IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 11 November 2022, accepted 30 November 2022, date of publication 2 December 2022,
date of current version 7 December 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3226564

RESEARCH ARTICLE

Multiscale Progressive Fusion of Infrared and Visible Images

SEONGHYUN PARK, (Graduate Student Member, IEEE), AND CHUL LEE, (Member, IEEE)
Department of Multimedia Engineering, Dongguk University, Seoul 04620, South Korea

Corresponding author: Chul Lee (chullee@dongguk.edu)

This work was supported by the National Research Foundation of Korea (NRF) funded by the Korean Government (MSIT) under Grant NRF-2022R1F1A1074402.

ABSTRACT Infrared and visible image fusion aims to generate more informative images of a given scene by combining multimodal images with complementary information. Although recent learning-based approaches have shown significant fusion performance, developing an effective fusion algorithm that can preserve complementary information while preventing bias toward either of the source images remains a significant challenge. In this work, we propose a multiscale progressive fusion (MPFusion) algorithm that extracts and progressively fuses multiscale features of infrared and visible images. The proposed algorithm consists of two networks, IRNet and FusionNet, which extract the intrinsic features of infrared and visible images, respectively. We transfer the multiscale information of the infrared image from IRNet to FusionNet to generate an informative fusion result. To this end, we develop the multi-dilated residual block (MDRB) and the progressive fusion block (PFB), which progressively combines the multiscale features from IRNet with those from FusionNet to fuse complementary features effectively and adaptively. Furthermore, we exploit edge-guided attention maps to preserve complementary edge information in the source images during fusion. Experimental results on several datasets demonstrate that the proposed algorithm outperforms state-of-the-art infrared and visible image fusion algorithms on both quantitative and qualitative comparisons.

INDEX TERMS Image fusion, infrared image, visible image, multiscale network, edge-guided attention map.

I. INTRODUCTION

Image fusion is a technique that combines multiple images captured from different sensors to generate a more informative image of a given scene that can facilitate subsequent processing [1], [2], [3], [4], [5]. A pair of infrared and visible images is the most commonly used combination of modalities because the images captured in the two wavelengths contain complementary information on a scene from different aspects, and thereby provide more robust and informative results together [2]. In particular, whereas visible images contain scene textures to facilitate human visual perception, their quality is easily affected by environmental conditions, such as illumination or weather. In contrast, because infrared images capture the thermal radiation of objects, they are robust against environmental conditions but have poor scene textures [6]. Infrared and visible image fusion techniques have been applied in various applications because of its practical usefulness and importance, including object tracking [7], [8], salient object detection [9], [10], [11],

(a) Infrared images (b) Visible images (c) Fused images

FIGURE 1. An example of object detection performance improvement using infrared and visible image fusion.

The associate editor coordinating the review of this manuscript and approving it for publication was Felix Albu.

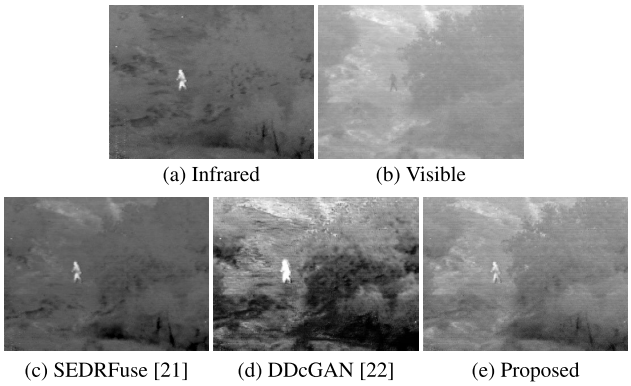


FIGURE 2. Comparison of infrared and visible image fusion results obtained by different algorithms. The proposed algorithm can better preserve complementary information in the source images.

and surveillance [12]. Figure 1 shows an example in which visible and infrared image fusion improves object detection performance.

The key challenge in image fusion is the development of effective feature extraction from each image and appropriate fusion rules to integrate them into the fused image. Various algorithms have recently been proposed to address this challenge. These algorithms can be broadly classified as model- and learning-based [2]. Model-based algorithms have been designed to extract image features based on different mathematical theories and then determine appropriate fusion rules on the basis of the extracted features [13], [14], [15], [16], [17], [18], [19], [20]. However, the extraction of faithful features using such manually designed models makes designing fusion rules difficult and computationally demanding.

With recent advances in deep learning, deep learning-based algorithms that employ convolutional neural networks (CNNs) [21], [23], [24], [25], [26], [27], [28], [29], [30], [31], [32] or generative adversarial networks (GANs) [22], [33], [34], [35], [36], [37] have been developed most actively. CNNs can extract high-level features from source images more effectively than traditional feature engineering, which is essential to generate informative fused images. Therefore, CNN-based fusion algorithms have been designed to learn to extract informative features and fuse them by characterizing the complex relations between source images and fused images. However, despite the powerful ability of CNNs to extract visual features, CNN-based algorithms may fail to preserve complementary information in either of the source images, thereby generating biased fusion results [38]. Further, single-scale feature extraction [23], [24], [25] hardly utilizes both global and local information simultaneously, which leads to a loss of spatial information in the source images. GAN-based algorithms generate fused images that preserve the pixel value distributions of both infrared and visible images. Although GAN-based algorithms have achieved improved performance, they have also exhibited limited ability to highlight discriminative regions in source images and generated undesirable artifacts and noise [22], [33], [34], [35]. Figure 2 shows an example of a pair of infrared and visible images and the fusion results obtained by SEDRFuse [21] and DDcGAN [22], which are representative CNN- and GAN-based algorithms, respectively. The result of SEDRFuse in Figure 2(c) is biased toward the infrared image, whereas that of DDcGAN in Figure 2(d) is biased toward the visible image. In contrast, the proposed algorithm achieves a desirable balance between the two source images by better preserving the dominant infrared objects and rich visible details.

Recently, several transformer-based fusion algorithms [39], [40], [41], [42], [43] have been developed to capture inter-domain long-range dependencies with self-attention mechanism. For example, in [39] and [40], local and global features respectively extracted by CNNs and transformers were integrated to take advantages of both models. In [41] and [42], both self-attention and cross-attention were utilized in pure transformers without CNNs. However, transformer-based fusion algorithms generally demand considerable computational resources to capture long-range dependencies, limiting their applicability to high-resolution images.

In this work, to address the aforementioned limitations of conventional algorithms and better preserve the complementary information in source images, we propose a multiscale progressive fusion algorithm, called MPFusion, for infrared and visible image pairs. The proposed algorithm is composed of two networks: IRNet, which extracts multiscale features from infrared images, and FusionNet, which extracts features from visible images and then progressively fuses the features extracted from both images. To this end, we develop the multi-dilated residual block (MDRB) and the progressive fusion block (PFB) to fuse the multiscale features extracted from IRNet with those from FusionNet. In addition, we develop edge-guided attention maps to faithfully preserve the complementary edge information in the source images during fusion. Experimental results show that the proposed MPFusion algorithm substantially outperforms state-of-the-art infrared and visible image fusion algorithms [21], [22], [29], [30], [31], [32], [33], [38], [44], [45], [46] on several datasets.

The main contributions of this work are summarized as follows:

- We propose the MPFusion algorithm for infrared and visible image fusion to extract and progressively fuse multiscale features of source images; thus, the MPFusion algorithm can preserve both global and local information in source images.
- We develop two new blocks, called MDRB and PFB, which are designed to improve fusion performance by effectively exploiting multiscale features. Specifically, MDRB progressively extracts intrinsic features of source images, whereas PFB progressively fuses their complementary information.
- We develop an adaptive channel fusion strategy that adaptively combines infrared and visible features to better exploit information on the statistical characteristics of the source images.

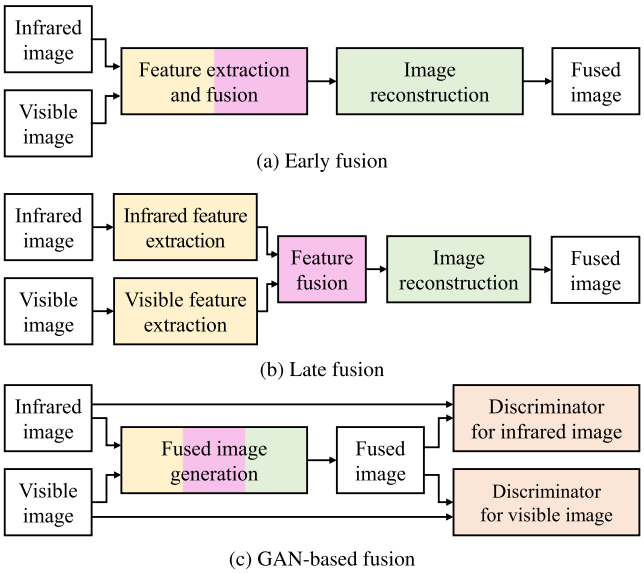


FIGURE 3. Illustration of different architectures for infrared and visible image fusion.

- We experimentally show that the proposed MPFusion algorithm outperforms state-of-the-art fusion algorithms on multiple datasets.

The remainder of this paper is organized as follows. Section II briefly reviews related work. Section III describes the proposed MPFusion algorithm for infrared and visible image fusion, and Section IV discusses the experimental results. Finally, Section V concludes the paper.

II. RELATED WORK

A. MODEL-BASED FUSION

Model-based algorithms have been developed based on different mathematical or algorithmic models for feature extraction and fusion rules [2]. For example, multiscale transform-based algorithms [13], [14] decompose each source image into multiscale representations, fuse them in a transform domain, and then obtain a fused image using the inverse multiscale transform. Sparse representation-based algorithms [15], [16], [17], [18] learn to construct over-complete dictionaries to represent the fused image. Saliency-based algorithms [47], [48] estimate salient areas of source images to improve the visual quality of the fused images by preserving important features in the source images. Finally, hybrid algorithms [49], [50] combine other model-based algorithms to improve fusion performance. However, model-based feature extraction complicates image fusion tasks, and considerable attention is required to ensure the completeness of features [30]. For a more detailed survey on model-based fusion, the reader is referred to [2].

B. LEARNING-BASED FUSION

Inspired by recent successes in deep learning-based computer vision and image processing tasks, extensive research has been conducted on learning-based infrared and visible image fusion. Learning-based fusion algorithms can be broadly categorized into three groups based on how they extract and fuse the features of each image. Figure 3 compares the three architectures commonly used for learning-based image fusion. The two architectures in Figures 3(a) and (b) use CNNs or transformers comprising feature extraction, feature fusion, and image reconstruction, whereas that in Figure 3(c) uses GANs.

Figure 3(a) shows an early fusion architecture, which performs feature extraction and feature fusion simultaneously, followed by image reconstruction in an end-to-end manner. Owing to its effectiveness in removing the correlation between two source images, many algorithms [25], [26], [29], [30], [38], [42], [44], [51], [52] using this architecture have recently been proposed. In particular, several researches [38], [44] have focused on designing network architectures for effective extraction of useful features from source images and their fusion. However, since algorithms in early fusion extract and fuse features simultaneously using a common block without considering different modalities, the intrinsic features and complementary information of source images may not be fully exploited. Thus, algorithms in this category may generate biased fused images.

