fig. 2. To sum up, our major contributions are four-fold:

- We propose a novel illumination-guided infrared and visible image fusion framework, which can round-the-clock fuse the meaningful information of source images through perceiving illumination situations.
- We combine the cross-modality differential aware fusion module with the halfway fusion strategy to integrate complementary and common information at various stages.
- We construct a new benchmark dataset for the training and evaluation of infrared and visible image fusion, termed as Multi-Spectral Road Scenarios (MSRS), on the basis of the MFNet dataset. It is available at https://github.com/Linfeng-Tang/MSRS.
- Extensive experiments demonstrate the superiority of our algorithm over state-of-the-art competitors. Compared with the alternatives, our method could adaptively fuse the complementary and common information according to illumination conditions.

The remainder of this paper is organized as follows. Section 2 briefly describes the related works of image fusion and illumination-aware-based computer vision applications. In Section 3, we introduce

Fig. 2. An example of illumination imbalance. From left to right: infrared image, visible image, the fused results of DenseFuse [6], FusionGAN [7], and our proposed PIAFusion. The visible image contains abundant information, such as texture details in the daytime (top row). But salient targets and textures are all included in the infrared image at nighttime (bottom row). Existing methods ignore the illumination imbalance issues, causing detail loss and thermal target degradation. Our algorithm can adaptively integrate meaningful information according to illumination conditions.

our proposed PIAFusion in detail, including the problem analysis, loss functions and network architectures. Section 4 illustrates the fusion performance of our method in comparison with other alternatives on three benchmark datasets, 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 including traditional image fusion methods, AE-based image fusion methods, CNN-based image fusion methods and GAN-based image fusion methods. Then, some vision missions based on illumination-aware are briefly introduced.

2.1. Infrared and visible image fusion

2.1.1. Traditional image fusion

Feature extraction, fusion and reconstruction are three essential elements of typically traditional image fusion methods. Moreover, the key to these algorithms lies in feature extraction and fusion since feature reconstruction is the inverse operation of feature extraction.

Numerous feature extraction functions have been proposed to achieve a better feature representation, which can be roughly divided into three categories, *i.e.*, multi-scale transforms, sparse representation and subspace clustering. Multi-scale transforms such as Laplacian pyramid, discrete wavelet [32], shearlet [33], curvelet [34] and non-subsampled contourlet [35] transforms are the most classical feature extraction techniques. Sparse representation including joint sparse representation [36], convolutional sparse representation [16] and latent low-rank representation [13], is another feature extraction method, which is consistent with human visual perception. Subspace clustering, *e.g.*, independent component analysis [15], principal component analysis [37], and non-negative matrix factorization [38] can extract mutually independent subcomponents. In addition, some specific fusion rules, including element-wise addition [27], element-wise weight summation [6] and element-wise maximum [28], are applied to integrate the extracted features/representations.

Beyond the typical approaches, some optimization-based methods also offer a new perspective for the image fusion community. In particular, gradient transfer fusion (GTF) has laid a solid foundation for CNN-based methods and GAN-based methods, which defines the objective function of image fusion as the overall intensity fidelity and texture structure preservation [17]. In addition, some hybrid models combine the advantages of different components to improve fusion performance [18,39]. Specifically, Hou et al. proposed a novel infrared and visible image fusion via combining non-subsampled shearlet transform (NSST), visual saliency and multi-objective artificial bee colony (MOABC) optimizing spiking cortical mode (SCM), which could alleviate some issues such as edge blurring, low contrast and loss of details [39].

2.1.2. AE-based image fusion

With the excellent nonlinear fitting ability of the neural network, deep learning has become the new favorite for various computer vision tasks. The image fusion community has also explored data-driven methods, and one of the successful examples is the approach based on auto-encoder. AE-based methods inherit the baseline of typically traditional image fusion algorithms, *i.e.*, feature extraction, fusion and reconstruction. Prabhakar et al. creatively achieved feature extraction via two convolutional layers with fixed parameters [27]. The feature reconstruction was also replaced by three pre-trained convolutional layers. Moreover, the element-wise addition fusion rule was employed to merge the feature maps.

Nevertheless, two convolutional layers fail to extract features containing semantic information. Li et al. strengthened the encoder and introduced dense connection to extract deep features and achieve features reuse [6]. Except for dense connection, Li et al. also introduced the multi-scale encoder–decoder network architectures and nest connection to extract more comprehensive features [20,21]. Unfortunately, the hand-crafted fusion rules severely restrict the improvement of fusion performance.

2.1.3. CNN-based image fusion

To avoid the limitations of hand-crafted fusion rules, researchers have developed end-to-end CNN-based image fusion algorithms. Zhang et al. explored an end-to-end framework to maintain the proportion of gradient and intensity through the intensity and gradient paths [40]. They modeled a general loss function and adjusted the hyper-parametric for different image fusion missions. In order to prompt the network to extract and merge the meaningful features purposefully, Ma et al. defined desired information for infrared and visible image fusion via introducing a salient target mask [24]. Their network can simultaneously fuse infrared and visible images and achieve salient target detection. Moreover, Xu et al. trained a unified model for multi-fusion tasks with elastic weight consolidation considering the cross-fertilization between different image fusion missions [30]. However, the fusion network cannot exhibit its full potential performance since the specificity of fusion tasks, *i.e.*, without ground truth.

2.1.4. GAN-based image fusion

Since the adversarial loss constrains networks at the distribution level, generative adversarial networks are perfectly suited for unsupervised tasks such as image fusion [7,41] and image-to-image translation [42,43]. Ma et al. innovatively viewed the infrared and visible image fusion as a game between the generator and discriminator. The discriminator forces the generator to synthesize fused images containing more textures [7]. However, a single discriminator tends to break the balance of data distributions between infrared and visible images. Therefore, Ma et al. further proposed a dual-discriminator conditional

generative adversarial network (DDcGAN) to maintain the balance of distribution between different source images [26]. Subsequently, Li et al. integrated a multi-scale attention mechanism into the GAN-based fusion framework [44]. The dual-discriminator can focus more on the attention regions with the guidance of attention loss function. Moreover, Ma et al. proposed a GAN-based fusion method with multi-classification, which transforms image fusion into a multi-distribution simultaneous estimation [45]. The multi-classifier prompts the fused images to have significant contrast and rich texture in a more balanced manner. However, the image fusion community still faces many challenges, such as the existing benchmark datasets only contain limited scenarios, and both traditional methods and data-driven approaches ignore illumination variations.

Noteworthy, there is a vital priori in the practical multi-modal datasets, *i.e.*, visible images contain a great deal of meaningful information in the daytime, while infrared images capture richer information at nighttime. Unfortunately, existing image fusion approaches do not take this priori into account in the modeling process. In particular, CNN-based and GAN-based methods define infrared and visible image fusion as intensity maintenance of infrared images and texture preservation of visible images. This assumption is reasonable in the daytime. Nevertheless, fused images should retain more texture details of infrared images to enhance the description of scenes at nighttime. Therefore, a robust fusion algorithm should perceive the illumination conditions and adaptively integrate meaningful information with the guidance of illumination.

2.2. Illumination aware-based vision applications

In fact, several practical computer vision applications have taken illumination factors into account when modeling. Wang et al. proposed a global illumination-aware and detail-preserving network (GLADNet) for low-light image enhancement [46]. GLADNet first calculates a global illumination estimation for the low-light image, then adjusts the illumination guided by the estimation and uses a concatenation with the original image to enhance details. In addition, Sakkos et al. developed a triple multi-task generative adversarial network to integrate features with varying illumination into the segmentation branch, which vastly improves the performance of foreground segmentation [47]. Furthermore, many researchers explored the feasibility of exploiting illumination information to boost multispectral pedestrian detection performance. For instance, Li et al. proposed an illumination-aware faster R-CNN to adaptively fuse infrared and visible sub-networks via a gate function defined over the output of illumination-aware network [48]. Coincidentally, Guan et al. presented a multispectral pedestrian detection framework based on illumination-aware pedestrian detection and semantic segmentation [49]. They utilized a novel illumination-aware weighting mechanism to describe lighting conditions and integrated the illumination information into dual-stream CNN to obtain human-related features under diverse illumination situations. Moreover, MBNet promotes the optimization process and improves the performance of the detector in a more flexible and balanced manner [31]. In particular, the illumination aware feature alignment module is utilized to select the complementary information according to illumination conditions adaptively.

In this work, we propose a progressive infrared and visible fusion framework based on illumination-aware to eliminate illumination imbalance issues. More specifically, we first develop an illumination-aware sub-network to estimate illumination conditions. Then, the illumination probability is exploited to construct the illumination-aware loss. Finally, the elaborate progressive network adaptively integrates meaningful information of infrared and visible images with the guidance of illumination-aware loss.

3. Methodology

In this section, we comprehensively describe the illumination aware based progressive infrared and visible image fusion framework. Firstly, we provide the problem analysis of our proposed PIAFusion. Then, we provide the illumination-aware loss of the fusion network and the cross-entropy loss of the illumination-aware sub-network in detail. Finally, the architectures of progressive fusion network and illumination-aware sub-network are presented.

