IR-MSDNet: Infrared and Visible Image Fusion Based On Infrared Features and Multiscale Dense Network

Asif Raza, Jingdong Liu, Yifan Liu, Jian Liu, Zeng Li, Xi Chen, Hong Huo, and Tao Fang

***Abstract*—Infrared (IR) and visible images are heterogeneous data, and their fusion is one of the important research contents in the remote sensing field. In the last decade, deep networks have been widely used in image fusion due to their ability to preserve high-level semantic information. However, due to the lower resolution of IR images, deep learning-based methods may not be able to retain the salient features of IR images. In this article, a novel IR and visible image fusion based on IR Features & Multiscale Dense Network (IR-MSDNet) is proposed to preserve the content and key target features from both visible and IR images in the fused image. It comprises an encoder, a multiscale decoder, a traditional processing unit, and a fused unit, and can capture incredibly rich background details in visible images and prominent target details in IR features. When the dense and multiscale features are fused, the background details are obtained by utilizing attention strategy, and then combined with complimentary edge features. While IR features are extracted by traditional quadtree decomposition and Bezier interpolation, and further intensified by refinement. Finally, both the decoded multiscale features and IR features are used to reconstruct the final fused image. Experimental evaluation with other state-of-the-art fusion methods validates the superiority of our proposed IR-MSDNet in both subjective and objective evaluation metrics. Additional objective evaluation conducted on the object detection (OD) task further verifies that the proposed IR-MSDNet has greatly enhanced the details in the fused images, which bring the best OD results.**

***Index Terms*—Feature attention, image fusion, multiscale feature fusion, object detection (OD), remote sensing.**

Manuscript received January 14, 2021; revised March 1, 2021; accepted March 2, 2021. Date of publication March 11, 2021; date of current version April 2, 2021. This work was supported by the National Key Research and Development Program of China under Grant 2018YFB0505000, in part by the National Science and Technology Major Project under Grant 21-Y20A06-9001-17/18, and in part by the Science Fund for Creative Research Groups of the National Natural Science Foundation of China under Grant 61221003. *(Corresponding author: Hong Huo.)*

Asif Raza, Jingdong Liu, Yifan Liu, Jian Liu, Zeng Li, Hong Huo, and Tao Fang are with the Department of Automation, and Key Laboratory of System Control and Information Processing, Ministry of Education, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: asifraza151@sjtu.edu.cn; sjtuljd@sjtu.edu.cn; liu-yifan@sjtu.edu.cn; zengli@sjtu.edu.cn; liujian@sjtu.edu.cn; huohong@sjtu.edu.cn; tfang@sjtu.edu.cn).

Xi Chen is with the Key Laboratory of Geographic Information Science (Ministry of Education), School of Geographic Sciences, Shanghai 200241, China, and also with the Shanghai Key Laboratory of Multidimensional Information Processing, East China Normal University, Shanghai 200062, China (e-mail: xchen@geo.ecnu.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2021.3065121

I. Introduction

REMOTE sensing image fusion has been studied for decades, because complementary information from multisource remote sensing images in the fused image is of great help to various remote sensing applications such as surveillance, object detection (OD), etc. [1], [2]. The fusion of infrared (IR) image and visible image is one of the important tasks in remote sensing field. Its main purpose is to extract features from multisource images, and then fuse them to generate fused images with prominent IR target and rich background details. Recently, many fusion methods have been proposed for this purpose, which can be broadly divided into traditional methods [3]–[7] and deep learning-based methods [8]–[12].

In addition to the spatial domain method, most traditional methods are based on signal processing techniques. These signal processing methods mainly include multiscale-based methods [3] and learning-based methods, which have achieved good results. Different from multiscale transformation, learning approaches are usually based on representation, such a as sparse representation [7] and dictionary learning [4]. Although these direct methods can avoid information loss during image fusion, they are usually complicated and time consuming, especially for online learning.

Recent advances in deep convolutional neural networks (CNN) in remote sensing has provided better potential for image fusion in learning and extracting high-level semantic information than traditional methods. Prabhakar *et al*. [8] proposed a novel CNN-based fusion framework for a multiexposure image fusion task. A fusion framework that utilizes multilevel deep features for image fusion was proposed [9]. Li *et al*. [12] proposed a fusion network using dense block in an autoencoder manner. Ma *et al*. [11] introduced a novel method for image fusion using a generative adversarial network (GAN) for IR and visible image fusion. Though the fusion performance of these CNN or GAN methods is better than existing methods, there are still some drawbacks in IR and visible image fusion.

First, common deep networks cannot efficiently extract salient features of IR images, because subsampling at each layer will weaken or smooth their features of IR images, and will be submerged in multiscale features of visible images. Second, because of the characteristic of IR image features, most network framework are not necessarily suitable for IR and visible image

fusion. Finally, IR features need to be further enhanced for image reconstruction after.

To overcome these above drawbacks, a novel IR and visible image fusion based on IR Features & Multiscale Dense Network (IR-MSDNet) is proposed, which makes full use of respective advantages of deep learning and traditional handcrafted feature, especially IR feature extraction, to obtain a better fusion result. It preserves full background details and key target features from both visible and IR images.

Major contributions of this article are described as follows:

1) In IR-MSDNet, an efficient encoder is designed for both visible and IR images, which is capable of preserving both the target details in IR images and the rich background details in visible images, and multiscale decoder can reconstruct the initial fusion image
2) IR features are extracted by traditional quadtree decomposition [13] and Bezier interpolation [14], and further enhanced by refinement. These IR features are combined with the initial fused image to produce the final fusion image directly. In this way, IR features will be not submerged in multiscale features of visible images.
3) To the best of our knowledge, it is the first time to combine deep learning with traditional methods for remote sensing image fusion, opening up a new perspective of visible and IR image fusion.