The late fusion architecture in Figure 3(b) extracts the features of each source image separately using CNNs dedicated to each of the two modalities, and then fuses them using a fusion scheme. As late fusion algorithms fuse the features extracted using independently trained networks, they can preserve the intrinsic features of each image. Most researches have focused on the design of elaborate network architectures for end-to-end fusion capable of both better feature extraction and feature fusion [21], [23], [24], [32], [39], [40], [41], [43]. In addition, in [31], an algorithm was developed to generate weight maps for effective fusion using pretrained networks. The proposed MPFusion algorithm similarly performs feature extraction and fusion separately. However, it extracts and progressively fuses multiscale features [53] to preserve complementary information in the source images more effectively, thereby being capable of avoiding bias toward either of the source images.

Finally, several GAN-based algorithms [22], [33], [34], [35], [36], [37] have recently been proposed, which generate a fused image by preserving the pixel value distributions in the source images through an adversarial game between a generator and a discriminator, as shown in Figure 3(c). In particular, as using a single discriminator may fail to preserve pixel value distributions of both images [33], a GAN architecture with two discriminators was developed to overcome the limitations of a single discriminator [22]. In addition, based on the observation that GAN-based algorithms have limited ability to highlight discriminative regions in source images, attempts have been made to incorporate an attention mechanism into GANs [34], [35]. However, because GAN-based algorithms also jointly perform feature extraction and fusion implicitly, they may generate fused images that are biased toward either of the source images as well.

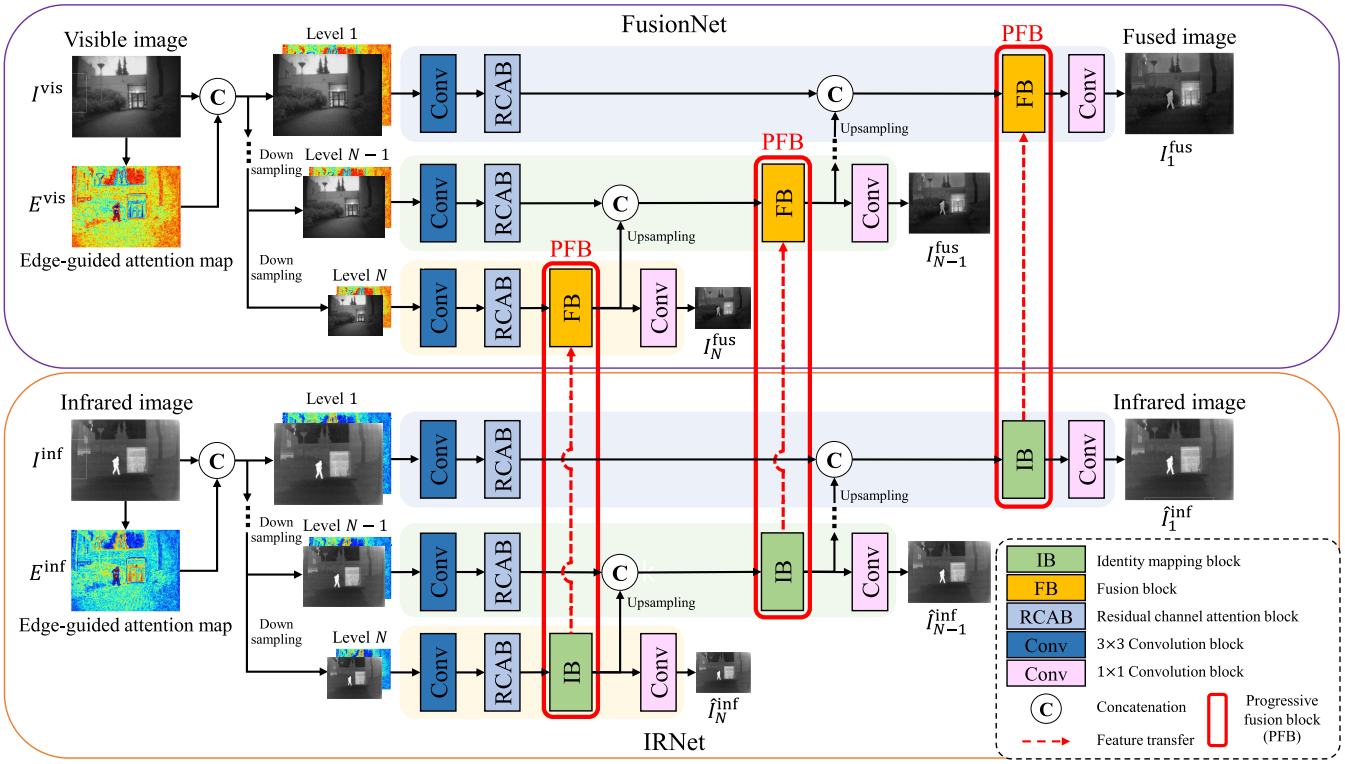


FIGURE 4. Overview of the proposed MPFusion algorithm. Given an infrared image I^{inf} and a visible image I^{vis} and their corresponding edge-guided attention maps E^{inf} and E^{vis}, respectively, IRNet extracts the multiscale intrinsic features of I^{inf} and FusionNet yields a fused image I_1^{fus}. IRNet is trained first, and subsequently FusionNet is trained with PFB, which feeds IRNet information progressively into FusionNet.

III. PROPOSED ALGORITHM

Figure 4 shows an overview of the proposed MPFusion algorithm, which consists of two networks. IRNet extracts multiscale features of an infrared image I^{inf} and FusionNet outputs the fused result of the infrared and visible images, denoted by I^{inf} and I^{vis}, respectively. The edge-guided attention maps E^{inf} and E^{vis} for I^{inf} and I^{vis} respectively, are used to improve the fusion performance by preserving the edge information in the source images. Features extracted by IRNet are fed into FusionNet progressively through the PFB at each level. Note that IRNet is trained separately to extract the intrinsic features of the infrared image, and FusionNet is trained with fixed IRNet. This training strategy improves the fusion performance and ensures the stable training of FusionNet, as will be discussed in Section IV-E.

A. EDGE-GUIDED ATTENTION MAPS

As mentioned previously, infrared and visible image fusion algorithms fuse source images in the feature domain rather than in the image domain. Thus, fine texture details in the source images may be lost during fusion due to unfaithful feature extraction, resulting in a blurry output. Note that the attention mechanism selectively focuses on important parts of the input data to improve the performance of CNNs, and the image details are well represented by the gradient of the image [54], [55]. Based on this observation, we define the edge-guided attention map as the relative magnitude of the gradient of the infrared and visible images. More specifically, we obtain the edge-guided attention maps E^{inf} and E^{vis} for infrared and visible images, respectively, as

$$E^{inf} = \frac{|\nabla I^{inf}|}{|\nabla I^{inf}|+|\nabla I^{vis}|}, \tag{1}$$

$$E^{vis} = \frac{|\nabla I^{vis}|}{|\nabla I^{inf}|+|\nabla I^{vis}|}, \tag{2}$$

where ∇ denotes the gradient operator and the division is element-wise. As shown in Figure 4, the edge-guided attention maps are concatenated with the corresponding source images; then, they are downsampled to construct multiscale inputs. The edge-guided attention maps force the network to focus more on the complementary edge information in the source images, thereby improving the fusion performance, as will be discussed in Section IV-E.

B. NETWORK ARCHITECTURE

As shown in Figure 4, both IRNet and FusionNet are multiscale networks that extract image features at multiple levels and successively add the features of the previous level to generate output images. Note that multiscale features of source images have been shown to be effective in infrared and visible image fusion [56], [57]. Each level of the networks is responsible for a particular aspect of the source images: a higher-level network for local details while a lower-level network for global structures. At each level of the networks, the residual channel attention block (RCAB) [58] is used first

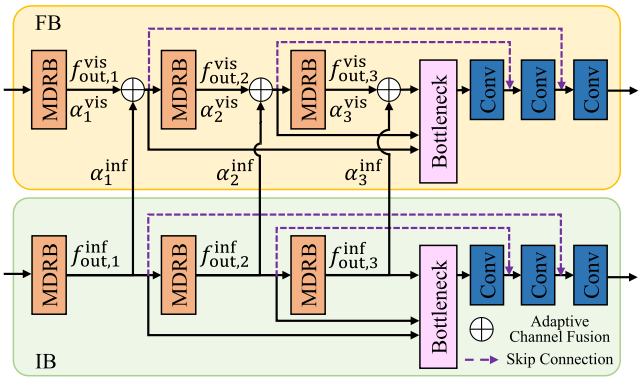


FIGURE 5. Architecture of the proposed PFB. The infrared features f^{inf}_{out} in IB are progressively fed into FB to be added with the visible features f^{vis}_{out}. The MDRB is detailed in Figure 7.

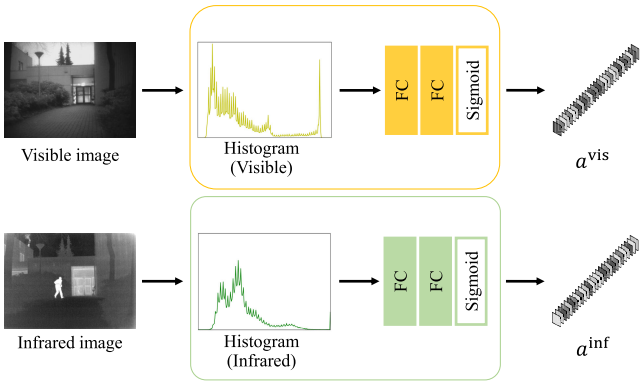


FIGURE 6. Architecture of the adaptive channel-weight generation network.

to force the network to focus on more informative features by adaptively rescaling channel-wise features. Then, the features extracted by IRNet are fed into FusionNet through PFB, which is composed of the identity mapping block (IB) in IRNet and the fusion block (FB) in FusionNet, to progressively fuse the features of both networks. The architecture of the proposed PFB will be described in detail in subsequent sections. Finally, the output images at each level are generated by applying a 1×1 convolutional layer. In this work, the level of the networks is fixed to $N = 3$; its effects will be discussed in Section IV-E.

We feed the IRNet features unidirectionally into FusionNet for the progressive fusion of infrared and visible images. To this end, we develop the PFB to combine the information from both source images. Figure 5 shows the architecture of the proposed PFB, which consists of IB in IRNet and FB in FusionNet. IB extracts features to generate the input infrared image, whereas FB extracts those of the input visible image and fuses them with the IB features. Both IB and FB have three MDRBs to preserve both global and local information in the source images, a bottleneck layer for dimensionality reduction, and three convolutional layers.

C. ADAPTIVE CHANNEL FUSION

In Figure 5, the infrared features f^{inf}_{out} in the IB and the visible features f^{vis}_{out} in the FB are fused and then fed into the next layer of the FB. An addition or concatenation can be used to fuse these features, as in [59]. However, such straightforward approaches may fail to fully exploit the different characteristics of the source images, degrading the fusion performance, as will be discussed in Section IV-E. Thus, we develop an adaptive channel fusion strategy that adaptively combines two features by exploiting information on the statistical characteristics of the source images during fusion. Specifically, we construct two weight maps α^{inf} and $\alpha^{vis} \in \mathbb{R}^{(N_{MDRB} \times N_C) \times 1 \times 1}$ for the infrared and visible images, respectively, where N_{MDRB} and N_C are respectively the numbers of MDRB and its output channels. It has been observed that pixel value distributions of source images are essential for representing texture details in images for fusion [25]. Therefore, we use the histograms of the source images to construct weight maps to consider their pixel value distributions more effectively. More specifically, we employ a simple network with two fully connected (FC) layers followed by a sigmoid activation function to learn a weight map for each image, which takes the normalized histogram of each image as input. Figure 6 illustrates the architecture of the adaptive channel-weight generation network.