3.1. Problem analysis

Given an infrared image I_{ir} and a visible image I_{vi}, the fused image I_f can be generated via feature extraction, integration and reconstruction. In order to improve fusion performance, we design an illumination-aware loss to constrain the above three steps. The progressive illumination-guided infrared and visible image fusion network is presented in Fig. 3.

Considering that illumination imbalance influences information distribution, we develop an illumination-aware sub-network to estimate the illumination of visible images. Given a visible image I_{vi}, the illumination-aware process is defined as:

$$\left\{P_d, P_n\right\} = N_{IA}(I_{vi}), \tag{1}$$

where N_{IA} indicates the illumination-aware sub-network, P_d and P_n denote the probability that an image belongs to the day and night, respectively. It is worth mentioning that P_d and P_n are non-negative scalars. Since most information is concentrated in visible images in the daytime, and infrared images contain more meaningful information at nighttime, the illumination probability reflects the richness of information in source images from the side. Therefore, we exploit the illumination probability to calculate illumination-aware weight that represents the contribution of source images via an illumination assignment mechanism. To simplify the calculation, our illumination assignment mechanism employs a simple normalization function defined as follows:

$$\begin{aligned} W_{ir} &= \frac{P_n}{P_d + P_n}, \\ W_{vi} &= \frac{P_d}{P_d + P_n}, \end{aligned} \tag{2}$$

where W_{ir} and W_{vi} represent the contribution of the infrared image and visible image to the fusion process, respectively.

In addition, we also design a progressive fusion network to integrate complementary and common information completely. More specifically, we first apply a feature encoder E_F to extract deep features from infrared and visible images, which can be represented as,

$$\left\{F_{ir}, F_{vi}\right\} = \left\{E_F(I_{ir}), E_F(I_{vi})\right\}, \tag{3}$$

where F_{ir} and F_{vi} denote infrared features and visible features. Moreover, we define F_{ir}^i and F_{vi}^i as features extracted by the ith convolutional layer from the infrared and visible image, respectively.

Furthermore, we propose a cross-modality differential aware fusion (CMDAF) module to compensate for differential information. In particular, the F_{ir}^i and F_{vi}^i can be formulated with the common part and complementary part as follows:

$$\begin{aligned} F_{ir}^i &= \frac{F_{ir}^i + F_{ir}^i}{2} + \frac{F_{vi}^i - F_{vi}^i}{2} = \frac{F_{ir}^i + F_{vi}^i}{2} + \frac{F_{ir}^i - F_{vi}^i}{2}, \\ F_{vi}^i &= \frac{F_{vi}^i + F_{vi}^i}{2} + \frac{F_{ir}^i - F_{ir}^i}{2} = \frac{F_{vi}^i + F_{ir}^i}{2} + \frac{F_{vi}^i - F_{ir}^i}{2}, \end{aligned} \tag{4}$$

where the common part represents the common features and the complementary part reflects the complementary features of different modalities. Eq. (4) explains the principle of differential decomposition, which is analogous to the differential amplification circuit. The key idea of CMDAF module is to fully integrate complementary information

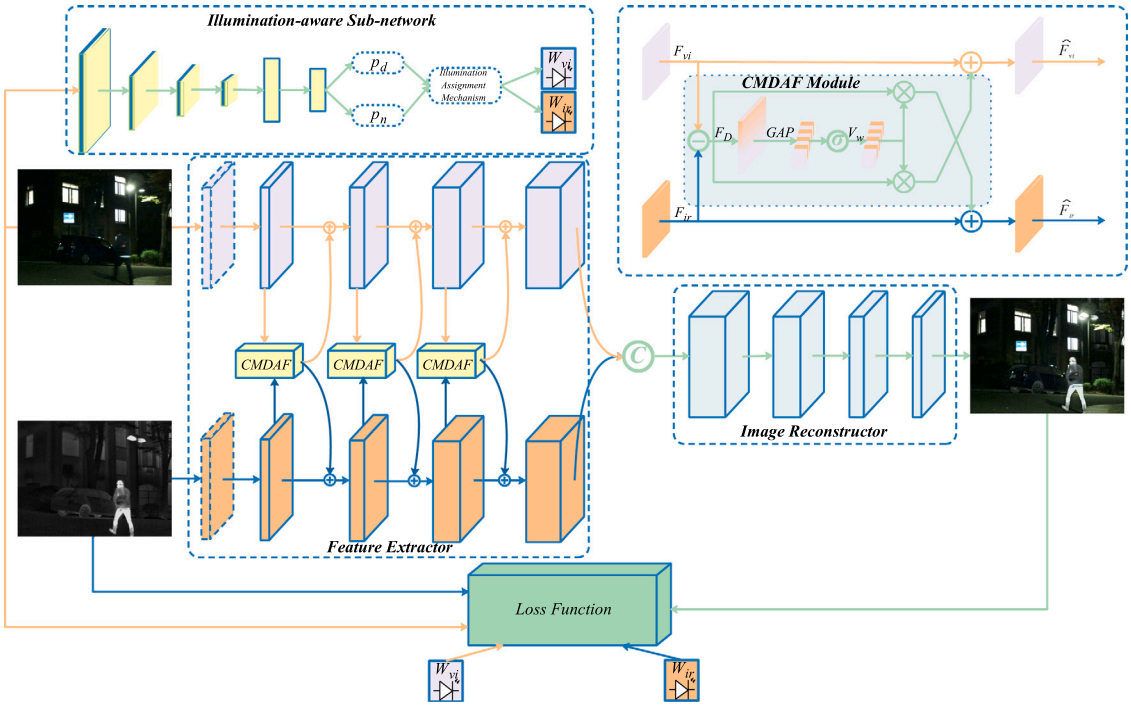

Fig. 3. The overall framework of the progressive infrared and visible image fusion algorithm based on illumination-aware.

with channel weighting. Therefore, the CMDAF module can be exactly defined as:

$$\begin{aligned}\hat{F}_{ir}^{i} &= F_{ir}^{i} \oplus \delta(GAP(F_{vi}^{i} - F_{ir}^{i})) \odot (F_{vi}^{i} - F_{ir}^{i}),\\ \hat{F}_{vi}^{i} &= F_{vi}^{i} \oplus \delta(GAP(F_{ir}^{i} - F_{vi}^{i})) \odot (F_{ir}^{i} - F_{vi}^{i}),\end{aligned} \tag{5}$$

where $\oplus$ refers to element-wise summation, $\odot$ indicates channel-wise multiplication, $\delta(\cdot)$ and $GAP(\cdot)$ denote the sigmoid function and Global Average Pooling, respectively. Eq. (5) means a global average pooling compresses the complementary features into a vector. Then, the vector is normalized to $[0, 1]$ via sigmoid function to generate the channel weights. Finally, the complementary features are multiplied by the channel weights, and the results are added to the original features as modal supplementary information. Therefore, our feature encoder can extract and pre-integrate complementary features of different modality with the CMDAF module.

In addition, the common and complementary features of infrared and visible images are completely merged via a halfway fusion manner, *i.e.*, concatenation. The halfway fusion strategy is expressed as follows:

$$F_f = C(F_{ir}, F_{vi}), \tag{6}$$

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 an image reconstructor $\mathcal{R}_I$, which is presented as Eq. (7):

$$I_f = \mathcal{R}_I(F_f). \tag{7}$$

3.2. Loss function

3.2.1. Loss function of progressive fusion network

In order to promote that our progressive fusion framework adaptively integrates meaningful information according to illumination conditions, we innovatively propose the illumination-aware loss. The illumination-aware loss $\mathcal{L}_{illum}$ is accurately defined as follows:

$$\mathcal{L}_{illum} = W_{ir} \cdot \mathcal{L}_{int}^{ir} + W_{vi} \cdot \mathcal{L}_{int}^{vi}, \tag{8}$$

where $\mathcal{L}_{int}^{ir}$ and $\mathcal{L}_{int}^{vi}$ denote the intensity loss of infrared and visible images, respectively. W_{ir} and W_{vi} are the illumination-aware weights, defined in Eq. (2). The specific definition of intensity loss terms is as follows.

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:

$$\begin{aligned}\mathcal{L}_{int}^{ir} &= \frac{1}{HW}\left\|I_f - I_{ir}\right\|_1,\\ \mathcal{L}_{int}^{vi} &= \frac{1}{HW}\left\|I_f - I_{vi}\right\|_1,\end{aligned} \tag{9}$$

where H and W are the height and width of the input image, respectively, $\|\cdot\|_1$ stands for the l_1-norm. In fact, the intensity distribution of the fused image should be consistent with different source images according to illumination situations. Hence, we use illumination-aware weights, *i.e.*, W_{ir} and W_{vi} to adjust the intensity constraints of the fused image.