The rest of this article is organized as follows. Section II describes the related work. Section III introduces the proposed IR-MSDNet in details. Section IV includes the discussion made on the experiments, and results. Section V concludes this article.

II. Related Work

A. Traditional Fusion Methods

The spatial domain approach can be roughly divided into pixel-based, block-based, and region-based approaches. Pixel-based image fusion method extracts image features by preserving the spatial consistency of final fused images, such as dense scale-invariant feature transform (DSIFT) [19], image matting (IM) [20], and guided filtering (GF) [21]. In block-based image fusion method [22], [23], the images are divided into the same number of blocks, and then the blocks are fused by fusion rules. The number and size of blocks directly affect the fusion results. The region-based method depends on image segmentation, so its performance also depends on the efficiency of image segmentation [24], [25].

The multiscale transformation methods, as a typical representative of traditional fusion methods, are usually used to decompose the images into multiscale representations, and then fuse the multiscale representations according to certain fusion rules. Finally, the fused image is obtained by the inverse transformation of multiscale representations. Laplacian pyramid [15], discrete wavelet transforms [16], dual tree complex wavelet transforms [17], and curvelet transform [18], are example of among multiscale transformation methods.

In short, in order to improve the quality of fusion traditional fusion methods generally require more manual intervention, and the fusion rules adopted are relatively complex, therefore, there are inevitably problems such as low efficiency and high computational cost.

B. Deep Learning Based Fusion Methods

Deep learning method has attracted extensive attention since its appearance, and has been successfully applied to a wide range of remote sensing applications, such as image fusion [26]. Liu *et al*. [27] first utilized CNNs as backbone to achieve a rich fusion result based on a decision map indicating the rules of image fusion. Nonetheless, this method had a limitation of training strategy only for multifocus images. Li *et al*. [12] proposed a novel autoencoder network for image fusion. It includes an encoder, fusion layers, and a decoder. The encoder and decoder are trained by all input images, and then deep features extracted by the encoder are adaptively fused. Zhang *et al*. [10] proposed a general end-to-end fusion network, which is simple and effective to produce fused images, but lacks expertise in IR images due to generalizability for different types of images. Ma *et al*. [11] introduced a GAN architecture for IR and visible image fusion. During the training, the source images features were concatenated to the generator network, and the fused image was obtained. However, sometimes it is the strong adversarial ability of GAN that may cause the IR image to suppress the visible image with its content after image fusion. In a word, the abovementioned deep networks are generally designed, and cannot efficiently extract salient features from IR images, as their features will be weakened by down sampling due to their lower resolution.

III. Proposed IR-MSDNET

IR-MSDNet, as shown in Fig. 1, comprises an encoder, a multiscale decoder, a traditional processing unit, and a fused unit. Suppose IV and IR represent visible and IR images, respectively, where IV and IR images have been preregistered according to [9], and fed to the encoder and traditional processing unit. The encoder is used to extract dense multiscale features from visible and IR images for the initial image fusion, respectively. The multiscale decoder is designed to reconstruct the initial fused image with richer background detail. In addition, to increase the detail of the initial fused image, traditional processing unit is designed to extract edge features from visible and IR images, especially IR features with focusing on target details from IR images. Because the traditional processing method focuses on the IR image's target detail to extract the IR feature, visible image can be used to refine the IR feature. In the fused unit, the final fused image is generated by fusing initial fused image and IR feature.

Therefore, it can be seen from the Fig. 1 that, in our proposed IR-MSDNet, the visible and IR image fusion includes two main processes. One process is to construct an encoder and decoder network to realize the initial fusion of visible and IR images, the other is to further fuse IR image features extracted by the traditional methods with the initial fusion image to compensate for the loss of IR image details caused by CNN.

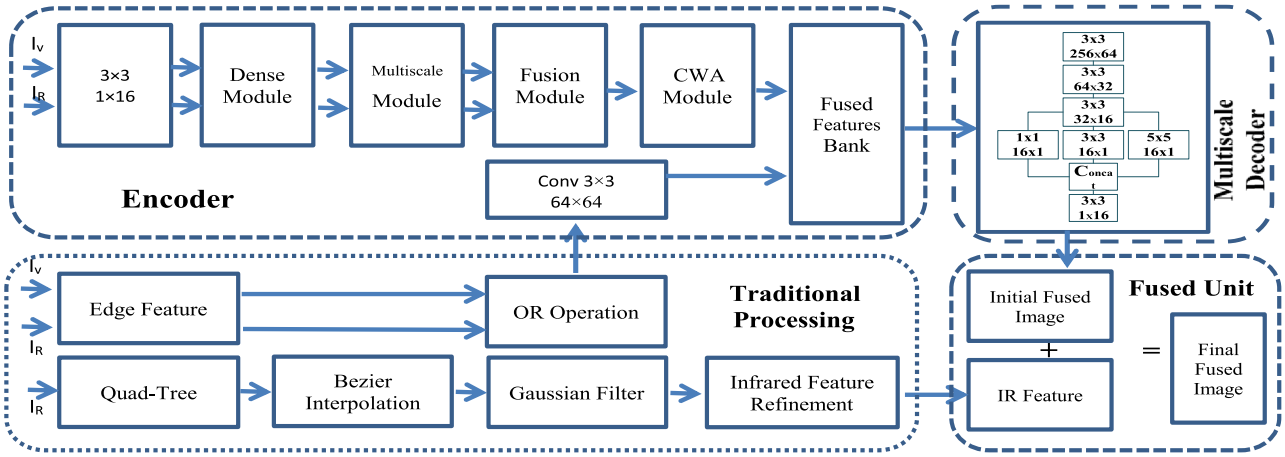


Fig. 1. Details of proposed IR-MSDNet fusion.