Next, after each MDRB, the PFB fuses the features of the IB with those of the FB and then feeds the fused features into the next layer of the FB. More specifically, let f^{inf}_{out} and f^{vis}_{out} denote the output features of MDRB in IB and FB, respectively; then, the input feature of the next layer f^{fus}_{out} in FB by adaptive channel fusion is obtained by

$$f^{fus}_{out} = \frac{\alpha^{inf} \odot f^{inf}_{out} + \alpha^{vis} \odot f^{vis}_{out}}{\alpha^{inf} + \alpha^{vis}}, \tag{3}$$

where $\odot$ denotes channel-wise multiplication and the division is also channel-wise. This strategy enables FusionNet to fuse the infrared and visible features progressively and stably, while preserving the intrinsic features of each image.

D. MDRB

It is important to extract features by fully exploiting the characteristics of the input image and to feed them through the network without loss for high-quality image generation. The multiscale residual block (MSRB) [60] using convolution kernels of different sizes has been frequently employed for feature extraction. However, MSRB requires high computational and memory complexities to increase the receptive field. In this work, inspired by MSRB, we develop MDRB to extract deep features at different scales by employing dilated convolution [61], which can expand the receptive field using the same number of parameters. Figure 7 shows the architecture of the proposed MDRB. MDRB adds the input features f_{t-1} to the output features of two shared bypass networks that use kernels with different dilation rates r, generating the output features f_t. MDRB provides better fusion results than MSRB by faithfully preserving both global and local

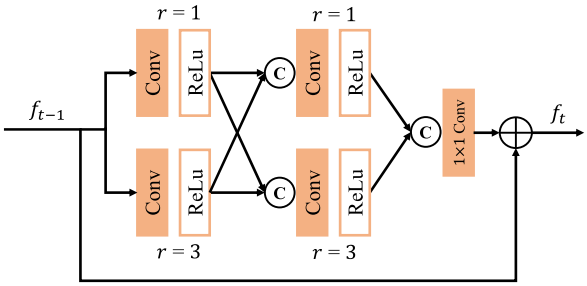


FIGURE 7. Architecture of the proposed MDRB. The features are shared by two bypass networks with different dilated convolutions with dilation rates r.

information with fewer parameters using dilated convolutions, as will be discussed in Section IV-E.

E. LOSS FUNCTIONS

To train IRNet and FusionNet, we define the IR loss $\mathcal{L}_{\text{IR}}$ and fusion loss $\mathcal{L}_{\text{fus}}$, respectively, as will be described subsequently.

1) IR LOSS

To train IRNet, we define the IR loss $\mathcal{L}_{\text{IR}}$ as the weighted sum of the data loss $\mathcal{L}_{\text{id}}$ and structure loss $\mathcal{L}_{\text{s}}$ between an input infrared image and its estimated version as

$$\mathcal{L}_{\text{IR}} = \mathcal{L}_{\text{id}} + \lambda_{\text{s}}\mathcal{L}_{\text{s}}, \tag{4}$$

where λ_{s} is a hyper-parameter that balances these two losses. We employ the ℓ_2 norm as the infrared data loss as

$$\mathcal{L}_{\text{id}} = \frac{1}{N}\sum_{k=1}^{N} \left\| \hat{I}_k^{\text{inf}} - I_k^{\text{inf}} \right\|_2, \tag{5}$$

where $\hat{I}_k^{\text{inf}}$ and I_k^{inf} represent the estimated image and the corresponding input image, respectively, at the kth network level. The structure loss is defined as

$$\mathcal{L}_{\text{s}} = 1 - \text{SSIM}\big(\hat{I}^{\text{inf}}, I^{\text{inf}}\big), \tag{6}$$

where SSIM$(\cdot)$ denotes the structural similarity index [62] between the two images.

2) FUSION LOSS

We define the fusion loss $\mathcal{L}_{\text{fus}}$ to train FusionNet as a weighted sum of the data loss $\mathcal{L}_{\text{fd}}$, spatial loss $\mathcal{L}_{\text{sp}}$, and perceptual loss $\mathcal{L}_{\text{p}}$ as

$$\mathcal{L}_{\text{fus}} = \mathcal{L}_{\text{fd}} + \lambda_{\text{sp}}\mathcal{L}_{\text{sp}} + \lambda_{\text{p}}\mathcal{L}_{\text{p}}, \tag{7}$$

where λ_{sp} and λ_{p} are hyper-parameters that control the relative impacts of the three losses. The fusion data loss is defined as

$$\mathcal{L}_{\text{fd}} = \frac{1}{N}\sum_{k=1}^{N} w_{\text{inf}} \cdot \left\| I_k^{\text{fus}} - I_k^{\text{inf}} \right\|_2 + w_{\text{vis}} \cdot \left\| I_k^{\text{fus}} - I_k^{\text{vis}} \right\|_2, \tag{8}$$

where w_{inf} and w_{vis} denote the weights that control the contributions of the input infrared and visible images, respectively, to the fused image. We employ the spatial consistency loss [63] to preserve the spatial characteristics of the source images, which is given by

$$\mathcal{L}_{\text{sp}} = \frac{1}{K}\sum_{i=1}^{K}\sum_{j\in\Omega(i)} \Bigg(w_{\text{inf}} \cdot \Big(\Big| I_i^{\text{fus}} - I_j^{\text{fus}} \Big| - \Big| I_i^{\text{inf}} - I_j^{\text{inf}} \Big| \Big)^2 + w_{\text{vis}} \cdot \Big(\Big| I_i^{\text{fus}} - I_j^{\text{fus}} \Big| - \Big| I_i^{\text{vis}} - I_j^{\text{vis}} \Big| \Big)^2 \Bigg), \tag{9}$$

where K denotes the number of regions and $\Omega(i)$ denotes the neighboring regions of i. Finally, to compare the high-level differences between the source images and fused image, we employ the perceptual loss [64] as

$$\mathcal{L}_{\text{p}} = \sum_{k=2,4,6} \Big(w_{\text{inf}} \cdot \Big\| \phi^k(I^{\text{fus}}) - \phi^k(I^{\text{inf}}) \Big\|_1 + w_{\text{vis}} \cdot \Big\| \phi^k(I^{\text{fus}}) - \phi^k(I^{\text{vis}}) \Big\|_1 \Big), \tag{10}$$

where ϕ^k denotes the feature map from the kth layer of the pretrained VGG-16 network [65].

IV. EXPERIMENTAL RESULTS

A. TRAINING

We first train IRNet, which is then fixed to train FusionNet.

IRNet: We use the Adam optimizer [66] with a learning rate of 10^{-4} and a batch size of 8 for 16 epochs. The hyper-parameter λ_{s} in (4) is fixed to 100.

FusionNet: We also use the Adam optimizer with the same settings as in IRNet with a batch size of 4 for 25 epochs. The hyper-parameters λ_{sp} and λ_{p} in (7) are fixed to 0.05 and 0.5, respectively, and w_{inf} and w_{vis} in (8)–(10) are all set to 0.5.

Training dataset: We use only the KAIST dataset [67] for training, which contains 95,000 well-aligned color-thermal image pairs with a resolution of 640 × 512. We augment the dataset by converting the RGB color to grayscale and randomly cropping 20,000 256 × 256 patches.

B. EXPERIMENTAL SETTINGS

1) DATASETS

Although we strictly use a single training dataset, we evaluate the performance of the proposed algorithm on various datasets to test its effectiveness and generalization ability.

KAIST [67]**:** The KAIST dataset contains well-aligned 95,000 image pairs captured using special camera devices with a resolution of 640 × 512. We randomly chose 200 pairs, which were not used for training.

TNO [68]**:** The TNO dataset contains multispectral nighttime scene images of various resolutions, ranging from 280 × 280 to 768 × 576, registered with multiband camera systems. We use the test set constructed by Li and Wu [38], which contains 20 image pairs.

RoadScene [30]**:** The RoadScene dataset contains aligned visible and infrared image pairs chosen by Xu et al. [30] from the FLIR dataset, which contains image pairs captured using

TABLE 1. Quantitative comparison of the fusion results on the KAIST, TNO, and RoadScene datasets using eight quality metrics. For each metric, the best result is shown in boldface, whereas the second-best is underlined. For each algorithm, the average ranking is reported.

	En (↑)	$Q^{AB/F}$ (↑)	SCD (↑)	MS-SSIM (↑)	FMI_{dct} (↑)	FMI_w (↑)	NIQE (↓)	BRISQUE (↓)	Avg. rank
KAIST dataset									
GTF [45]	6.3580	**0.7518**	0.4234	0.8203	0.3413	0.3641	3.5283	**38.2529**	6.75
VggML [31]	6.4160	0.4249	1.4121	0.9012	0.3604	0.3686	3.7877	43.3369	7.50
DenseFuse [38]	6.4038	0.3834	1.4087	0.8957	0.3620	0.3709	3.6522	43.9859	7.63
FusionGAN [33]	6.6596	0.1561	1.1354	0.6332	0.1139	0.1564	4.6517	46.9863	11.38
IFCNN [29]	6.8241	0.6364	1.4728	0.9548	0.3454	0.3719	3.4436	38.6872	4.25
SEDRFuse [21]	7.1142	0.5629	1.6481	0.9367	0.3192	0.3643	3.1701	39.0740	4.50
DDcGAN [22]	**7.3284**	0.5333	1.6090	0.9212	0.3233	0.3743	3.4185	41.6222	4.50
U2Fusion [30]	7.0414	0.5928	1.5590	**0.9584**	0.3299	0.3484	3.4185	41.6222	6.00
DRF [32]	6.9918	0.5065	1.1592	0.9108	0.3115	0.3495	3.6315	44.1095	8.00
IVFusion [46]	7.0371	0.3581	1.1164	0.8075	0.3023	**0.4152**	3.6891	42.0342	7.88
RFN-Nest [44]	7.0654	0.6259	1.6361	0.9542	0.2661	0.2972	4.3555	45.4504	7.25
MPFusion (Proposed)	6.9261	0.6410	**1.6518**	0.9568	**0.3699**	0.3971	**3.1684**	38.9736	**2.38**
TNO dataset									
GTF [45]	6.6371	0.4181	1.0154	0.8142	**0.4201**	**0.4339**	4.5705	27.8935	5.75
VggML [31]	6.1639	0.3661	1.6352	0.8749	0.4023	0.4143	3.8179	33.1756	6.63
DenseFuse [38]	6.6308	0.3719	1.7605	0.9113	0.4033	0.4144	3.8186	33.7323	5.75
FusionGAN [33]	6.3922	0.1586	1.4013	0.7418	0.1136	0.1533	4.5843	40.6689	11.38
IFCNN [29]	6.5736	0.4993	1.7114	0.9046	0.3713	0.3995	3.9199	**26.7474**	4.88
SEDRFuse [21]	6.7008	0.4211	1.7941	0.9016	0.3413	0.3831	4.9258	33.0565	6.63
DDcGAN [22]	**7.2892**	0.3330	1.6760	0.5996	0.3263	0.3784	3.8332	28.6108	6.88
U2Fusion [30]	6.7655	0.3923	1.7663	**0.9217**	0.3912	0.4105	3.9285	30.6118	4.88
DRF [32]	6.8607	0.3836	1.5064	0.7674	0.3371	0.3840	**3.5330**	29.6288	6.38
IVFusion [46]	7.1757	0.2964	1.1457	0.7930	0.2966	0.3341	4.4002	32.1631	8.75
RFN-Nest [44]	6.8648	0.4093	**1.8140**	0.8984	0.3286	0.3290	4.3229	38.1917	6.75
MPFusion (Proposed)	6.8350	**0.5090**	1.8011	0.9139	0.3913	0.4285	4.0837	28.8070	**3.38**
RoadScene dataset									
GTF [45]	7.4961	0.3527	1.0201	0.7416	0.3826	0.3807	3.1525	31.6851	7.38
VggML [31]	6.8342	0.4174	1.3685	0.8648	0.3875	0.4260	2.7139	29.4141	4.63
DenseFuse [38]	6.8250	0.3918	1.3633	0.8585	**0.3891**	**0.4272**	**2.6651**	28.6425	5.00
FusionGAN [33]	6.9395	0.1706	0.9707	0.5893	0.1193	0.1762	4.4951	41.4441	11.25
IFCNN [29]	7.1218	**0.5446**	1.4391	0.8957	0.3467	0.4090	2.9810	**25.5031**	4.00
SEDRFuse [21]	6.3916	0.4037	1.5853	0.8289	0.2885	0.3477	4.6119	43.2400	8.88
DDcGAN [22]	7.4112	0.3273	1.5660	0.7533	0.2949	0.3312	2.9088	29.5907	7.38
U2Fusion [30]	7.3437	0.4046	1.6051	0.8310	0.3467	0.3854	2.7196	27.3004	4.25
DRF [32]	7.2929	0.4158	1.4021	0.7664	0.3192	0.3810	2.9226	26.9762	6.13
IVFusion [46]	**7.5172**	0.3254	1.1326	0.7689	0.3264	0.2920	3.4968	30.8065	8.00
RFN-Nest [44]	7.3006	0.3348	**1.6849**	0.8663	0.3065	0.2900	4.2401	44.6952	7.63
MPFusion (Proposed)	7.2792	0.4982	1.6670	**0.8986**	0.3377	0.3980	2.7968	26.9476	**3.38**