The illumination loss drives the progressive fusion network to dynamically preserve intensity information from source images based on illumination conditions, yet it does not maintain an optimal intensity distribution for fused images. To this end, we further introduce the auxiliary intensity loss, which is represented as:

$$\mathcal{L}_{aux} = \frac{1}{HW}\left\|I_f - \max(I_{ir}, I_{vi})\right\|_1, \tag{10}$$

where $\max(\cdot)$ denotes the element-wise maximum selection.

Moreover, we expect the fused image to maintain the best intensity distribution and preserve abundant texture details simultaneously. Based on extensive experiments, we find that the optimal texture of the fused image can be expressed as the maximum aggregate of infrared and visible image textures. Therefore, a texture loss is introduced to force the fused image to contain more texture information, which is defined as follows:

$$\mathcal{L}_{texture} = \frac{1}{HW}\left\|\left|\nabla I_f\right| - max(\left|\nabla I_{ir}\right|, \left|\nabla I_{vi}\right|)\right\|_1, \tag{11}$$

where ∇ indicates the gradient operator which measures the texture information of an image. In this paper, the Sobel operator is exploited to compute the gradient. $|\cdot|$ refers to the absolute operation.

Finally, the full objective function of progressive fusion network is a weighted combination of the illumination loss, auxiliary intensity loss

Table 1
Kernel size, output channels and activation function of all convolutional layers in the progressive fusion network.

	Feature Extractor			Image Reconstructor		
	Kernel Size	Output Channels	Activation Function	Kernel Size	Output Channels	Activation Function
Layer1	1 × 1	16	Leaky Relu	3 × 3	256	Leaky Relu
Layer2	3 × 3	16	Leaky Relu	3 × 3	128	Leaky Relu
Layer3	3 × 3	32	Leaky Relu	3 × 3	64	Leaky Relu
Layer4	3 × 3	64	Leaky Relu	3 × 3	32	Leaky Relu
Layer5	3 × 3	128	Leaky Relu	1 × 1	1	Tanh

and texture loss, which is expressed as follows:

$$\mathcal{L}_{fusion} = \lambda_1 \cdot \mathcal{L}_{illum} + \lambda_2 \cdot \mathcal{L}_{aux} + \lambda_3 \cdot \mathcal{L}_{texture}, \tag{12}$$

In conclusion, our progressive fusion network can dynamically retain the optimal intensity distribution according to the illumination scenario with the guidance of illumination loss and auxiliary intensity loss. And it could obtain the ideal texture detail guided by the texture loss. Therefore, the fusion network can round-the-clock fuse the meaningful information from infrared and visible images.

3.2.2. Loss function of illumination-aware sub-network

In our method, the fusion performance heavily relies on the accuracy of the illumination-aware sub-network. The illumination-aware network is a classifier essentially, which calculates the probability of an image belonging to the daytime and nighttime. Therefore, we employ the cross-entropy loss $\mathcal{L}_{IAN}$ to constrain the training process of the illumination-aware sub-network, which is represented as follows:

$$\mathcal{L}_{IAN} = -z \log \sigma(y) - (1 - z) \log(1 - \sigma(y)), \tag{13}$$

where z denotes the illumination label of the input image, $y = [P_n, P_d]$ indicates the output of the illumination-aware sub-network, and σ refers to the softmax function, which normalizes the illumination probability to $[0, 1]$.

3.3. Network architecture

3.3.1. Progressive fusion network architecture

As illustrated in Fig. 3, we employ the end-to-end CNN-based framework as our backbone and our progressive network consists of the feature extractor and image reconstructor.

There are five convolutional layers in the feature extractor that aims to extract complementary and common features adequately. Firstly, the 1 × 1 convolutional layer is designed to reduce the modal difference between infrared and visible images. Thus, we train the first layer for infrared and visible images separately. Afterwards, four convolutional layers with shared weights are employed to extract the deep features of infrared and visible images. It is worth noting that the output of the 2nd, 3rd and 4th layers is followed by a CMDAF module to exchange modal complementary features. The CMDAF module enables our network to integrate the complementary information at the feature extraction stage in a progressive manner. Hence, Our feature extractor could completely extract the common and complementary features from infrared and visible images. More details of all convolutional layers in the feature extractor, such as kernel size, output channel and activation function, are presented in Table 1. The kernel size of all convolutional layers is 3 except the first layer. All layers of the feature extractor employ Leaky Relu as the activation function.

Subsequently, the deep features extracted from infrared and visible images are concatenated, which are the input of the image reconstructor. The image reconstructor contains five convolutional layers, which is responsible for fully integrating common and complementary information and generating the fused image. The detailed configuration of the image reconstructor is exhibited in Table 1. The kernel size of all layers is 3 × 3 except the last layer, whose kernel size is 1 × 1. Moreover, the image reconstructor gradually reduces the channels number of feature maps in the image reconstruction process. All convolutional layers in the image reconstructor adopt Leaky Relu as the activation function except the last layer, whose activation function is Tanh.

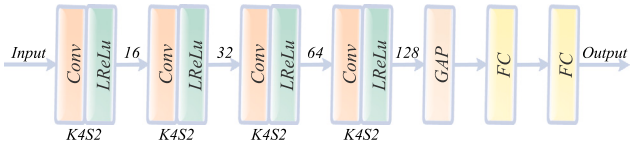

Fig. 4. Illumination-aware sub-network architecture. K4S2 indicates that the convolutional kernel size is 4 and stride is set to 2, LReLU denotes Leaky Relu activation function, GAP refers to Global Average Pooling and FC represents Fully Connected Layer.

As we know that information loss is a catastrophic issue in the image fusion process. Hence, the padding is set as same in all convolutional layers of the progressive fusion network, and the stride is set to 1 except for the first and last layers. Consequently, our network does not introduce any down-sampling, and the size of fused images is consistent with source images.

3.3.2. Illumination-aware sub-network architecture

The illumination-aware sub-network aims to estimate the illumination distribution of the scenario, whose input is visible images and output is illumination probability. The architecture of the illumination-aware sub-network is presented in Fig. 4. It consists of four convolutional layers, a global average pooling and two fully connected layers. The 4 × 4 convolutional layer with stride setting to 2 compresses spatial information and extracts illumination information. All convolutional layers employ Leaky Relu as the activation function and set the padding to same. Then, a global average pooling operation is utilized to integrate illumination information. Finally, two fully connected layers calculate the illumination probabilities from illumination information.

4. Experimental validation

In this section, we first present the dataset construction, experimental configurations and implementation details. Then, some comparative experiments and generalization experiments are performed to demonstrate the superiority of our algorithm. Next, we analyze the running efficiency of different methods. In addition, we demonstrate not only the ability of our approach to adaptive feature selection, but also the potential of our approach for high-level vision tasks. Some ablation studies are performed to verify the effectiveness of our specific designs, including the illumination-aware loss and cross-modality differential aware fusion module.

4.1. Dataset construction

We construct a new multi-spectral dataset for infrared and visible image fusion based on the MFNet dataset [5]. The MFNet dataset contains 1,569 image pairs (820 taken at the daytime and 749 taken at nighttime) with spatial resolution is 480 × 640. However, There are many misaligned image pairs in the MFNet dataset and most infrared

images are low signal-to-noise and low contrast. To this end, we first collect 715 daytime image pairs and 729 nighttime image pairs via removing 125 misaligned image pairs. Moreover, an image enhancement algorithm based on dark channel prior [50] is leveraged to optimize the contrast and signal-to-noise of infrared images. As a result, the released new Multi-Spectral Road Scenarios (MSRS) dataset contains 1,444 pairs of aligned infrared and visible images with high quality.

4.2. Experimental configurations

To completely evaluate our proposed method, we performed extensively qualitative and quantitative experiments on the MSRS, RoadScene [30], TNO [29] datasets. We compare our algorithm with nine state-of-the-art methods, including two traditional methods, *i.e.*, GTF [17], and MDLatLRR [13], three AE-based method, *i.e.*, DenseFuse [6], DRF [51], and CSF [11], one GAN-based methods, *i.e.*, FusionGAN [7], three CNN-based methods, *i.e.,* IFCNN [22] PMGI [40], and U2Fusion [30]. The implementations of all these nine methods are publicly available, and we set the parameters as reported in the original papers.

Four evaluation metrics are selected to quantify the evaluation, including mutual information (MI) [52], standard deviation (SD), visual information fidelity (VIF) [53] and Q_{abf}. The MI metric measures the amount of information transferred from the source images to the fused image from the information theory perspective. The SD metric reflects the distribution and contrast of the fused image from the statistical perspective. The VIF metric evaluates the information fidelity of the fused image from the perspective of the human visual system. The Q_{abf} measures the amount of edge information transferred from source images to the fused image. Moreover, the larger MI, SD, VIF and Q_{abf} indicate better fusion performance.