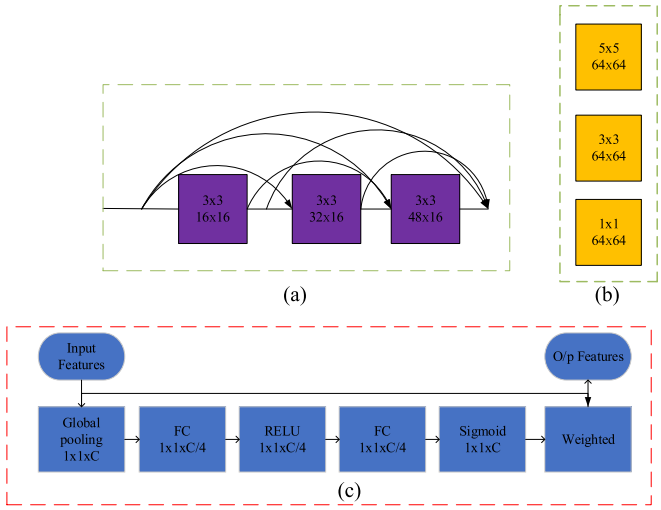


Fig. 2. Details of dense module, multi-scale module and attention module of IR-MSDNet.

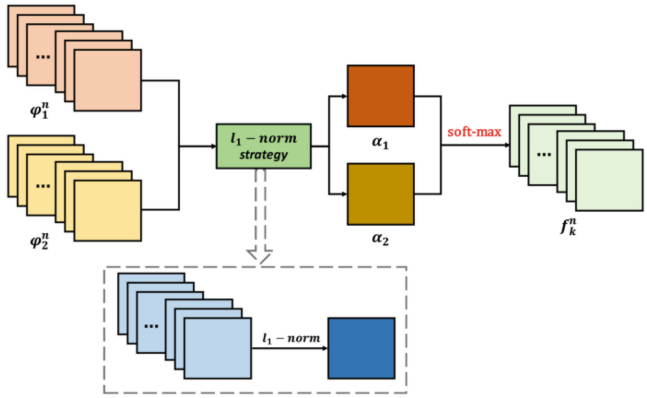


Fig. 3. Diagram of l1-norm and soft-max strategy.

A. Encoder

The encoder contains different modules for feature extraction, including dense module, multiscale module, feature fusion module, channel-wise attention (CWA) [28] module and fused feature bank. In order to receive a visible or IR image of arbitrary size, this image is first processed via the convolutional layer, where kernel size is 3×3, and the stride is 1. Before feature fusion, different multiscale features were extracted from visible and IR images, respectively, through dense module and multiscale module. In the encoder, details of dense module, multiscale module, and CWA module are shown in Fig. 2, respectively.

The dense module [shown in Fig. 2(a)] is made up of three cascaded convolutional layers, namely, the output of the previous layer is the input of the next layer. The convolution kernel of 3x3 is usually used to extract coarse features in this dense module. In multiscale module [shown in Fig. 2(b)], the size of convolution kernels varies from 5 × 5, 3 × 3 to 1 × 1, which not only preserves the details in the dense module, but also extracts features from rough to coarse. Therefore, these multiscale features are necessary for image fusion [29].

In the fusion module, l_1-norm and soft-max strategy [8] has been chosen for fusing the multiscale features, as shown in the Fig. 3. In order to improve fusion efficiency, block based averaging method is adopted to avoid any misregistration between multiscale feature maps, making fusion consistent. Let $\varphi_k^n\ (x,\ y)$ be multiscale features at $(x,\ y)$ position, where $k \in \{I_R, I_v\}$ corresponds to IR image or visible image, and $n \in \{1,\ 3,\ 5\}$ represents one of multiscale corresponding to certain kernel size. According to [30], the l_1-norm of $\varphi_k^n\ (x,\ y)$ is defined as the activity level measurement for fusing multiscale features. The initial activity level measurement ${\alpha_k}'$ will be calculated as follows:

$$ {\alpha_k}'\ (x,\ y) = \left|\left|\ \varphi_k^n\,(\mathrm{x},\mathrm{y})\right|\right|_1. \tag{1} $$

The final activity level measurement for entire multiscale features would be

$$ \alpha_k\,(x,\ y) = \frac{\sum_{\mathrm{a}\ =\ -\mathrm{p}}^{p} \sum_{\mathrm{b}\ =\ -\mathrm{r}}^{p} \alpha_k'\,(x+a, y+b)}{(2p+1)^2} \tag{2} $$

where p represents the size of entire block size, and it is recommended to choose a small value for p, such as $p = 1$.

Let f_k^n denotes fused feature map:

$$f_k^n(\mathrm{x},\mathrm{y}) = \sum_1^k w_k^n(x,\ y) * \varphi_k^n(\mathrm{x},\ \mathrm{y}) \tag{3}$$

$$w_k^n(\mathrm{x},\ \mathrm{y}) = \frac{\alpha_k(x,\ y)}{\sum_1^k \alpha_k(x,\ y)} \tag{4}$$

where w_k^n represents weight map, and $*$ represents convolution operation. Then, the sum of $f_{IR}^n(x,y) + f_{Iv}^n(x,y)$ is result of the dense multiscale fused features.