real cameras.[1] We use its test set, which contains 221 image pairs with resolutions of up to 563 × 459.

2) ALGORITHMS USED FOR COMPARISONS

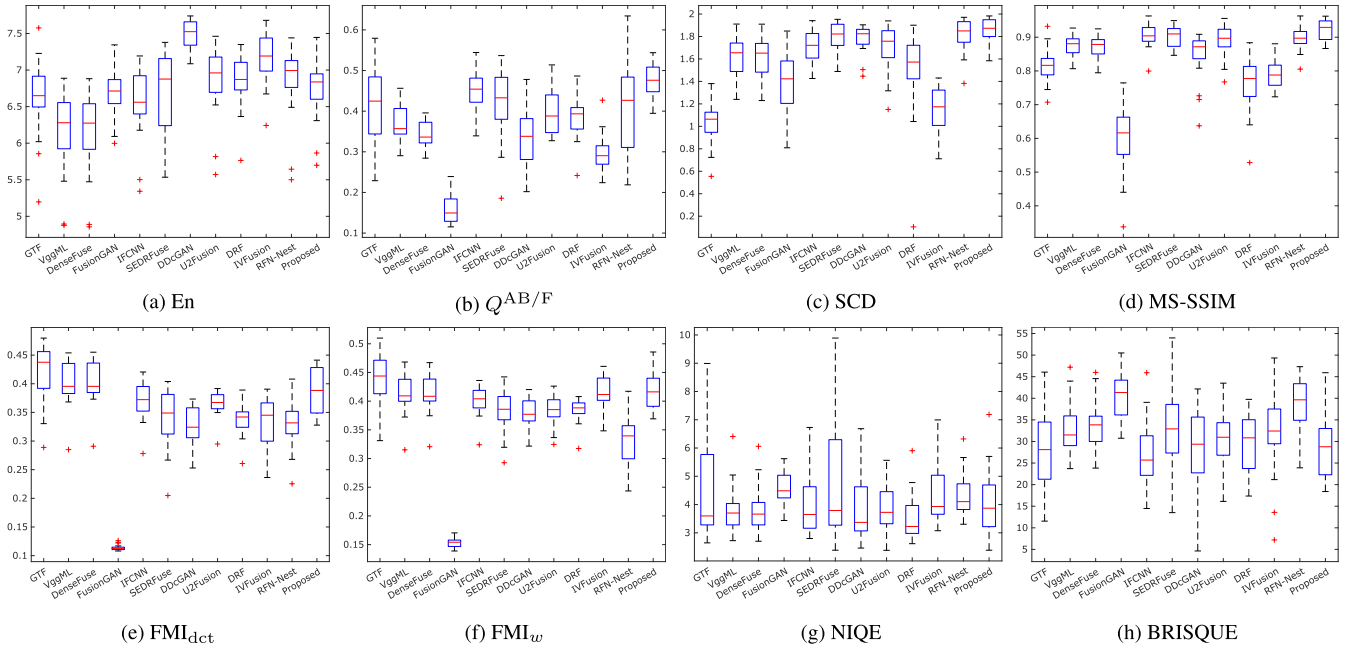
We compare the fusion performance of the proposed algorithm with those of conventional algorithms: GTF [45], VggML [31], DenseFuse [38], FusionGAN [33], IFCNN [29], SEDRFuse [21], DDcGAN [22], U2Fusion [30], DRF [32], IVFusion [46], and RFN-Nest [44]. We retrained the learning-based algorithms [21], [22], [30], [32], [33], [38], [44] with the parameter settings recommended by the respective authors using the KAIST dataset, except for VggML [31] and IFCNN [29], which use pretrained networks, and model-based GTF [45] and IVFusion [46]. The results of the conventional algorithms were obtained by executing the codes provided by their respective authors. The source codes and pretrained models are available on our project website.[2]

C. QUANTITATIVE ASSESSMENT

We use eight frequently used objective quality metrics to evaluate the fusion performance: entropy (En) [69], total edge information ($Q^{AB/F}$) [70], sum of the correlations of differences (SCD) [71], multiscale structural similarity (MS-SSIM) [72], mutual information for discrete cosine features (FMI_{dct}) [73] as well as wavelet features (FMI_w) [73], natural image quality evaluator (NIQE) [74], and blind/referenceless image spatial quality evaluator (BRISQUE) [75]. The scores for En, $Q^{AB/F}$, SCD, MS-SSIM, FMI_{dct}, and FMI_w are computed between the fused image and the input visible and infrared images, and then averaged. As the ground-truths are unavailable for infrared and visible image fusion, we also use two blind objective quality metrics:

[1] https://www.flir.ca/oem/adas/adas-dataset-form/

[2] https://github.com/seonghyun0108/MPFusion



(a) En (b) $Q^{\mathrm{AB/F}}$ (c) SCD (d) MS-SSIM

(e) $\mathrm{FMI}_{\mathrm{dct}}$ (f) FMI_{w} (g) NIQE (h) BRISQUE

FIGURE 8. Comparison of box plots for the eight quality metrics in Table 1 on the TNO dataset.

NIQE and BRISQUE. Higher En, $Q^{\mathrm{AB/F}}$, SCD, MS-SSIM, $\mathrm{FMI}_{\mathrm{dct}}$, and FMI_{w} scores imply better results, whereas lower NIQE and BRISQUE scores indicate better performance.

Table 1 compares the quantitative performances. The proposed MPFusion algorithm provides the highest or second-highest $Q^{\mathrm{AB/F}}$, SCD, and MS-SSIM scores for each dataset, implying that the proposed algorithm can better preserve the structures and details of the input images via multiscale and progressive feature fusion. DDcGAN yields relatively high scores in terms of the information theory-based metrics, *i.e.*, En, $\mathrm{FMI}_{\mathrm{dct}}$, and FMI_{w}, but achieves lower scores on the fidelity-based metrics $Q^{\mathrm{AB/F}}$ and MS-SSIM. This is because the GAN-based DDcGAN tends to generate noise and artifacts in fused images, which increase the amount of information conveyed but degrade the visual quality of the resulting images. Note that, because the information theory-based metrics quantify the amount of information in the images, DDcGAN yields high scores of the information theory-based metrics. The results of DDcGAN indicate that each quality metric assesses different aspects of image quality. Thus, one algorithm may outperform the others in terms of a single metric; however, it may perform poorly in terms of other metrics. Therefore, we evaluate the overall performance of the algorithms by employing a ranking-based assessment. Specifically, we obtain the ranking of each algorithm in each metric, and the average rankings are presented in the rightmost column of Table 1. The proposed algorithm consistently yields the best average rankings for all datasets with large margins, which confirms its effectiveness. In addition, the proposed algorithm exhibits similar tendencies across the quality metrics for all the datasets. This confirms the superior generalization ability of the proposed algorithm compared with the algorithms used for comparison.

Finally, Figure 8 shows the box plots for the eight quality metrics in Table 1 using all test images in the TNO dataset. The red lines and crosses denote median values and outliers, respectively. The proposed MPFusion algorithm achieves the highest median values for $Q^{\mathrm{AB/F}}$, SCD, and MS-SSIM scores in Figures 8(b), (c), and (d), respectively. In addition, the proposed algorithm yields the smallest number of outliers. This indicates that the proposed algorithm is more stable and robust than the conventional algorithms.

D. QUALITATIVE ASSESSMENT

Figure 9 compares the fusion results obtained by each algorithm on the KAIST dataset. GTF in Figure 9(c) loses the fine textures in the input images. In Figures 9(d), (e), (g), (h), and (k), VggML, DenseFuse, IFCNN, SEDRFuse, and DRF, respectively, yield relatively blurry results losing texture details, *e.g.*, the license plate in the second row. The GAN-based algorithms FusionGAN and DDcGAN in Figures 9(f) and (i), respectively, generate undesirable artifacts and noise that alter the image characteristics, *e.g.*, around the car in the second row. IVFusion in Figure 9(l) over-enhances the contrast of the input images. U2Fusion and RFN-Nest in Figures 9(j) and (m), respectively, preserve fine details in the visible images, but lose those in infrared images, *e.g.*, the trees in the third row. In contrast, the proposed algorithm in Figure 9(n) generates a fused image that preserves the fine textures of both input images faithfully without noticeable artifacts.