4.3. Implementation details

We train our progressive fusion model and illumination-aware model on the MSRS dataset. We select 427 images of daytime scenes and 376 images of night scenes to train the illumination-aware sub-network. The crop-and-decompose data augmentation is employed to generate sufficient training data. Specifically, we crop these images into 64×64 patches with stride set to 64. Therefore, we collect a total of 29,960 daytime patches and 26,320 nighttime patches. Moreover, we utilize 376 daytime image pairs, *i.e.*, 26,320 patches and 376 nighttime image pairs, *i.e.*, 26,320 patches to learn parameters of the progressive fusion model. All image patches are normalized to $[0, 1]$ before being fed into networks. We use one-hot labels as the reference of illumination-aware sub-network, and the labels for the daytime scene and nighttime scene is set to 2D vector $[1, 0]$ and $[0, 1]$, respectively.

The illumination-aware sub-network and progressive fusion network are trained in sequence. More specifically, we first train the illumination-aware sub-network. After that, the pre-trained illumination-aware network is utilized to calculate the illumination probabilities and construct the illumination-aware loss when training the progressive fusion network. The batch size is set to b, the train steps in one epoch are set as p, and it takes M epochs to train a model. For the illumination-aware sub-network, we empirically set $b_1 = 128$, $M_1 = 100$, and $p_1 = 438$. For the progressive fusion network, b_2 is set as 64, M_2 is set to 30, and $p_2 = 819$. The model parameters are updated by the Adam optimizer with the learning rate first initialized to 0.001 and then decayed exponentially. For hyper-parameters of Eq. (12), we set as $\lambda_1 = 3$, $\lambda_2 = 7$ and $\lambda_3 = 50$. We further summarize the training procedure of PIAFusion in Algorithm 1. The proposed method is implemented on the TensorFlow platform [54]. All experiments are conducted on an NVIDIA TITAN V GPU and 2.00 GHz Intel (R) Xeon (R) Gold 5117 CPU. It is worth emphasizing that source images are fed directly into the progressive fusion network at the test stage.

Algorithm 1: Training procedure

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

for M_1 *epochs* **do**
  for p_1 *steps* **do**
    Select b_1 visible images $\{I_{vi}^1, I_{vi}^2, \ldots, I_{vi}^{b_1}\}$;
    Generate the illumination probability P_n and P_d with the illumination-aware sub-network N_A;
    Calculate the cross-entropy loss $\mathcal{L}_{IAN}$ according to Eq. (13);
    Update the parameters of the illumination-aware sub-network N_A by Adam Optimizer: $\bigtriangledown_{N_A}(\mathcal{L}_{IAN})$;
  end
end
for $M_2 epochs$ **do**
  for p_2 *steps* **do**
    Select b_2 infrared images $\{I_{ir}^1, I_{ir}^2, \ldots, I_{ir}^{b_2}\}$;
    Select b_2 visible images $\{I_{vi}^1, I_{vi}^2, \ldots, I_{vi}^{b_2}\}$;
    Generate the illumination probability P_n and P_d with the illumination-aware sub-network N_A;
    synthesize the fused images $\{I_f^1, I_f^2, \ldots, I_f^{b_2}\}$ with the progressive fusion network;
    Calculate the total loss $\mathcal{L}_{fusion}$ according to Eq. (12);
    Update the parameters of the progressive fusion network by Adam Optimizer: $\bigtriangledown_{E_F,R_I}(\mathcal{L}_{Illum}(E_F, R_I)$;
  end
end

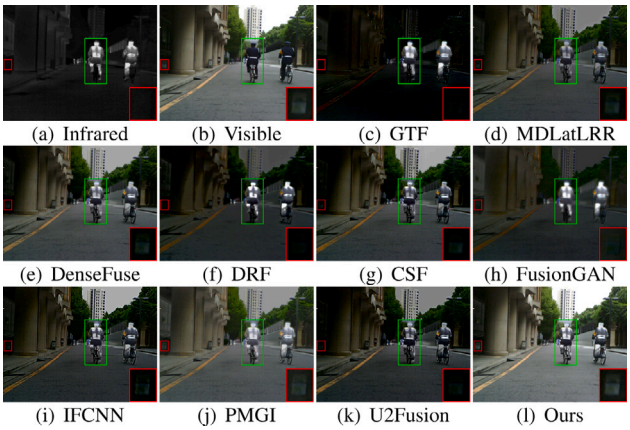

Fig. 5. Qualitative comparison of PIAFusion with 9 state-of-the-art methods on the daytime scene (*00537D*). For a clear comparison, we select a texture region (*i.e.*, the red box) in each image and zoom in it in the bottom right corner and highlight a salient region (*i.e.*, the green box). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

We use a specific fusion strategy [27] to retain color information in fused images since the MSRS and RoadScene dataset contain color visible images. More specifically, we first convert visible images to the YCbCr color space. Then, different fusion methods are employed to merge the Y channel of visible images and infrared images. Finally, the fused image can be converted into the RGB color space via combining with Cb and Cr channels of visible images.

4.4. Comparative experiment

In order to comprehensively evaluate the performance of our method, we compare the proposed PIAFusion with other nine methods on the MSRS dataset.

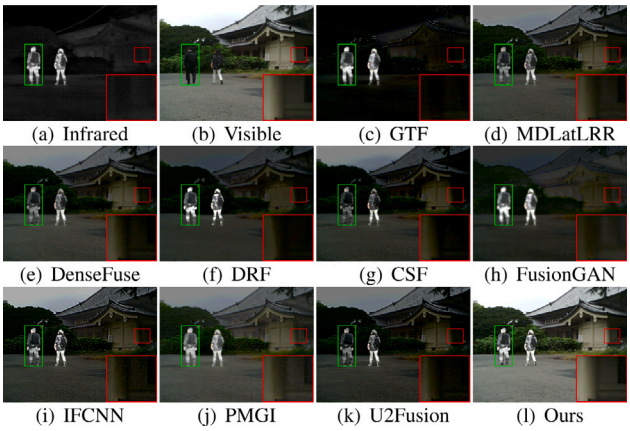

Fig. 6. Qualitative comparison of PIAFusion with 9 state-of-the-art methods on the daytime scene (*00633D*). For a clear comparison, we select a texture region (*i.e.*, the red box) in each image and zoom in it in the bottom right corner and highlight a target area (*i.e.*, the green box). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

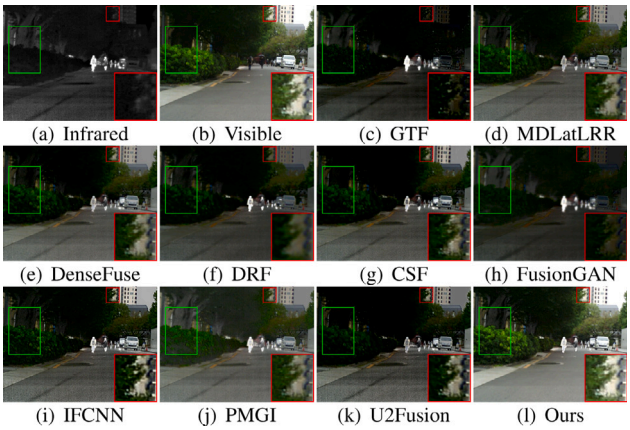

Fig. 7. Qualitative comparison of PIAFusion with 9 state-of-the-art methods on the daytime scene (*00556D*). For a clear comparison, we select a region (*i.e.*, the red box) with abundant texture in each image and zoom in it in the bottom right corner and highlight a salient region (*i.e.*, the green box). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

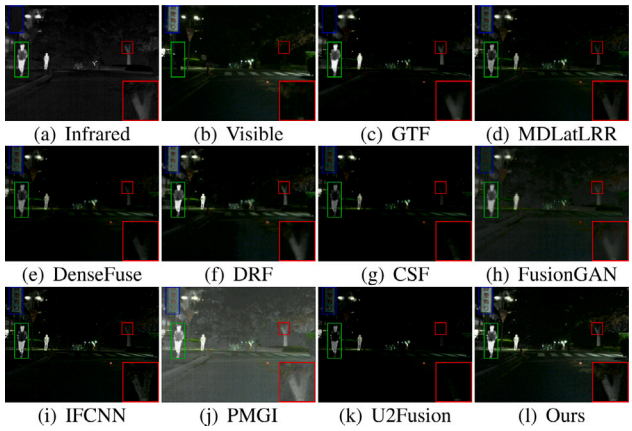

Fig. 8. Qualitative comparison of PIAFusion with 9 state-of-the-art methods on the nighttime scene (*00881N*). For a clear comparison, we select a texture region (*i.e.*, the red box) in each image and zoom in it in the bottom right corner and highlight a detailed area (*i.e.*, the blue box) and a target area (*i.e.*, the green box). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

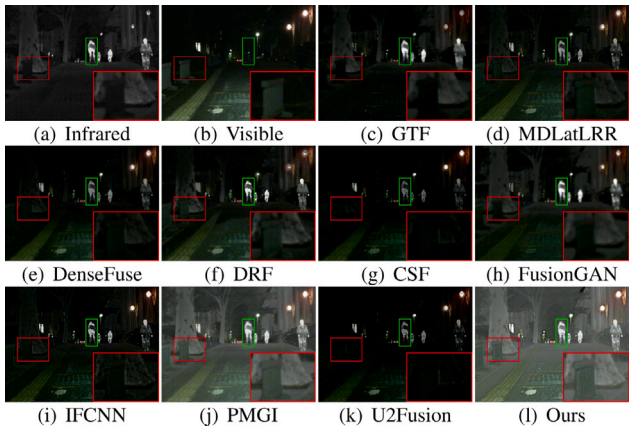

Fig. 9. Qualitative comparison of PIAFusion with 9 state-of-the-art methods on the nighttime scene (*01023N*). For a clear comparison, we select a texture region (*i.e.*, the red box) in each image and zoom in it in the bottom right corner and highlight a target area (*i.e.*, the green box). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

4.4.1. Qualitative results

To intuitively demonstrate the adaptability of the illumination-aware-based fusion method to illumination variation, we select three daytime images and two nighttime images for subjective evaluation. The visualized results are presented in Figs. 5–9.