After the above feature fusion, to further fully exploit the dense multiscale fused features, CWA is used to further treat the fused features by acquiring different channel weights, which make rich the enhanced features with both more salient high-level features and more channel information. Actually, in CWA module [shown in Fig. 2(c)], CWA is carried out by a series of

Processes, which include the global pooling, FC (fully connection layer), RELU activation, FC (fully connection layer), and sigmoid activation. Finally, the combined weighted features are multiplied by the fused features to obtain the processed. Suppose F is fused features and F' is enhanced salient representation obtained by CWA, which can be expressed as follows:

$$F' = \sigma 1\left(f_c 2\left(\delta\ \left(f_c 1\left(g\left(F\right),\beta_1\right)\right),\ \beta_2\right)\right) \tag{5}$$

$$F' = C\ (F,\ \beta)\ \text{or}\ F\beta \tag{6}$$

where β_1 and β_2 are the weights of two FC with their tasks f_c1 and f_c2, respectively. g is the global pooling operator. β is the combine total weight for channel attention and δ denotes the RELU function. $C(.)$ denotes the channel -wise multiplication between fused feature map and total weight β.

As is known to all, in image data fusion, edge is one of the most important information of the fused image. However, in the CNN, the edge feature information of the image is easy to be blurred or smoothed by pooling operation, so that the reconstructed image edge details after feature fusion are not rich. Therefore, in addition to the above features from the CNN fusion and processed by CWA, edge features extracted by traditional methods from visible and IR images are also added to our model. In order to facilitate the subsequent decoder processing, such edge features of visible and IR images need to be processed through a fixed convolutional layer of size 3×3 after the fusion of OR operation (discussed in the following sections), as shown Fig. 1. In the fused feature bank, two types of above features are being collected in concatenated way to build up a rich and detailed encoder.

B. Multiscale Decoder

At the end of the encoder, complimentary edge features are concatenated with rich fused features from attention module in fused feature bank. The initial fused image is then reconstructed in multiscale decoder.

In our proposed multiscale decoder, there are five learnable convolution layers. Each convolution layer is enclosed by ReLU function. In order to avoid vanishing gradients, which sometimes occur in many networks, a smooth training a skip connection strategy is used for smooth training. Thus, the three layers are designed via skip connections. In this decoder, including the multiscale layers, a unified 3×3 kernel size has been implemented. Similar to encoder module, the same size of convolution kernels, which varies from $5 \times 5, 3 \times 3$ to 1×1, has been chosen to preserve the details of all scales. The details of multiscale decoder are shown in Fig. 1. All the feature maps with multiscale are concatenated before the final layers to reconstruct a rich initial fused image.

C. Training Encoder and Decoder Networks

In our IR-MSDNet, encoder (except fusion layer) and decoder networks are mainly trained. In the training stage, the aim is to obtain the optimal weights to train the autoencoder network, so that it has the ability of deeply rich feature extraction and more abundant reconstruction.

In the training period, the main key is to train the network to reconstruct the initial fused image from visible and IR source images. The dense and multiscale modules extract the features from the visible and IR images, respectively, and then multiscale features are concatenated and fed to the CWA module. Finally, the features via the CWA module and edge features are concatenated, forming the fused features bank, then fed to the multiscale decoder for decoding.

The training will enable the encoder and decoder to obtain their final parameters and weights by loss functions. Let the total loss L_T be the sum of two kind of loss functions. The first loss function is structural similarity (SSIM) loss, and the second pixel loss, which are describe as follows:

$$L_T = \lambda\ (1\ -\ \mathrm{SSIM}\left(\mathrm{Op};\ I\right)) + \left(\left|\left|\mathrm{Op} - I\right|\right|_2\right) \tag{7}$$

$$\mathrm{Lssim} = 1\ -\ \mathrm{SSIM}\left(\mathrm{Op};\ I\right), \tag{8}$$

$$Lp = \left|\left|\mathrm{Op} - I\right|\right|_2 \tag{9}$$

where Op and I denotes the respective fused image and source images. $||\cdot||$ is Frobenius norm. SSIM means the structural similarity of two images. λ represents the tradeoff parameter for total loss to build the final output images from source images. In addition, this tradeoff parameter is utilized to handle the efficiency factor in terms of early training of the network. Due to its importance, λ early training leads to the optimal weights with fast convergence. Its effect has been explained briefly in Section IV-C.

D. Traditional Processing

Different from the abovementioned extracting features of encoders, there are two modules in this traditional processing unit. The first important module is responsible for extracting edge features from visible and IR images, respectively, while the second module directly extracts IR features from IR image through a series of mini process modules.

Edge features are extracted by canny edge detectors [31] from both visible and IR images, and then combined by OR operation as shown in Fig. 4. In order to feed edge features to encoder, this combined edge feature is convoluted by the kernel of size 3×3 and forward toward to the fused feature bank.

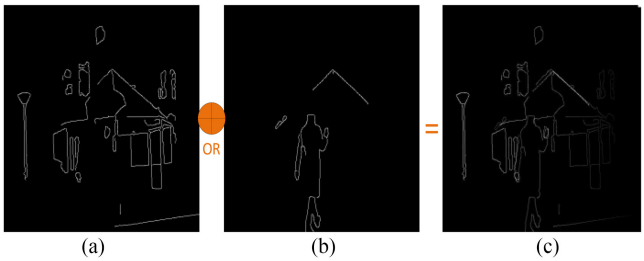


Fig. 4. Edge features from the visible image, IR image and the after OR operation from left to right. (a) Visible image. (b) Infrared image. (c) Result by OR operation.