Figure 10 shows the fused images from the TNO dataset. GTF, IFCNN, SEDRFuse, and DRF in Figures 10(c), (g), (h), and (k), respectively, fail to effectively retain complementary information in both input images; the fused images contain more information from the infrared images while losing

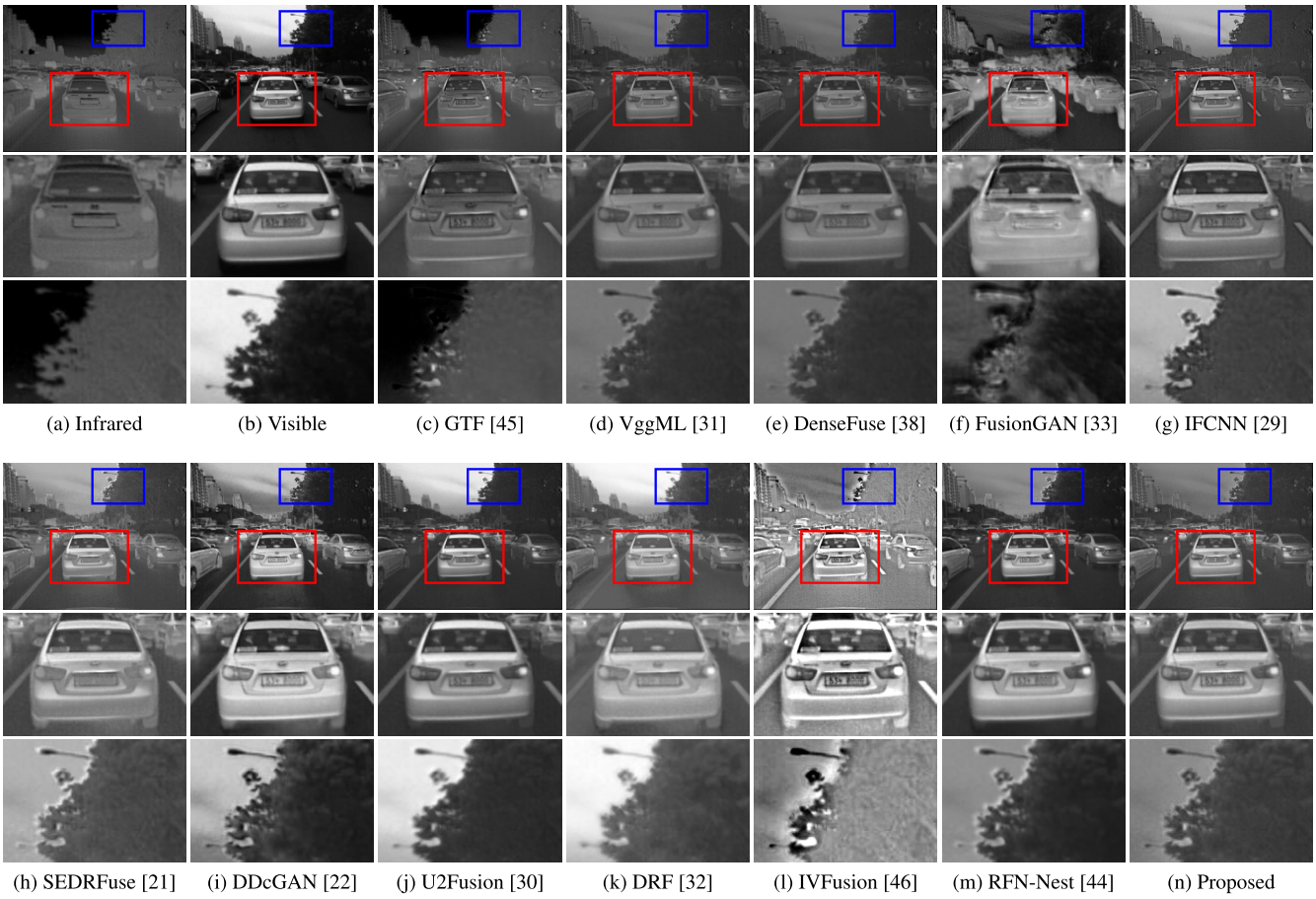


FIGURE 9. Comparison of fusion results and their magnified parts on the KAIST dataset. The second and third rows show the magnified parts for the red and blue rectangles, respectively, in the first row.

visual information from the visible images. In Figures 10(d), (e), (j), and (m), VggML, DenseFuse, U2Fusion, and RFN-Nest provide fused images with less artifacts, but the results are blurred, losing texture information. FusionGAN, DDcGAN, and IVFusion in Figures 10(f), (i), and (l), respectively, generate severe noise components, degrading the image quality. On the contrary, the proposed algorithm in Figure 10(n) preserves the fine textures in the source images faithfully, *e.g.*, the bushes in the third row.

Finally, Figure 11 shows the fusion results of the RoadScene dataset; they exhibit similar tendencies to the results in Figures 9 and 10. In Figures 11(c), (d), (e), and (g), GTF, VggML, DenseFuse, and IFCNN provide poor detail. FusionGAN, SEDRFuse, DDcGAN, DRF, and IVFusion in Figures 11(f), (h), (i), (k), and (l), respectively, provide fusion results with severe artifacts and noise that degrade the quality of images. U2Fusion in Figure 11(j) preserves the texture information in the visible images, but loses the background information in the infrared images, *e.g.*, the clouds in the second row. RFN-Nest in Figure 11(m) loses the object contours in the fused images, *e.g.*, the car and windows in the third row. In contrast, the proposed algorithm in Figure 11(n) provides fused images that preserve fine details in each source image.

TABLE 2. Impacts of edge-guided attention maps on fusion performance. Average rankings of fusion results are reported.

Edge-guided maps		Avg. rank
IRNet	FusionNet	
		2.50
✓		3.00
	✓	2.75
✓	✓	1.75

E. ABLATION STUDIES

We conduct several ablation studies to analyze the effects of the key components of the proposed algorithm on fusion performance. All experiments are performed for the TNO dataset [68] using all metrics used in the previous section to compute the average rankings.

1) EDGE-GUIDED ATTENTION MAPS

To analyze the effectiveness of the edge-guided attention maps, we train the proposed networks using different settings. Table 2 compares the results. The absence of an edge-guided attention map provides poor results because

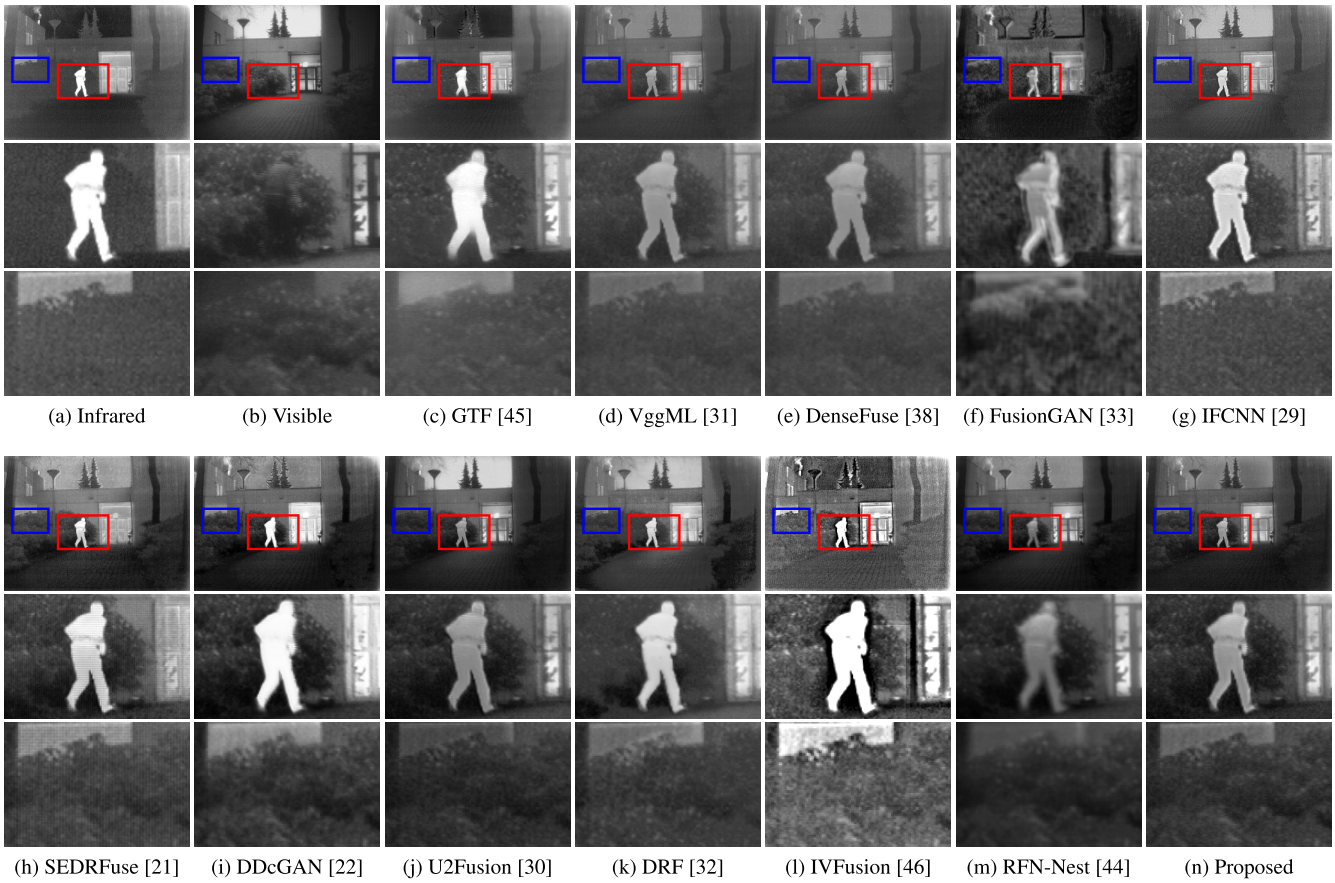


FIGURE 10. Comparison of fusion results and their magnified parts on the TNO dataset. The second and third rows show the magnified parts for the red and blue rectangles, respectively, in the first row.

the complementary edge information in the source images cannot be fully exploited. Using an edge-guided attention map in only one of the two networks worsens the performance because only the edge information of a single image is emphasized, which causes the networks to generate fusion results biased toward the image with the attention map. Finally, using edge-guided attention maps in both networks yields the best performance by selectively focusing on complementary edge information.

Figure 12 visually compares the fusion results. Using an edge-guided attention map in either of the two networks generates biased results toward the source images, as shown in Figures 12(d) and (e). Using edge-guided attention maps in both networks achieves the best performance, as shown in Figure 12(f), by forcing the networks to focus on the complementary edge information in the source images.

2) FUSION STRATEGIES

We analyze the effectiveness of the proposed adaptive channel fusion described in Section III-C by training the proposed networks with different fusion strategies. We choose five conventional handcrafted fusion strategies, as described in [59]. Table 3 compares the fusion performances. The proposed fusion strategy outperforms all the handcrafted fusion strategies, because it adaptively fuses features by considering the statistical characteristics of the source images using the input histograms.

TABLE 3. Impacts of the fusion strategies on the fusion performance. Average rankings of fusion results are reported.

	Product	Concat.	Max	Mean	Addition	Proposed
Avg. rank	4.25	3.50	3.75	3.75	3.25	2.50

3) LOSS FUNCTIONS

We train IRNet and FusionNet using different combinations of losses to analyze the effectiveness of each loss function. Table 4 quantitatively compares the results. First, using only ($\mathcal{L}_{\text{id}}$, $\mathcal{L}_{\text{fd}}$) provides the worst performance. Second, $\mathcal{L}_{\text{s}}$ improves the fusion performance. Third, the addition of either $\mathcal{L}_{\text{sp}}$ or $\mathcal{L}_{\text{p}}$ significantly improves the fusion performance. Finally, the combination of all the losses yields the best fusion performance by a large margin.

4) NETWORK LEVEL

To analyze the effectiveness of the levels of the proposed networks, we train the proposed networks with different levels.