In the daytime scenes, thermal radiation information of infrared images can be used as supplementary information for visible images. Therefore, a superior fusion algorithm should preserve the texture details of visible images while highlighting salient targets without introducing spectral contamination. As shown in Figs. 5 and 6, GTF and FusionGAN fail to retain detail information of visible images. Although MDLatLRR, DenseFuse, DRF, CSF, IFCNN, and U2Fusion integrate the texture information of visible images with salient target information in infrared images, the background regions suffer from varying degrees of spectral contamination. We zoom in on a background area with the red box to illustrate this phenomenon. Only our method and PMGI completely preserve texture details. However, PMGI, MDLatLRR, DenseFuse and U2Fusion weaken the infrared targets. Moreover, PMGI also introduces additional information into the fused results in some cases, which is shown in the green box of Fig. 7.

In the nighttime scenes, visible images only contain limited detail information. Fortunately, infrared images contain both salient targets and extensive textural details that can complement the textures of visible images. However, it is a challenge to adaptively integrate texture information from both infrared and visible images according to illumination scenarios. In Fig. 8, although all algorithms retain salient targets, MDLatLRR, DenseFuse, CSF, IFCNN and U2Fusion fail to clearly show the tree trunk hidden in the dark. In addition, GTF ignores words, *i.e.*, the blue box, in the visible image, and PMGI introduces noises during the fusion process. FusionGAN and DRF blur the targets and words, respectively. Our method could effectively integrate complementary information of infrared and visible images. A similar phenomenon occurs in Fig. 9. Except for our method and PMGI, all methods cannot simultaneously contain information about the trunk and fence. Therefore, our progressive fusion can adaptively fuse meaningful information according to illumination conditions, benefiting from the illumination-aware loss.

4.4.2. Quantitative results

The quantitative results of four complementary metrics on 150 image pairs are presented in Table 2. It is worth noting that our method exhibits significant superiority in all four metrics. The best MI metric means PIAFusion transfers the most information from source images

Table 2
Quantitative comparisons of the four metrics, *i.e.*, MI, SD, SF, Q_{abf}, on 150 image pairs from the MSRS dataset. RED indicates the best result and *BLUE* represents the second best result.

	MI	SD	VIF	Q_{abf}
GTF	2.1719 ± 0.2346	6.3358 ± 1.1538	0.6217 ± 0.0113	0.3406 ± 0.0034
MDLatLrr	2.5924 ± 0.1750	7.5274 ± 2.7317	0.7354 ± 0.0057	0.5349 ± 0.0034
DenseFuse	*2.7184 ± 0.2145*	7.8514 ± 2.8166	*0.8195 ± 0.0122*	0.4960 ± 0.0071
DRF	2.1372 ± 0.0893	7.7047 ± 1.6202	0.6287 ± 0.0130	0.1615 ± 0.0019
CSF	2.3958 ± 0.1218	7.3685 ± 2.3912	0.7341 ± 0.0089	0.3961 ± 0.0051
FusionGAN	1.9067 ± 0.1005	5.9565 ± 1.0401	0.5477 ± 0.0207	0.1427 ± 0.0021
IFCNN	2.0389 ± 0.1247	7.0242 ± 2.3154	0.7134 ± 0.0080	*0.5388 ± 0.0022*
PMGI	2.2114 ± 0.1423	*7.8803 ± 1.1846*	0.7243 ± 0.0247	0.4233 ± 0.0033
U2Fusion	2.2537 ± 0.2418	7.1052 ± 4.2118	0.6752 ± 0.0114	0.4572 ± 0.0081
PIAFusion	**4.4188 ± 1.3596**	**8.8035 ± 2.8142**	**1.1644 ± 0.0106**	**0.6416 ± 0.0051**

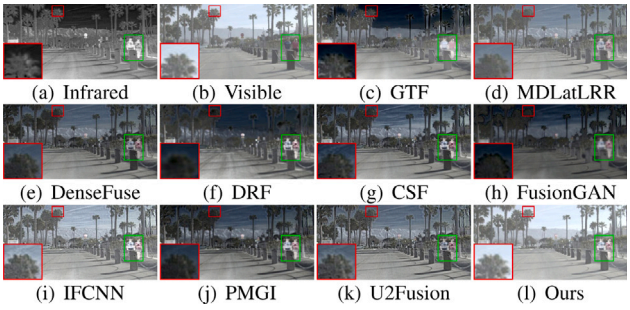

Fig. 10. Qualitative comparison of STDFusionNet with 9 state-of-the-art methods on *FLIR_04269*. For a clear comparison, we select a small area (*i.e.*, the red box) with abundant texture in each image and zoom in it in the bottom right corner and highlight a salient region (*i.e.*, the green box). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

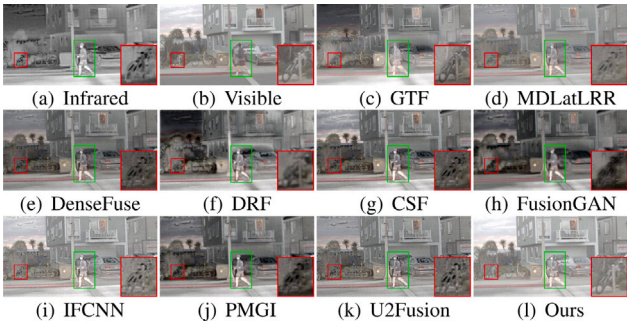

Fig. 11. Qualitative comparison of STDFusionNet with 9 state-of-the-art methods on *FLIR_06307*. For a clear comparison, we select a small area (*i.e.*, the red box) with abundant texture in each image and zoom in it in the bottom right corner and highlight a salient region (*i.e.*, the green box). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

to the fused image according to illumination conditions. Moreover, our proposed method achieves the highest SD and VIF, indicating that our fused images have high contrast and satisfying visual effects. In addition, our PIAFusion shows the best Q_{abf}, which implies that more edge information is preserved in the fused results, benefiting from our proposed cross-modality differential aware fusion module.

4.5. Generalization experiment

It is public knowledge that generalization performance is an important factor in evaluating the data-driven approach. Therefore, we provide generalization experiments performed on the RoadScene and TNO datasets to verify the generalizability of our PIAFusion. Notably, our fusion model is trained on the MSRS dataset and tested directly on the RoadScene and TNO datasets.

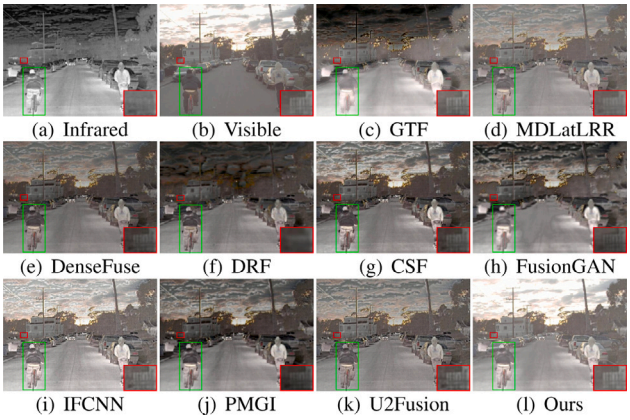

Fig. 12. Qualitative comparison of STDFusionNet with 9 state-of-the-art methods on *FLIR_06570*. For a clear comparison, we select a small area (*i.e.*, the red box) with abundant texture in each image and zoom in it in the bottom right corner and highlight a salient region (*i.e.*, the green box). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

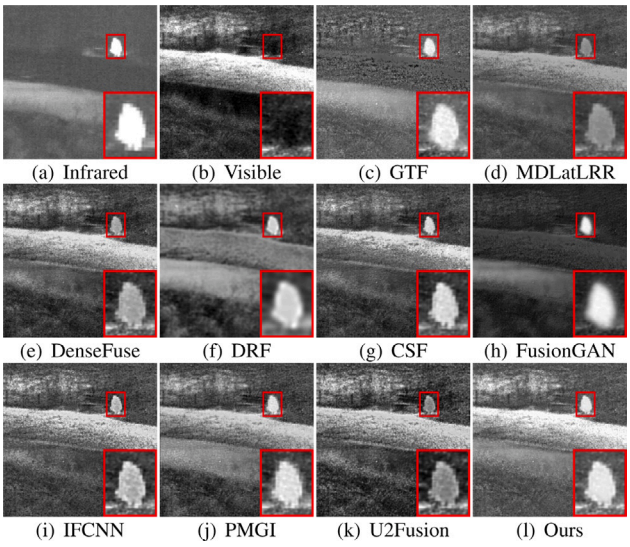

Fig. 13. Visualized results of PIAFusion compared with 9 state-of-the-art algorithms on *bench* scene of the TNO dataset. For a clear comparison, we highlight the salient area (*i.e.*, the red box). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