The second and most important module in the traditional processing unit is IR feature extraction, which is also core heart of our scheme, as shown in Fig. 1. In order to extracts IR features directly, a series of processes include Quad-tree decomposition, Bezier interpolation, Gaussian filter, and IR feature refinement.

Initially, the quadtree decomposition technique [13] is adopted to pay more attention to the approximate outline of IR target entity. Because Quadtree decomposition is time-efficient, it helps to select appropriate control points, with which lots of noises can be actively suppressed. In quadtree decomposition, threshold T_{quad} and the area size are two vital parameters. T_{quad} is utilized to control whether the area size would be more decayed or not. Typically, a small upper limit is designated to prevent the variation. Usually the location of control points is expressed as $(a,\ b)$. These coordinates are consistently appraised from individual area in the quadtree framework.

The second step is to construct artificially background by Bezier interpolation [14]. Bezier interpolation is one of the best methods to reconstruct a large-scale matrix, which can be an image in our case. Thus, the method first interpolates to some identified control points, and then the interpolation can be adapted to estimate the contour of the object.

After that, the Bezier plane of individual area can be reconstructed through approximation of x and y coordinates and grey values. These approximations directly correspond to 16 control points:

$$Q\ (a,\ b)\ =\ \text{AMRO}^T B^T \tag{10}$$

where $(a;\ b)$ indicates the position of an interpolated point. *(A; B)* indicates the variable interpolation factor, which is connected to *(a; b)*. O indicates the constant interpolation factor matrix. R indicates 4×4 matrix with 16 control points. T stands for vector or matrix transpose $B,\ M$, and R are then defined as follows: $A\ =\ [a^3, a^2, a^1, a^0], B\ =\ [b^3, b^2, b^1, b^0,]$, where $0\ \leq\ a,\ b \leq 1$.

$$O = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -3 & 3 & 0 & 0 \\ 3 & -6 & 3 & 0 \\ -1 & 3 & -3 & 1 \end{bmatrix}$$

$$R = \begin{bmatrix} R_{11} & R_{12} & R_{13} & R_{14} \\ R_{21} & R_{22} & R_{23} & R_{24} \\ R_{31} & R_{32} & R_{33} & R_{34} \\ R_{41} & R_{42} & R_{43} & R_{44} \end{bmatrix}.$$

Although IR background (I_{BIR}) might be almost perfectly restored by joining the Bezier plane of individual area in the quadtree framework, the linear combination of all Bezier planes may hinder some IR key objects. Diverse control points are utilized in the sewed areas, therefore, the IR background is flattened by a Gaussian filter.

$$I_{\text{FBIR}} =\ I_{\text{BIR}} * g\,(s,\ \Phi) \tag{11}$$

where s and Φ represent the size and omega parameter of the Gaussian filter, respectively. In most cases, the linked Bezier areas are much similar. Therefore, a minor smoothing gradation can be reasonable for producing a flatten background. After that, a flatten and expected IR background image I_{FBIR} is obtained. Then, the bright IR could be easily obtained by difference between background image and IR image I_R.

$$\text{IR}\ =\ \max\ (I_R -\ I_{\text{FBIR}}) \tag{12}$$

To further refine IR features, IR features is subtracted from the cross product of the estimated background (difference between I_R and I_V) and a suitable minimizing ratio α. Consequently, a lot of useless background details can be almost removed, whereas the beneficial IR features are preserved.

$$\text{IR}'\ =\ \text{IR} - \alpha * \max\,(I_R - I_V, 0) \tag{13}$$

where α signifies the parameter that control the background degradation factor within a range of [0,1], and $\alpha = 0.6$ in our experiments. After the improvement and enhancement of IR features, data fusion with initial fused image can be carried out.

E. Data Fusion

In the fused unit, the final fusion image $I_{\text{Final Fused}}$ is obtained by pixel-level fusing of the initial fused image $I_{\text{Initial Fused}}$ with IR_{Final} features.

$$I_{\text{Final Fused}} =\ \text{IR}_{\text{Final}} + I_{\text{Intial Fused}} \tag{14}$$

IR_{Final} features are obtained by suppressing the initial IR features while preserving the visible information to overcome the fused image suffering from overexposure:

$$\text{IR}_{\text{Final}}\ =\ \forall *\ \text{IR}' \tag{15}$$

where $\forall$ denotes the feature suppression ratio and it can be calculated as follows:

$$\forall\ = \frac{\text{Avg}}{255} \tag{16}$$

where Avg denotes the mean average of the 0.5% highest of the addition of initial fused image and IR feature image. This process is like an average scaling of the grey intensities. After the improvement's steps, final IR features are now feasible for fusion. Fig. 5 and Fig. 6 show the complete process from IR features extraction to the final fusion.