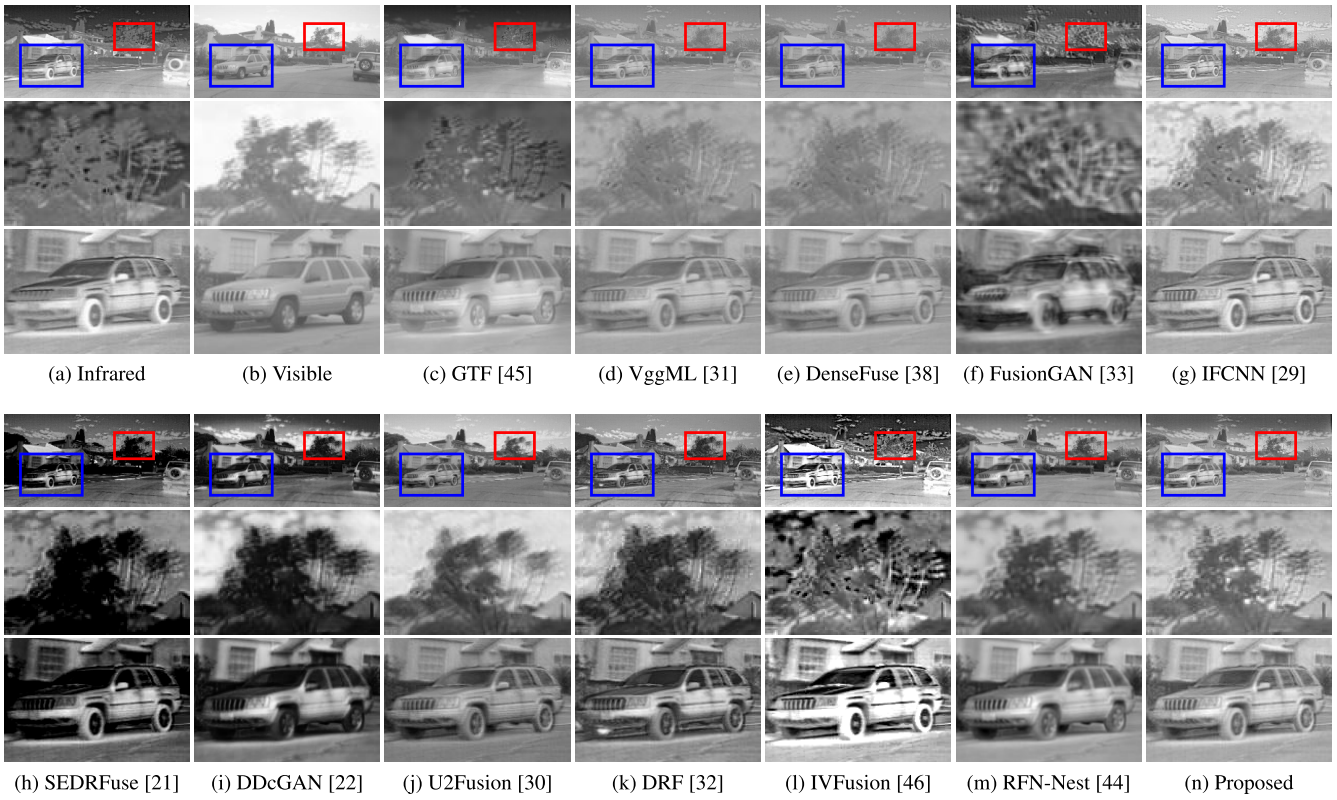


FIGURE 11. Comparison of fusion results and their magnified parts on the RoadScene dataset. The second and third rows show the magnified parts for the red and blue rectangles, respectively, in the first row.

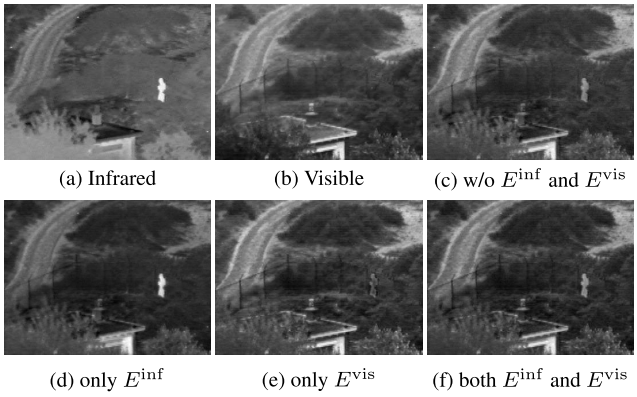


FIGURE 12. Comparison of fusion results according to the different settings of the edge-guided maps.

TABLE 4. Impacts of the losses on the fusion performance. Average rankings of fusion results are reported.

$\mathcal{L}_{\text{id}}, \mathcal{L}_{\text{fd}}$	$\mathcal{L}_{\text{s}}$	$\mathcal{L}_{\text{sp}}$	$\mathcal{L}_{\text{p}}$	Avg. rank
✓				3.88
✓	✓			3.75
✓	✓	✓		2.75
✓	✓		✓	3.00
✓	✓	✓	✓	1.63

TABLE 5. Impacts of the network level *N* on the fusion performance. Average rankings of fusion results are reported.

N	1	2	3	4	5
Avg. rank	3.88	3.63	2.00	2.75	2.75

Table 5 compares the fusion performances. As the network level N increases, performance improves by extracting more meaningful features. However, increasing the level excessively decreases fusion performance. This is because less structural features are extracted from excessively small images, which are then fed to the next level, causing the propagation less informative features.

5) MDRB

To analyze the effectiveness of the proposed MDRB, we train the proposed networks using three feature extraction blocks: MSRB [60], short skip connection (SSC) [76], and the proposed MDRB. Table 6 compares the fusion performance of each feature extraction block. The proposed MDRB yields the best fusion performance because it captures global information better with larger receptive fields than MSRB and SSC, while requiring a slightly larger number of parameters than SSC.

In addition, Table 7 compares the fusion performance according to the number of MDRBs. As the number of MDRBs increases, performance improves, because more salient features can be extracted. However, increasing the number of MDRBs excessively saturates the performance,

(a) (0.4, 0.6) (b) (0.45, 0.55) (c) (0.5, 0.5) (d) (0.55, 0.45) (e) (0.6, 0.4)

FIGURE 13. Comparison of fusion results of the proposed MPFusion algorithm for different combinations of (w_{inf}, w_{vis}) values.

TABLE 6. Comparison of fusion performance using the three feature extraction blocks. Average rankings of fusion results and the number of parameters are reported.

	MSRB	SSC	MDRB
Avg. rank	2.38	2.00	1.63
# Params (M)	3.71	1.69	2.37

TABLE 7. Impact of the number of MDRBs on the fusion performance. Average rankings of fusion results are reported.

# MDRBs	1	2	3	4	5
Avg. rank	3.50	2.50	2.13	2.88	4.00

TABLE 8. Impact of the training strategy on the fusion performance. Average rankings of fusion results are reported.

Training strategy	Avg. rank
Joint training	1.75
Separate training	1.25

while still increasing the required computational and memory complexities.

6) TRAINING STRATEGIES

As mentioned in Section III, IRNet is first trained separately, and then FusionNet is trained with fixed IRNet. To analyze the effectiveness of this separate training strategy, we train the proposed networks using different training strategies. Table 8 compares the fusion performances. The separate training strategy provides higher fusion performance than joint training. This is because separate training focuses on extracting the intrinsic features of each source image, which better preserves complementary information in the source images during fusion.

F. EFFECTS OF PARAMETERS w_{inf} AND w_{vis} ON FUSION PERFORMANCE

As discussed in Section III-E, the parameters w_{inf} and w_{vis} in (8)–(10) control the contributions of the infrared and visible images, respectively, to the fused image. We evaluate the effects of these parameters on the fusion performance.

TABLE 9. Comparisons of computational complexity in terms of the number of network parameters, GFLOPs, and runtime in seconds.

	# Params (M)	GFLOPs	Runtime
DenseFuse [38]	0.07	57.76	0.0293
FusionGAN [33]	0.93	588.83	0.0854
IFCNN [29]	0.08	4.25	0.0130
SEDRFuse [21]	3.47	146.86	0.1177
DDcGAN [22]	1.10	1058.03	0.1372
U2Fusion [30]	0.66	432.18	0.0767
DRF [32]	16.05	5158.29	0.1854
RFN-Nest [44]	4.79	82.74	0.0334
Proposed (MPFusion)	2.37	177.61	0.0929

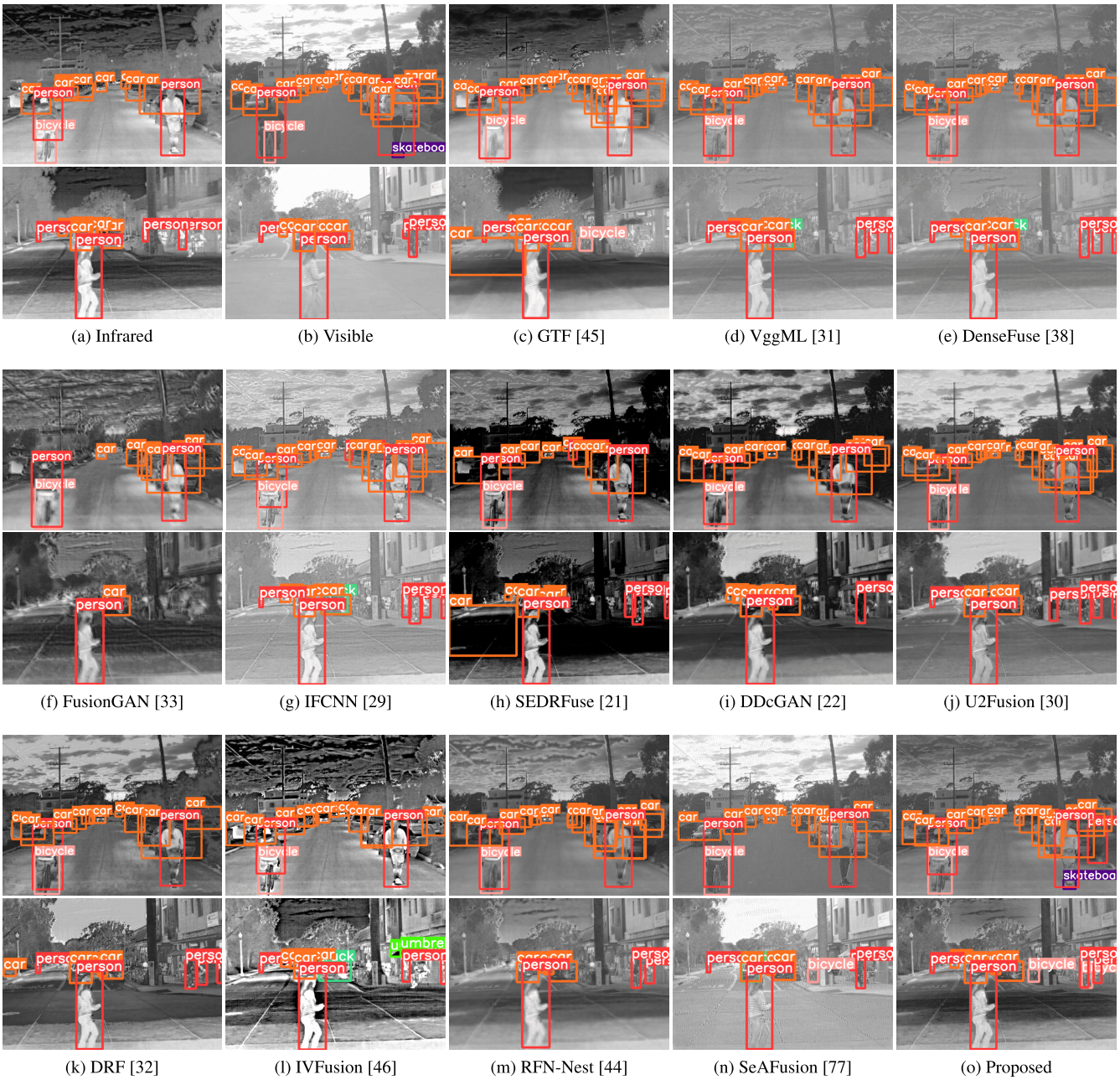
Figure 13 shows the fused images for several combinations of w_{inf} and w_{vis}. The fusion performance is considerably affected by the values of w_{inf} and w_{vis}. More specifically, when $(w_{\text{inf}}, w_{\text{vis}}) = (0.4, 0.6)$, the fusion results contain information mainly from visible images. However, as w_{inf} increases and w_{vis} decreases, the infrared images contribute to the fusion results more aggressively. This indicates that the selection of w_{inf} and w_{vis} significantly affects fusion performance. Therefore, to achieve the best fusion performance, we fixed both w_{inf} and w_{vis} to 0.5 in this work.