4.5.1. Qualitative results

The qualitative comparisons of different algorithms on the RoadScene dataset are presented in Figs. 10–12. One can notice that all algorithms maintain the intensity distribution in the area of salient objects. However, GTF and FusionGAN fail to retain the sharp edges

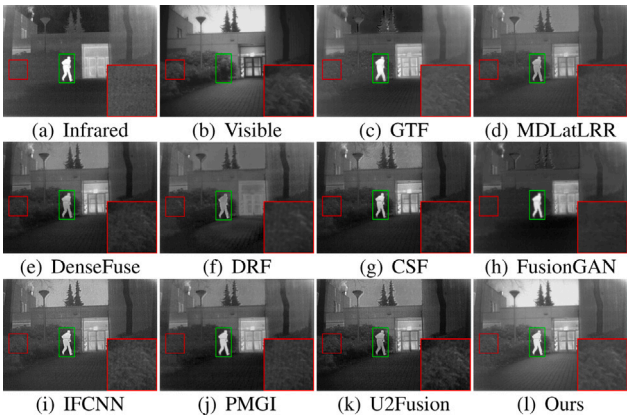

Fig. 14. Visualized results of PIAFusion compared with 9 state-of-the-art algorithms on *Kaptein_1123* scene of the TNO dataset. For a clear comparison, we select a texture area (*i.e.*, the red box) in each image and zoom in it in the bottom right corner and highlight a salient region (*i.e.*, the green box). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

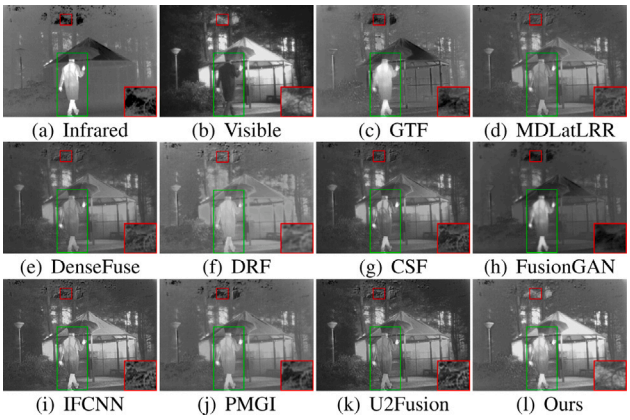

Fig. 15. Visualized results of PIAFusion compared with 9 state-of-the-art algorithms on *Kaptein_1654* scene of the TNO dataset. For a clear comparison, we select a texture area (*i.e.*, the red box) in each image and zoom in it in the bottom right corner and highlight a salient region (*i.e.*, the green box). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

of targets. Moreover, other methods, especially GTF, FusionGAN, DRF, CSF and PMGI, are suffered from varying degrees of spectral contamination, which can be observed from the sky. As shown in the red box in Fig. 10, GTF, IFCNN, and U2Fusion introduce artifacts into fused results, which degrade the visual effect of fused images. In addition, other methods lose texture derails (*e.g.*, the red box in Fig. 11 and Fig. 12) of the visible images because of inadequate information integration. Our approach completely integrates complementary and common features of source images via the CMDAF module and halfway fusion manner. Therefore, the above-mentioned issues, *i.e.*, thermal target degradation, spectral contamination and texture blurring, do not appear in our fused results.

The visualized results of different methods on the TNO dataset are shown in Figs. 13–15. As can be seen from the red box in Fig. 13, MDLatLRR, DenseFuse and U2Fusion weaken the salient target. Moreover, DRF and FusionGAN blur the edge of thermal targets and suffer from serious spectral contamination in the background regions. Only our method, CSF, IFCNN and PMGI successfully maintain the intensity of salient targets and preserve the texture details of visible images. Similar phenomenons can be found in Figs. 14 and 15. It is worth emphasizing that our progressive fusion network effectively preserves the texture details of visible images compared with other alternatives, which can be observed from the red box in Figs. 14 and 15.

Thus, extensive visualized results on various datasets demonstrate the superiority of our algorithm in terms of significant target maintenance and texture preservation. We attribute the advantage to the following aspects. On the one hand, we define the illumination-aware loss to constrain the network to integrate meaningful information according to specific illumination conditions. On the other hand, we design a cross-modality differential aware fusion module to fuse complementary information of infrared and visible images completely.

4.5.2. Quantitative results

We also select 25 image pairs from the RoadScene and TNO datasets for quantitative evaluation. The comparative results of different approaches on the four metrics are presented in Figs. 16 and 17. From Fig. 16, one can find that PIAFusion presents distinctive superiority in MI and VIF metrics. Such results mean our methods could transfer more information from source images to fused images and generate pleasing fused results. On the Q_{abf} metric, our method also ranks first, indicating that PIAFusion preserves sufficient edge information. Moreover, the proposed approach only follows IFCNN by a narrow margin in the SD metric.

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

	MSRS	RoadScene	TNO
GTF	2.9756 ± 0.5128	1.3662 ± 0.4142	3.2837 ± 2.2344
MDLatLRR	123.4549 ± 2.7267	64.2259 ± 12.1389	136.968 ± 66.9715
DenseFuse	0.3039 ± 0.2389	0.6397 ± 0.5337	0.5514 ± 0.5928
DRF	24.461 ± 7.8751	12.4196 ± 0.9145	14.6332 ± 1.5757
CSF	2.8766 ± 0.3786	2.7107 ± 0.7504	3.5572 ± 1.8906
FusionGAN	0.0759 ± 0.1742	0.4602 ± 0.3090	0.3846 ± 0.5103
IFCNN	**0.0214 ± 0.1215**	**0.0724 ± 0.3088**	**0.0744 ± 0.2966**
PMGI	*0.0507 ± 0.1559*	*0.2376 ± 0.3314*	*0.1965 ± 0.3617*
U2Fusion	0.1453 ± 0.2093	0.7537 ± 0.3969	0.5881 ± 0.6956
Ours	0.0895 ± 0.2090	0.8119 ± 0.2989	0.6502 ± 0.9011

As shown in Fig. 17, PIAFusion ranks first in MI, VIF and Q_{abf} metrics by a significant margin. For the SD metric, our method only has comparable performance. This is justified since the TNO dataset mainly contains daytime scenes, and our method prefers to maintain the intensity of visible images in the daytime, thus reducing the contrast of fused images.

In conclusion, both qualitative and quantitative results demonstrate that PIAFusion has great generalization performance. In addition, our proposed approach effectively maintains the intensity distribution of target areas and preserves the texture details of the background region, benefiting from the proposed illumination-aware loss and CMDAF module.

4.6. Efficiency comparison

In order to evaluate different algorithms in a holistic manner, we provide the average running time of different methods in Table 3. One can observe that traditional methods consume more time to synthetic fused images. In particular, MDLatLRR obtains the low-rank representations via multi-scale decomposition, which is extremely time-consuming. In contrast, the data-driven approaches present noticeable advantages in the term of running efficiency, benefiting from GPU acceleration. Especially, IFCNN is the fastest algorithm on all datasets. Our progressive fusion network achieves per-fusion of complementary features via the cross-modality differential aware fusion module and integrates common and complementary information in a concatenation manner. Hence, our PIAFusion is relatively time-consuming. Fortunately, our method still has comparable running efficiency to other approaches.