IV. Experiments and Analysis

In this section, two remote sensing benchmark datasets are used in our experiments for verifying the effectiveness and

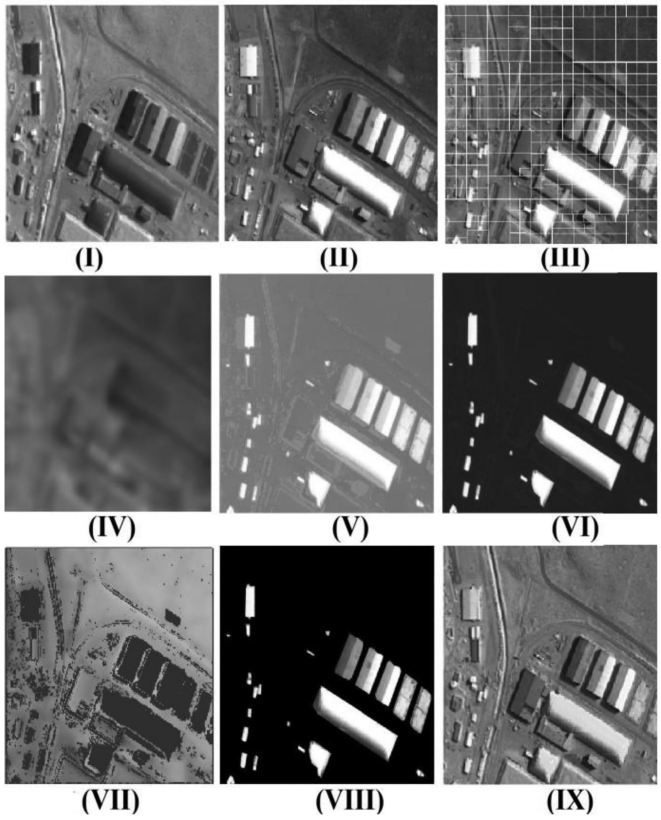


Fig. 5. Illustration of the IR feature extraction from start to the final fused image on Aerial dataset. (I), and (II) are a visible image and an IR image. (III) is Quadtree decomposition structure of the IR image. (IV) shows the self-constructed IR background by Quadtree structure and Bezier interpolation. (V) shows the initial IR features by subtracting (IV) from (II). (VI) shows the refined IR features by exposure suppression on (V). (VII) shows the valued background by subtracting (I) from (II). (VIII) shows the final IR features. (IX) shows the final fused image.

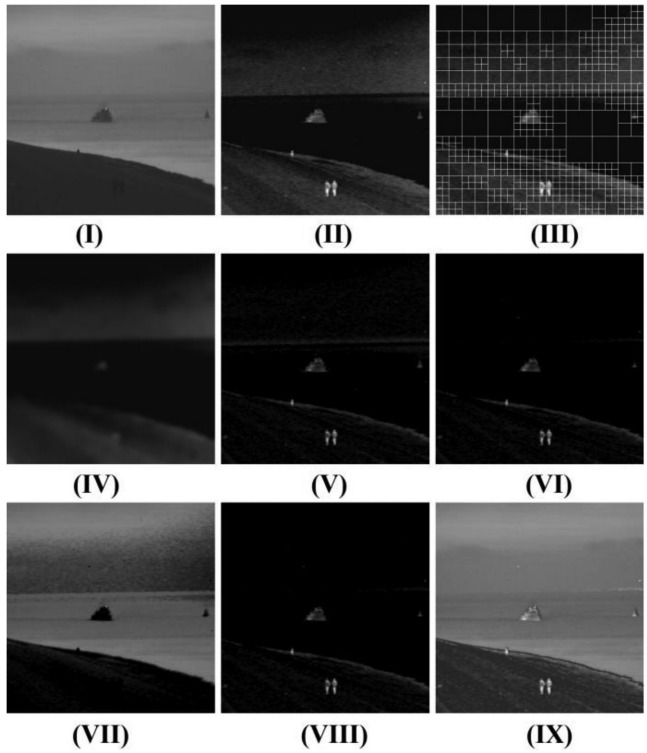


Fig. 6. Illustration of the IR feature extraction from start to the final fused image on TNO dataset. (I), and (II) are a visible image and an IR image. (III) is Quadtree decomposition structure of the IR image. (IV) shows the self-constructed IR background by Quadtree structure and Bezier interpolation. (V) shows the initial IR features by subtracting (IV) from (II). (VI) shows the refined IR features by exposure suppression on (V). (VII) shows the valued background by subtracting (I) from (II). (VIII) shows the final IR features. (IX) shows the final fused image.

superiority of the proposed IR-MSDNet subjectively, and then objectively by quantitative quality metrics with other state of arts fusion methods. An OD task is further performed to verify that the proposed IR-MSDNet has greatly enhanced the details in the fused images, which bring the best OD results.

A. Datasets

Two remotes sensing benchmark datasets are TNO [32][a] and Aerial Image dataset [4][b], which also commonly used by others algorithm in [6], [9]–[11]. TNO image fusion dataset comprises multispectral (visual and IR images) nighttime imagery related to different military circumstances, and they are registered with different multiband camera systems. The second Aerial image dataset of visible and IR images has been captured by a remote sensing platform. The size of each image on the Aerial image dataset is 512×512. Fig. 7 shows twenty pairs with different scenes on Aerial Image dataset, while Fig. 8 shows ten visible and IR image pairs of TNO dataset.

B. Compared Methods

In order to verify the effectiveness and superiority of our proposed IR-MSDNet, existing state-of-the-art traditional fusion methods have been considered for comparison. These methods include generalized joint sparse representation-based method (GJSR) [4], joint sparse representation-based method (JSR) [4], l0-generalized total variation model (GTVM) [6], JSR model with saliency detection method (JSRSD) [7], VGG–19 and multilayer fusion strategy (VggML) [9], a CNN-based fusion (DeepFuse) [8], GAN-based fusion algorithm (FusionGAN) [11], dense-block based fusion (DenseFuse) [12], and an end-to-end fusion network (IFCNN) [10]. Generally, the efficiency of the image fusion is typically assessed either subjectively or objectively. Most fusion metrics are usually based on features, such as edges and the amount of details, from the different source images into the fused image [33].