G. COMPLEXITY ANALYSIS

Table 9 compares the computational complexity in terms of the average runtime and number of giga floating-point operations per second (GFLOPs) to synthesize 200 paired KAIST images with a resolution 640 × 512 on an Nvidia RTX 2080Ti GPU and the number of network parameters. Although the proposed MPFusion consists of two networks, IRNet and FusionNet, it enables a graceful tradeoff between fusion performance and computational complexity.

H. OBJECT DETECTION PERFORMANCE EVALUATION

To verify the effectiveness of the proposed MPFusion algorithm in improving the performance of computer vision tasks, we apply an object detection algorithm to the fusion results obtained by each algorithm. Specifically, we use a YOLOv4



(a) Infrared (b) Visible (c) GTF [45] (d) VggML [31] (e) DenseFuse [38]

(f) FusionGAN [33] (g) IFCNN [29] (h) SEDRFuse [21] (i) DDcGAN [22] (j) U2Fusion [30]

(k) DRF [32] (l) IVFusion [46] (m) RFN-Nest [44] (n) SeAFusion [77] (o) Proposed

FIGURE 14. Comparison of object detection results on the RoadScene dataset.

model [78] pretrained using the COCO dataset [79] for the evaluation. In addition, in this evaluation, we also compare object detection performance on the fusion results of SeAFusion [77], which is an infrared and visible image fusion algorithm dedicated to high-level vision tasks.

Figure 14 shows examples of object detection results on the RoadScene dataset. The fusion results obtained by the proposed algorithm in Figure 14(o) yield an improved detection performance compared with either the infrared or visible images in Figures 14(a) and (b), respectively. For example, the bicycles, which are not detected in the second row of Figures 14(a) and (b), are detected in Figure 14(o). In addition, the fusion results of the proposed algorithm yield better detection performance than all the competing algorithms. For example, the skateboard in the first row of Figure 14(b) is detected only in Figure 14(o).

Figure 15 compares the precision-recall (PR) curves and reports the mean average precision (mAP) performance for the fusion results of each algorithm. A higher mAP value implies more accurate detection. The results show that the proposed algorithm achieves the best detection performance, *i.e.*, the highest mAP score. Further, the proposed algorithm exhibits 22.45% and 9.20% higher mAP values than the input infrared and visible images, respectively. Finally, it is worth noting that the proposed algorithm yields better performance than the dedicated SeAFusion [77]. These results indicate

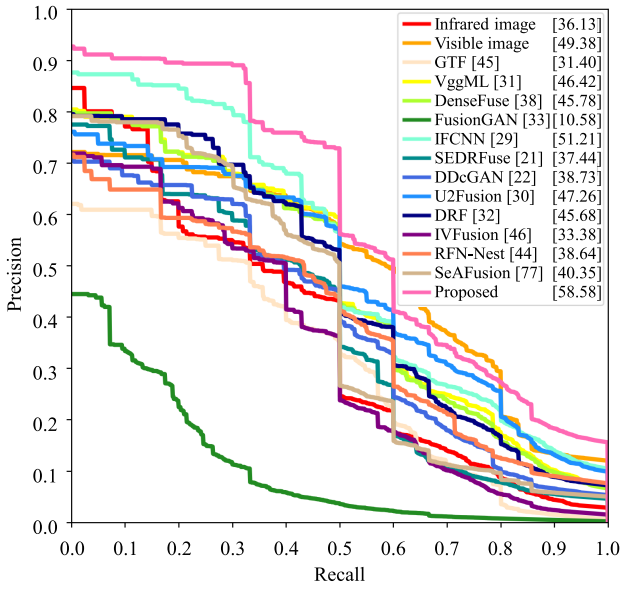


FIGURE 15. Comparison of PR curves and mAP values (%) for different fusion algorithms on the RoadScene dataset.

(a) Infrared images (b) Visible images (c) Fused images

FIGURE 16. Fusion results of the proposed algorithm for RGB visible images. The first and second rows show the results on the KAIST dataset, whereas the third and fourth rows show those on the RoadScene dataset.

that the fusion results of the proposed MPFusion algorithm consistently improve object detection performance. Hence, the proposed algorithm may exhibit positive impacts on computer vision applications under severe environmental conditions.

I. FUSION RESULTS FOR RGB IMAGES

We developed the proposed algorithm to fuse single-channel visible and infrared images for consistency with the TNO dataset [68]. However, the proposed algorithm can also be applied to fuse the visible and infrared images of RGB channels. To this end, we compute edge-guided attention maps in Figure 4 for each of the RGB channels for both visible and infrared images. Figure 16 shows the fusion results for images in the KAIST and RoadScene datasets. The color information of the visible images and the edge information of the infrared images are accurately preserved in the fused images.

V. CONCLUSION

We proposed a multiscale progressive fusion algorithm, called MPFusion, for infrared and visible image fusion. The proposed MPFusion algorithm consists of two networks: IRNet extracts multiscale features of the infrared image, and FusionNet extracts multiscale features of the visible image and progressively fuses them with those from IRNet. Specifically, we developed MDRB and PFB to improve fusion performance by progressively incorporating the multiscale features extracted from IRNet with those from FusionNet. Finally, we further improved the fusion performance by preserving the complementary edge information in the source images during fusion based on edge-guided attention maps. Extensive experiments demonstrated that the proposed algorithm outperforms state-of-the-art algorithms on several datasets. An important direction for future work is to develop more effective and sophisticated fusion schemes that can facilitate high-level vision tasks, such as segmentation, object detection, and re-identification.

REFERENCES

[1] C. L. Li, A. Lu, A. H. Zheng, Z. Tu, and J. Tang, "Multi-adapter RGBT tracking," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 2262–2270.
[2] J. Ma, Y. Ma, and C. Li, "Infrared and visible image fusion methods and applications: A survey," *Inf. Fusion*, vol. 45, pp. 153–178, Jan. 2019.
[3] P. Sun, W. Zhang, H. Wang, S. Li, and X. Li, "Deep RGB-D saliency detection with depth-sensitive attention and automatic multi-modal fusion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 1407–1417.
[4] H. Zhang, H. Xu, X. Tian, J. Jiang, and J. Ma, "Image fusion meets deep learning: A survey and perspective," *Inf. Fusion*, vol. 76, pp. 323–336, Dec. 2021.
[5] L. Jiao, R. Zhang, F. Liu, S. Yang, B. Hou, L. Li, and X. Tang, "New generation deep learning for video object detection: A survey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 8, pp. 3195–3215, Aug. 2022.
[6] R. E. Rivadeneira et al., "Thermal image super-resolution challenge results-PBVS 2022," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2022, pp. 418–426.
[7] C. Li, X. Liang, Y. Lu, N. Zhao, and J. Tang, "RGB-T object tracking: Benchmark and baseline," *Pattern Recognit.*, vol. 96, Dec. 2019, Art. no. 106977.
[8] X. Zhang, P. Ye, H. Leung, K. Gong, and G. Xiao, "Object fusion tracking based on visible and infrared images: A comprehensive review," *Inf. Fusion*, vol. 63, pp. 166–187, Nov. 2020.
[9] Q. Zhang, N. Huang, L. Yao, D. Zhang, C. Shan, and J. Han, "RGB-T salient object detection via fusing multi-level CNN features," *IEEE Trans. Image Process.*, vol. 29, pp. 3321–3335, 2020.
[10] Q. Zhang, T. Xiao, N. Huang, D. Zhang, and J. Han, "Revisiting feature fusion for RGB-T salient object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 5, pp. 1804–1818, May 2021.
[11] J. Ma, L. Tang, M. Xu, H. Zhang, and G. Xiao, "STDFusionNet: An infrared and visible image fusion network based on salient target detection," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–13, 2021.
[12] W. Su, Y. Huang, Q. Li, F. Zuo, and L. Liu, "Infrared and visible image fusion based on adversarial feature extraction and stable image reconstruction," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–14, 2022.

[13] S. Li, B. Yang, and J. Hu, "Performance comparison of different multi-resolution transforms for image fusion," *Inf. Fusion*, vol. 12, no. 2, pp. 74–84, Apr. 2011.

[14] H. Li, L. Liu, W. Huang, and C. Yue, "An improved fusion algorithm for infrared and visible images based on multi-scale transform," *Infr. Phys. Technol.*, vol. 74, pp. 28–37, Jan. 2016.

[15] S. Li, H. Yin, and L. Fang, "Group-sparse representation with dictionary learning for medical image denoising and fusion," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 12, pp. 3450–3459, Dec. 2012.

[16] X. Lu, B. Zhang, Y. Zhao, H. Liu, and H. Pei, "The infrared and visible image fusion algorithm based on target separation and sparse representation," *Infr. Phys. Technol.*, vol. 67, pp. 397–407, Nov. 2014.

[17] M. Yin, P. Duan, W. Liu, and X. Liang, "A novel infrared and visible image fusion algorithm based on shift-invariant dual-tree complex shearlet transform and sparse representation," *Neurocomputing*, vol. 226, pp. 182–191, Feb. 2017.

[18] Z. Zhu, H. Yin, Y. Chai, Y. Li, and G. Qi, "A novel multi-modality image fusion method based on image decomposition and sparse representation," *Inf. Sci.*, vol. 432, pp. 516–529, Mar. 2018.

[19] H. Li, X.-J. Wu, and J. Kittler, "MDLatLRR: A novel decomposition method for infrared and visible image fusion," *IEEE Trans. Image Process.*, vol. 29, pp. 4733–4746, 2020.

[20] H. Yan, J.-X. Zhang, and X. Zhang, "Injected infrared and visible image fusion via L_1 decomposition model and guided filtering," *IEEE Trans. Comput. Imag.*, vol. 8, pp. 162–173, 2022.

[21] L. Jian, X. Yang, Z. Liu, G. Jeon, M. Gao, and D. Chisholm, "SEDRFuse: A symmetric encoder–decoder with residual block network for infrared and visible image fusion," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–15, 2021.

[22] J. Ma, H. Xu, J. Jiang, X. Mei, and X.-P. Zhang, "DDcGAN: A dual-discriminator conditional generative adversarial network for multi-resolution image fusion," *IEEE Trans. Image Process.*, vol. 29, pp. 4980–4995, 2020.

[23] R. Hou, D. Zhou, R. Nie, D. Liu, L. Xiong, Y. Guo, and C. Yu, "VIF-Net: An unsupervised framework for infrared and visible image fusion," *IEEE Trans. Comput. Imag.*, vol. 6, pp. 640–651, 2020.