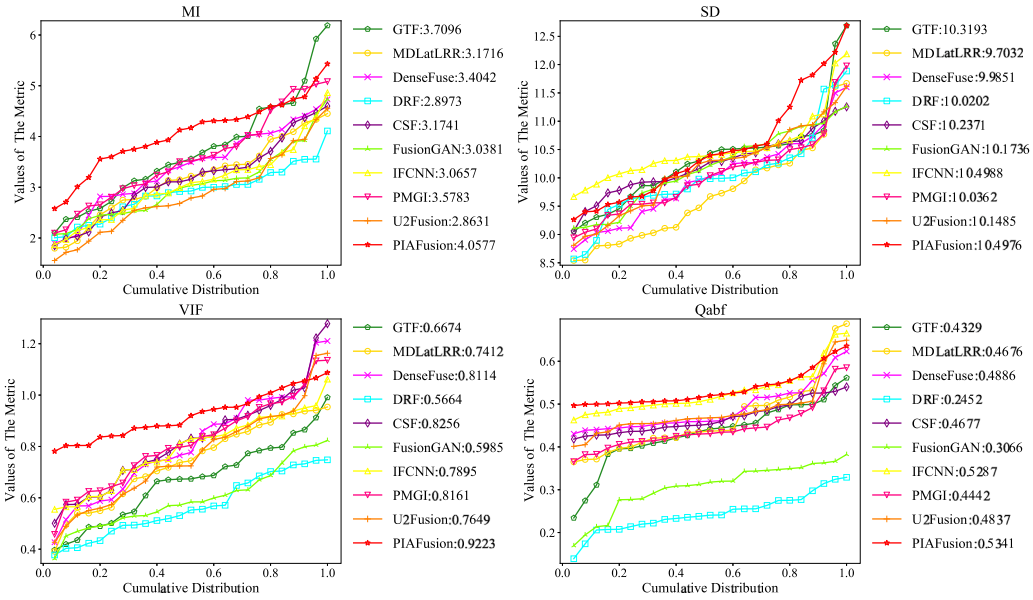

Fig. 16. Quantitative comparisons of the four metrics, *i.e.*, MI, SD, VIF and Q_{abf}, on 25 image pairs from the RoadScene dataset. The nine state-of-the-art methods such as GTF [17], MDLatLRR [13], DenseFuse [6], DRF [51], CSF [11], FusionGAN [7], IFCNN [22], PMGI [40], and U2Fusion [30] and are utilized for comparison. 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.

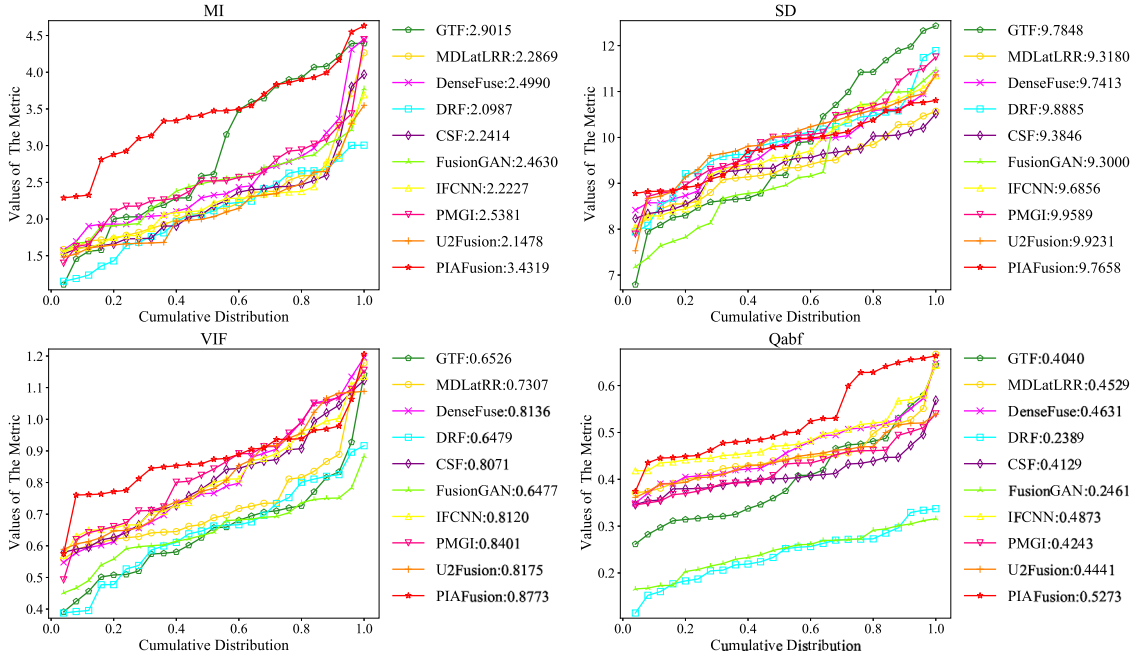

Fig. 17. Quantitative comparisons of the four metrics, *i.e.*, MI, SD, VIF, Q_{abf}, on 25 image pairs from the TNO dataset. The nine state-of-the-art methods such as GTF [17], MDLatLRR [13], DenseFuse [6], DRF [51], CSF [11], FusionGAN [7], IFCNN [22], PMGI [40], and U2Fusion [30] are used for comparison. 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.

4.7. Adaptive feature selection

The proposed framework could implicitly implement feature extraction, feature selection and image reconstruction based on illumination scenarios with the guidance of our elaborate loss function. To intuitively demonstrate the implicit feature selection function of the fusion network according to illumination changes, we provide partial feature maps after the feature integration phase in the daytime and nighttime scenes, respectively, as shown in Figs. 18 and 19. One can observe that our fusion model could purposefully achieve feature selection based on the illumination conditions with the guidance of illumination-aware loss. In the daytime scene, our fusion network is able to preserve the visible features well, while a little bit of information from the infrared features is integrated into the fusion feature maps as a supplement. On the contrary, in the nighttime scenario, the fused features mainly inherit the distribution of infrared feature maps, while a small number of fine-grained details of visible features are also integrated into the fused features. These visual results prove that our neural network can

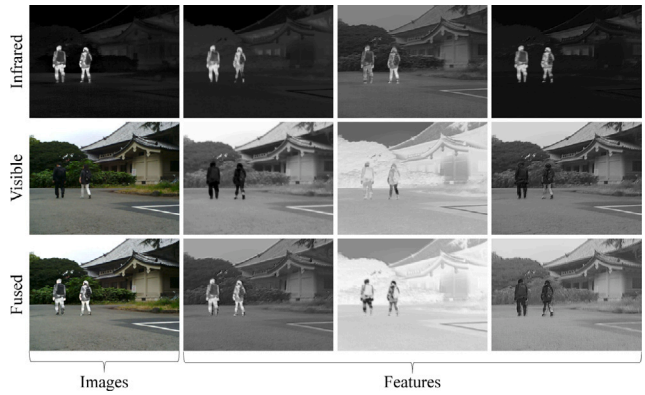

Fig. 18. Visualized results of images and feature maps in the daytime scene. The first column presents the infrared image, visible image and fused image, respectively. The next three columns show the feature maps corresponding to the infrared, visible and fused images in various channel dimensions.

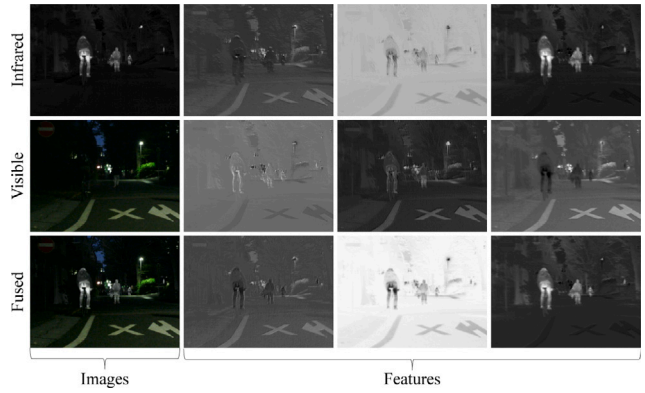

Fig. 19. Visualized results of images and feature maps in the nighttime scenario. The first column shows the infrared image, visible image and fused image, respectively. The following three columns present the feature maps corresponding to the infrared, visible, and fused images in various channel dimensions.

achieve effective feature selection as expected according to illumination changes.

4.8. Application to semantic segmentation

We investigate the positive role of infrared and visible image fusion on semantic segmentation [55]. Specifically, we train the semantic segmentation algorithm [56] on the source images and fused images, respectively. We select 1,000 images as the training set and test the segmentation performance of different models on 360 images. The visualized results and quantitative comparisons are presented in Fig. 20 and Table 4. Visible images contain a great deal of meaningful information in daytime scenes, so segmentation performed on visible images has high accuracy in daytime. However, the visible sensor fails to capture enough information at nighttime as poor illumination. Therefore, the segmentation network cannot effectively segment objects at nighttime, especially people. On the contrary, the infrared images capture thermal information, which can highlight targets, *e.g.*, pedestrians, even under extreme illumination conditions. Thus, the segmentation network can segment person more accurately on infrared images than on visible images. Nevertheless, infrared images contain limited information about vehicles and bicycles, which degrades the segmentation accuracy of vehicles and bicycles.