In this article, for objective evaluation, seven quantitative quality metrics are selected: entropy (En) [34]; Qabf [35] showing the quality of visual evidence found from the fusion image; FMIw and FMIdct [36] computing fast mutual information (FMI); a modified structural similarity SSIMa [37]; MS-SSIM [38] computing a modified structural similarity which only emphases on structural information, and to further analyses the quality of the fused image, and the standard deviation (SD) [39], which are utilized as quality metrics.

C. Related Parameters

In compared with the visible image dataset, the IR image dataset is relatively small, therefore, in this article, a publicly available larger dataset MS-COCO [40] is used for training

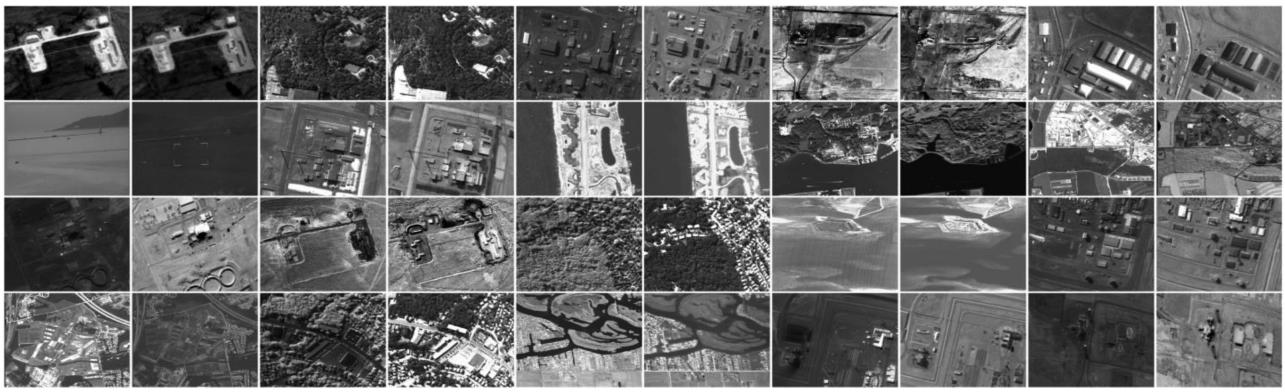
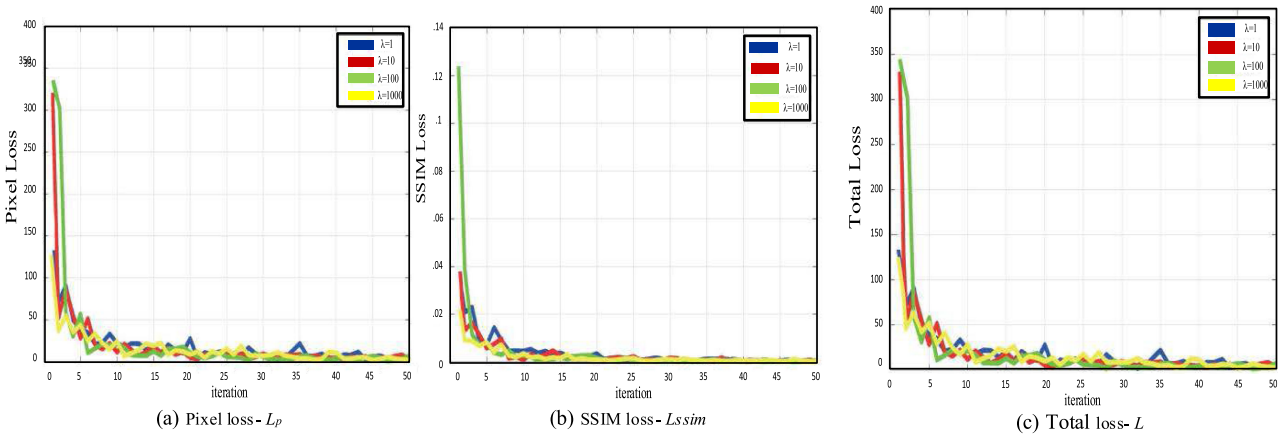


Fig. 7. Twenty pairs of test images from aerial image dataset. Each pair contains a visible image and an IR image from left to right.

Fig. 8. Images pairs form TNO dataset. Each pair contains a visible image and an IR image from left to right.



(a) Pixel loss- L_p (b) SSIM loss- L_{ssim} (c) Total loss- L

Fig. 9. Graph plot of pixel loss (a), SSIM loss (b) and total loss (c). x-axis indicates iterations, each single point represents 100 iteration. "blue" - $\lambda = 1$; "red" - $\lambda = 10$; "green" - $\lambda = 100$; "yellow" - $\lambda = 1000$.

the network following the convention set by other methods as [9], [12]. First, MS- COCO is converted to gray for training the network. Images are resized to 256 $\times$ 256 and converted to grayscale images. Learning rate, batch size and epochs are set as $1 * 10^{-4}$, 2 and 4, respectively. Some parameters for IR feature extraction are set as follows: such as, the threshold T_{quad} quadtree decomposition size, the kernel size $s = 11$ and sigma $\Phi = 5$ in Gaussian smoothing, the background suppression ratio, and the IR feature suppression ratio is set throughout the article as [41]. Fig. 9 shows the parameter in loss function evaluation, as discussed in the last para of Section III, the parameter $\lambda \in \{1, 10, 100, 1000\}$; it is observed that if λ is set to larger values the network converges faster. However, after 40 000 iterations, the optimal weights are achieved, no matter which loss weights are chosen. According to observation and feasibility, when λ is set 1000; the optimal weights are obtained after training the network, which means best values for quality metrics entropy (En), for the quality of visual evidence found from the fusion images (Q_{abf}), for computing fast mutual information (FMIw), and for a modified structural similarity (SSIMa).