[24] H. Jung, Y. Kim, H. Jang, N. Ha, and K. Sohn, "Unsupervised deep image fusion with structure tensor representations," *IEEE Trans. Image Process.*, vol. 29, pp. 3845–3858, 2020.

[25] H. Zhang, H. Xu, Y. Xiao, X. Guo, and J. Ma, "Rethinking the image fusion: A fast unified image fusion network based on proportional maintenance of gradient and intensity," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 7, pp. 12797–12804.

[26] H. Li, X.-J. Wu, and T. S. Durrani, "Infrared and visible image fusion with ResNet and zero-phase component analysis," *Infr. Phys. Technol.*, vol. 102, Nov. 2019, Art. no. 103039.

[27] H. Zhang and J. Ma, "SDNet: A versatile squeeze-and-decomposition network for real-time image fusion," *Int. J. Comput. Vis.*, vol. 129, no. 10, pp. 2761–2785, Oct. 2021.

[28] X. Liu, L. Li, F. Liu, B. Hou, S. Yang, and L. Jiao, "GAFNet: Group attention fusion network for PAN and MS image high-resolution classification," *IEEE Trans. Cybern.*, vol. 52, no. 10, pp. 10556–10569, Oct. 2022.

[29] Y. Zhang, Y. Liu, P. Sun, H. Yan, X. Zhao, and L. Zhang, "IFCNN: A general image fusion framework based on convolutional neural network," *Inf. Fusion*, vol. 54, pp. 99–118, Feb. 2020.

[30] H. Xu, J. Ma, J. Jiang, X. Guo, and H. Ling, "U2Fusion: A unified unsupervised image fusion network," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 1, pp. 502–518, Jan. 2022.

[31] H. Li, X.-J. Wu, and J. Kittler, "Infrared and visible image fusion using a deep learning framework," in *Proc. Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2018, pp. 2705–2710.

[32] H. Xu, X. Wang, and J. Ma, "DRF: Disentangled representation for visible and infrared image fusion," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–13, 2021.

[33] J. Ma, W. Yu, P. Liang, C. Li, and J. Jiang, "FusionGAN: A generative adversarial network for infrared and visible image fusion," *Inf. Fusion*, vol. 48, pp. 11–26, Aug. 2019.

[34] J. Li, H. Huo, C. Li, R. Wang, and Q. Feng, "AttentionFGAN: Infrared and visible image fusion using attention-based generative adversarial networks," *IEEE Trans. Multimedia*, vol. 23, pp. 1383–1396, 2021.

[35] J. Li, H. Huo, C. Li, R. Wang, C. Sui, and Z. Liu, "Multigrained attention network for infrared and visible image fusion," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–12, 2021.

[36] Y. Yang, J. Liu, S. Huang, W. Wan, W. Wen, and J. Guan, "Infrared and visible image fusion via texture conditional generative adversarial network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 12, pp. 4771–4783, Dec. 2021.

[37] H. Zhang, J. Yuan, X. Tian, and J. Ma, "GAN-FM: Infrared and visible image fusion using GAN with full-scale skip connection and dual Markovian discriminators," *IEEE Trans. Comput. Imag.*, vol. 7, pp. 1134–1147, 2021.

[38] H. Li and X.-J. Wu, "DenseFuse: A fusion approach to infrared and visible images," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2614–2623, May 2019.

[39] J. Li, J. Zhu, C. Li, X. Chen, and B. Yang, "CGTF: Convolution-guided transformer for infrared and visible image fusion," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–14, 2022.

[40] W. Tang, F. He, and Y. Liu, "YDTR: Infrared and visible image fusion via Y-shape dynamic transformer," *IEEE Trans. Multimedia*, early access, Jul. 20, 2022, doi: 10.1109/TMM.2022.3192661.

[41] J. Ma, L. Tang, F. Fan, J. Huang, X. Mei, and Y. Ma, "SwinFusion: Cross-domain long-range learning for general image fusion via swin transformer," *IEEE/CAA J. Automat. Sinica*, vol. 9, no. 7, pp. 1200–1217, Jul. 2022.

[42] Z. Wang, Y. Chen, W. Shao, H. Li, and L. Zhang, "SwinFuse: A residual swin transformer fusion network for infrared and visible images," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–12, 2022.

[43] S. Park, A. G. Vien, and C. Lee, "Infrared and visible image fusion using bimodal transformers," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2022, pp. 1741–1745.

[44] H. Li, X.-J. Wu, and J. Kittler, "RFN-Nest: An end-to-end residual fusion network for infrared and visible images," *Inf. Fusion*, vol. 73, pp. 72–86, Sep. 2021.

[45] J. Ma, C. Chen, C. Li, and J. Huang, "Infrared and visible image fusion via gradient transfer and total variation minimization," *Inf. Fusion*, vol. 31, pp. 100–109, Sep. 2016.

[46] G. Li, Y. Lin, and X. Qu, "An infrared and visible image fusion method based on multi-scale transformation and norm optimization," *Inf. Fusion*, vol. 71, pp. 109–129, Jul. 2021.

[47] J. Han, E. J. Pauwels, and P. De Zeeuw, "Fast saliency-aware multi-modality image fusion," *Neurocomputing*, vol. 111, pp. 70–80, Jul. 2013.

[48] Y. Yang, Y. Que, S. Huang, and P. Lin, "Multiple visual features measurement with gradient domain guided filtering for multisensor image fusion," *IEEE Trans. Instrum. Meas.*, vol. 66, no. 4, pp. 691–703, Apr. 2017.

[49] S. Li, X. Kang, and J. Hu, "Image fusion with guided filtering," *IEEE Trans. Image Process.*, vol. 22, no. 7, pp. 2864–2875, Jul. 2013.

[50] Y. Liu, S. Liu, and Z. Wang, "A general framework for image fusion based on multi-scale transform and sparse representation," *Inf. Fusion*, vol. 24, pp. 147–164, Jul. 2015.

[51] Y. Cui, H. Du, and W. Mei, "Infrared and visible image fusion using detail enhanced channel attention network," *IEEE Access*, vol. 7, pp. 182185–182197, 2019.

[52] Y. Long, H. Jia, Y. Zhong, Y. Jiang, and Y. Jia, "RXDNFuse: A aggregated residual dense network for infrared and visible image fusion," *Inf. Fusion*, vol. 69, pp. 128–141, May 2021.

[53] J. Gao, L. Jiao, F. Liu, S. Yang, B. Hou, and X. Liu, "Multiscale curvelet scattering network," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Oct. 15, 2021, doi: 10.1109/TNNLS.2021.3118221.

[54] O. Ulutan, S. Rallapalli, M. Srivatsa, C. Torres, and B. S. Manjunath, "Actor conditioned attention maps for video action detection," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 527–536.

[55] J.-X. Zhao, J.-J. Liu, D.-P. Fan, Y. Cao, J. Yang, and M.-M. Cheng, "EGNet: Edge guidance network for salient object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8779–8788.

[56] Z. Wang, J. Wang, Y. Wu, J. Xu, and X. Zhang, "UNFusion: A unified multi-scale densely connected network for infrared and visible image fusion," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 6, pp. 3360–3374, Jun. 2022.

[57] J. Liu, X. Fan, J. Jiang, R. Liu, and Z. Luo, "Learning a deep multi-scale feature ensemble and an edge-attention guidance for image fusion," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 1, pp. 105–119, Jan. 2022.

[58] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 286–301.

[59] K. R. Prabhakar, V. S. Srikar, and R. V. Babu, "DeepFuse: A deep unsupervised approach for exposure fusion with extreme exposure image pairs," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4724–4732.

[60] J. Li, F. Fang, K. Mei, and G. Zhang, "Multi-scale residual network for image super-resolution," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 517–532.
[61] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, May 2016, pp. 1–13.
[62] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
[63] C. Li, C. Guo, and C. C. Loy, "Learning to enhance low-light image via zero-reference deep curve estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 8, pp. 4225–4238, Aug. 2022.
[64] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4681–4690.
[65] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, May 2015, pp. 1–14.
[66] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, May 2015, pp. 1–15.
[67] S. Hwang, J. Park, N. Kim, Y. Choi, and I. S. Kweon, "Multispectral pedestrian detection: Benchmark dataset and baseline," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1037–1045.
[68] A. Toet. (2014). *TNO Image Fusion Dataset*. [Online]. Available: https://figshare.com/articles/dataset/TNO_Image_Fusion_Dataset/1008029
[69] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, no. 3, pp. 379–423, Jul. 1948.
[70] C. S. Xydeas and V. Petrović, "Objective image fusion performance measure," *Electron. Lett.*, vol. 36, no. 4, pp. 308–309, 2000.
[71] V. Aslantas and E. Bendes, "A new image quality metric for image fusion: The sum of the correlations of differences," *AEU—Int. J. Electron. Commun.*, vol. 69, no. 12, pp. 1890–1896, 2015.
[72] K. Ma, K. Zeng, and Z. Wang, "Perceptual quality assessment for multi-exposure image fusion," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 3345–3356, Nov. 2015.
[73] M. Haghighat and M. A. Razian, "Fast-FMI: Non-reference image fusion metric," in *Proc. IEEE Int. Conf. Appl. Inf. Commun. Technol. (AICT)*, Oct. 2014, pp. 1–3.
[74] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a 'completely blind' image quality analyzer," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 209–212, Mar. 2013.
[75] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Trans. Image Process.*, vol. 21, no. 12, pp. 4695–4708, Dec. 2012.
[76] S. Anwar and N. Barnes, "Real image denoising with feature attention," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3155–3164.
[77] L. Tang, J. Yuan, and 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*, vol. 82, pp. 28–42, Jun. 2022.
[78] A. Bochkovskiy, C.-Y. Wang, and H.-Y. Mark Liao, "YOLOv4: Optimal speed and accuracy of object detection," 2020, *arXiv:2004.10934*.
[79] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2014, pp. 740–755.

SEONGHYUN PARK (Graduate Student Member, IEEE) received the B.S. degree in electrical, electronic, and control engineering from Hankyong National University, Anseong, South Korea, in 2020. He is currently pursuing the M.S. degree with the Department of Multimedia Engineering, Dongguk University, Seoul, South Korea. His research interests include image processing and computational imaging.

CHUL LEE (Member, IEEE) received the B.S., M.S., and Ph.D. degrees in electrical engineering from Korea University, Seoul, South Korea, in 2003, 2008, and 2013, respectively.

From 2002 to 2006, he was with Biospace Inc., Seoul, where he involved in the development of medical equipment. From 2013 to 2014, he was a Postdoctoral Scholar with the Department of Electrical Engineering, Pennsylvania State University, University Park, PA, USA. From 2014 to 2015, he was a Research Scientist with the Department of Electrical and Electronic Engineering, The University of Hong Kong, Hong Kong. From 2015 to 2019, he was an Assistant Professor with the Department of Computer Engineering, Pukyong National University, Busan, South Korea. In March 2019, he joined the Department of Multimedia Engineering, Dongguk University, Seoul, where he is currently an Associate Professor. His current research interests include image processing and computational imaging with an emphasis on restoration and high dynamic range imaging.

Dr. Lee received the Best Paper Award from the *Journal of Visual Communication and Image Representation*, in 2014. He is also an Editorial Board Member of the *Journal of Visual Communication and Image Representation*.