It is worth mentioning that our progressive fusion network fully integrates the meaningful information of source images. Moreover, complementary features in multi-modal images facilitate enhancing the

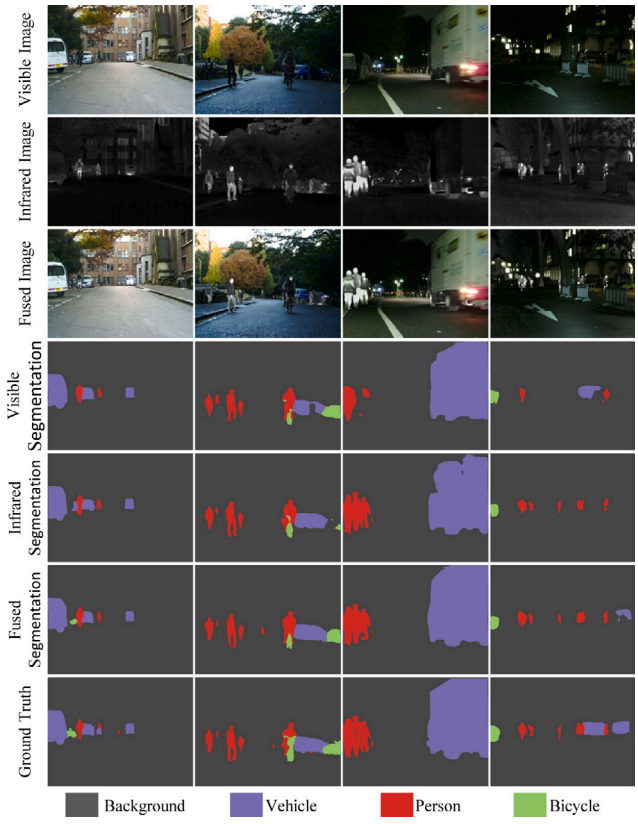

Fig. 20. Segmentation results for infrared, visible and fused images on the MSRS dataset.

Table 4
Segmentation performance (mIoU) of visible, infrared and fused images on the MSRS dataset. RED indicates the best result and *BLUE* represents the second best result.

	Background	Vehicle	Person	Bicycle	Mean
Visible Image	98.88%	*89.36%*	61.03%	**71.34%**	80.15%
Infrared Image	*98.93%*	88.38%	*71.29%*	68.16%	*81.69%*
Fused Image	**99.06%**	**90.11%**	**72.58%**	*71.20%*	**83.24%**

semantic information of fused images. As a result, our segmentation performance of the vehicle and pedestrian is better than single-modal images. Furthermore, the bicycle segmentation accuracy is comparable with visible images. Finally, the mean segmentation accuracy on fused is higher than on infrared and visible images.

4.9. Ablation studies

4.9.1. Illumination-aware loss analysis

We design an illumination-aware loss to guide the training of our progressive fusion network, considering that meaningful information is contained in different source images under diverse illumination conditions. To verify the rationality of illumination-aware loss, we perform an ablation experiment about illumination-aware loss. More specifically, we set $W_{ir} = W_{vi} = 0.5$ and keep other configurations in the ablation experiment. Some typical examples are presented in Fig. 21. It can be observed that the progressive fusion network fails to adaptively fuse meaningful information of infrared and visible images without the illumination-aware loss. In particular, the fused images are suffered from spectral pollution and texture blurring in the daytime scenes. Moreover, the progressive fusion model cannot maintain the contrast of thermal targets or compensate for the background regions with the detail information in infrared images at nighttime. On the contrary, our PIAFusion adaptively achieves contrast maintenance and texture preservation according to illumination conditions.

Fig. 21. Visualized results of ablation on four typical infrared and visible image pairs. From top to bottom: infrared images, visible images, fused results of PIAFusion, PIAFusion without illumination-aware loss, and PIAFusion without CMDAF.

4.9.2. Cross-modality differential aware fusion module analysis

Another essential component in our fusion model is the cross-modality differential aware fusion module, which pre-integrates complementary information of multi-modal images. Therefore, we also implement an ablation experiment about the CMDAF module and present the experimental results in Fig. 21. In the ablation experiment, we remove the CMDAF module from the feature extractor, which means features of infrared and visible images are merged via the halfway fusion strategy. From the results, we could find that fused images preserve the texture details of visible images in the daytime scenes, but they weaken the salient targets. In addition, fused images are similar to infrared images but ignore some detail information in visible images at nighttime. By contrast, our progressive fusion network simultaneously maintains the intensity distribution of salient targets and preserves the texture details of background areas.

Therefore, the proposed PIAFusion adaptively merges common and complementary information according to the illumination conditions, which is benefited from our special designs, *i.e.*, the illumination-aware loss and cross-modality differential aware fusion module.

5. Conclusion

In this work, we have proposed a progressive infrared and visible image fusion framework based on illumination-aware known as PIAFusion, which adaptively integrates meaningful information according to illumination conditions. We have designed an illumination-aware sub-network to estimate the illumination distribution of an input image and calculate the illumination probability. Furthermore, the illumination probability is employed to construct the illumination-aware loss. With the guidance of the illumination-aware loss, the fusion network adaptively merges common and complementary information via the cross-modality differential aware fusion module and the halfway fusion strategy. As a result, the progressive fusion network can around-the-clock generate fused images containing salient targets and abundant texture information. In addition, we have constructed a new infrared and visible dataset, called Multi-Spectral Road Scenarios (MSRS), for the training and benchmark evaluation of image fusion. Extensive experiments have been performed to demonstrate our superiority, including target maintenance, texture preservation and illumination adaptation. Moreover, the expanded experiments on semantic segmentation prove the potential of our proposed method for high-level vision tasks.

CRediT authorship contribution statement

Linfeng Tang: Conceptualization of this study, Methodology, Experiment, Writing. **Jiteng Yuan:** Experiment. **Hao Zhang:** Methodology. **Xingyu Jiang:** Methodology. **Jiayi Ma:** Conceptualization of this study, 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 Natural Science Foundation of Hubei Province (2019CFA037, 2020BAB113).

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, DenseFuse: A fusion approach to infrared and visible images, IEEE Trans. Image Process. 28 (5) (2018) 2614–2623.
[7] 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.
[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] H. Xu, H. Zhang, J. Ma, Classification saliency-based rule for visible and infrared image fusion, IEEE Trans. Comput. Imaging 7 (2021) 824–836.
[12] 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.
[13] 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.
[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] N. Cvejic, D. Bull, N. Canagarajah, Region-based multimodal image fusion using ICA bases, IEEE Sens. J. 7 (5) (2007) 743–751.
[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] 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.
[18] 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.
[19] 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.
[20] 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.
[21] 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.
[22] 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.

[23] 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.
[24] 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.
[25] 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.
[26] 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.
[27] 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.
[28] Z. Zhu, H. Yin, Y. Chai, Y. Li, G. Qi, A novel multi-modality image fusion method based on image decomposition and sparse representation, Inform. Sci. 432 (2018) 516–529.
[29] A. Toet, TNO Image fusion dataset, 2014, URL https://figshare.com/articles/dataset/TNO_Image_Fusion_Dataset/1008029.
[30] 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.
[31] K. Zhou, L. Chen, X. Cao, Improving multispectral pedestrian detection by addressing modality imbalance problems, in: Proceedings of the European Conference on Computer Vision, 2020, pp. 787–803.
[32] 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.
[33] 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.
[34] M. Choi, R.Y. Kim, M.-R. Nam, H.O. Kim, Fusion of multispectral and panchromatic satellite images using the curvelet transform, IEEE Geosci. Remote Sens. Lett. 2 (2) (2005) 136–140.
[35] 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.
[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] 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.
[38] 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.
[39] R. Hou, R. Nie, D. Zhou, J. Cao, D. Liu, Infrared and visible images fusion using visual saliency and optimized spiking cortical model in non-subsampled shearlet transform domain, Multimedia Tools Appl. 78 (20) (2019) 28609–28632.
[40] 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.
[41] 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.
[42] 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.
[43] 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.
[44] 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.
[45] 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.
[46] W. Wang, C. Wei, W. Yang, J. Liu, GLADNet: Low-light enhancement network with global awareness, in: Proceedings of the IEEE International Conference on Automatic Face & Gesture Recognition, 2018, pp. 751–755.
[47] D. Sakkos, E.S. Ho, H.P. Shum, Illumination-aware multi-task GANs for foreground segmentation, IEEE Access 7 (2019) 10976–10986.
[48] C. Li, D. Song, R. Tong, M. Tang, Illumination-aware faster R-CNN for robust multispectral pedestrian detection, Pattern Recognit. 85 (2019) 161–171.
[49] D. Guan, Y. Cao, J. Yang, Y. Cao, M.Y. Yang, Fusion of multispectral data through illumination-aware deep neural networks for pedestrian detection, Inf. Fusion 50 (2019) 148–157.
[50] K. He, J. Sun, X. Tang, Single image haze removal using dark channel prior, IEEE Trans. Pattern Anal. Mach. Intell. 33 (12) (2010) 2341–2353.
[51] H. Xu, X. Wang, J. Ma, DRF: Disentangled representation for visible and infrared image fusion, IEEE Trans. Instrum. Meas. 70 (2021) 5006713.
[52] G. Qu, D. Zhang, P. Yan, Information measure for performance of image fusion, Electron. Lett. 38 (7) (2002) 313–315.
[53] 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.
[54] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, et al., Tensorflow: A system for large-scale machine learning, in: Proceedings of the USENIX Symposium on Operating Systems Design and Implementation, 2016, pp. 265–283.
[55] L. Tang, J. Yuan, J. Ma, Image fusion in the loop of high-level vision tasks: A semantic-aware real-time infrared and visible image fusion network, Inf. Fusion 82 (2022) 28–42.
[56] 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.