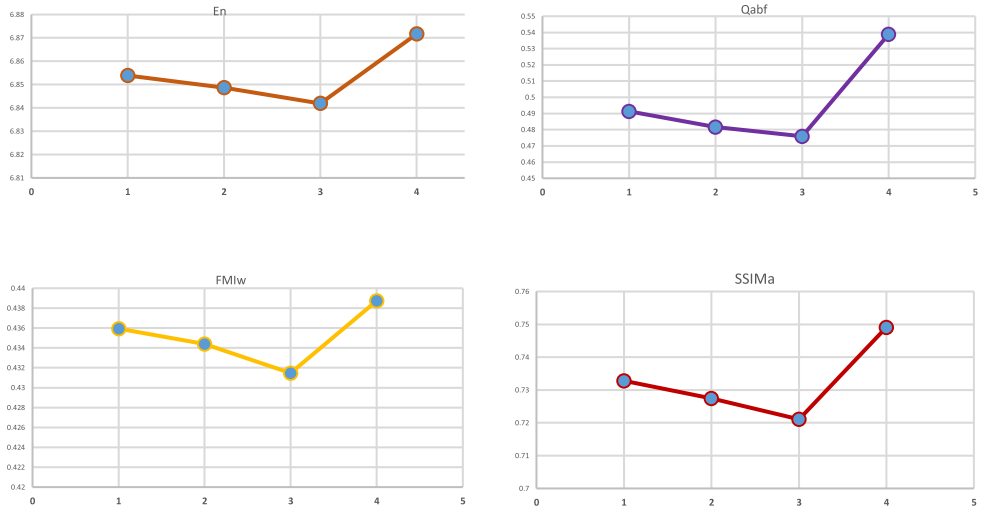


Fig. 10. Effect of different modules upon fusion results with different metrics (first row: En, Qabf ;second row: FMIw, SSIMa), the numbers on the *X*-axis in each graph represent four cases: *1 without multiscale features and CWA, *2 without edges features, *3 without IR features, and *4 IR-MSDNet.

TABLE I
EFFECT OF DIFFERENT MODULES UPON FUSION RESULTS WITH DIFFERENT METRICS

Condition	En	Qabf	FMIw	SSIMa
1	6.85381	0.49134	0.43594	0.73276
2	6.84864	0.48167	0.43438	0.72743
3	6.84189	0.47595	0.43147	0.72106
4	6.87164	0.53892	0.43874	0.74911

Condition *1 without multiscale features and CWA, *2 without edges features, *3 without IR features and *4 IR-MSDNet.

Graphical and numerical illustrations in Fig. 10, and Table I show the performance on TNO dataset with different modules of IR-MSDNet, respectively. There are four different condition cases: without multiscale features and CWA, without edges features, without IR features, and IR-MSDNet. The effectiveness of each module can be seen from their values of each metric in four cases. It is noticed that If without IR features, all metrics decline. It implies that IR features are the most important features. Moreover, edges feature has been also found to be more prominent than multiscale and CWA. Obviously, by taking all features into the fusion, the best fusion is given by IR-MSDNet.

D. Comparison and Analysis of Subjective Evaluations

To comprehensively evaluate the performance of IR MSDNet, traditional and deep learning-based methods (as mentioned in Section IV-B) are involved in these comparisons. Moreover, twenty image pairs of TNO dataset and Aerial image dataset are evaluated here. Figs. 11 (3)–(10) and 12 (3)–(10) are the fused images of "umbrella" from TNO and "Warehouse" from Aerial Image dataset.

Manifestly, the proposed IR-MSDNet differs from the rest comparative methods mainly in the target region and the region of the details in the background. In order to facilitate observation, a small region of the IR target is marked with a green frame, and the region of the details in the background is marked with a red frame. It can be seen that all the traditional fusion methods not only spot the target information existing in the IR images, but also contain some noises, which leads to blurred effects and not more salient in the fusion images. The fused images in Figs. 11 (3) and (4) and in 12 (4) and (5), hold several blocking artifacts, demonstrating in ringing around the salient features. Deep learning-based methods all have better human visual perception than traditional methods except Fig. 12(7), which produce a blur effect in Aerial Image dataset.

In a word, it can be seen from Figs. 11 and 12 that the proposed IR-MSDNet preserves more precise intensity information of the IR image and captures more textured from the visible image on both datasets.

E. Comparison and Analysis of Objective Evaluations

Objective evaluations for the proposed IR-MSDNet and all studied fusion methods is given in Table II on TNO dataset and Table III on Aerial image dataset. The value of the assessment metrics in boldface indicate the optimal ones and in underline for suboptimal ones. It can be seen from Table II that the objective evaluations are consistent with the conclusions from the subjective evaluations. In particular on TNO dataset, the proposed IR-MSDNet has optimal values for four of the seven metrics and suboptimal values for the other three. It also has the optimal and suboptimal values with least average value for $\mathrm{FMI}_{\mathrm{dct}}$ on aerial image dataset that can be seen from Table III. Different datasets may have slight inferior performances, but for Qabf, the proposed IR-MSDNet is considerably superior compared with other methods, which means that IR-MSDNet can produce fused images with rich details. The IR-MSDNet has significant advantages on image fidelity, which is consistent

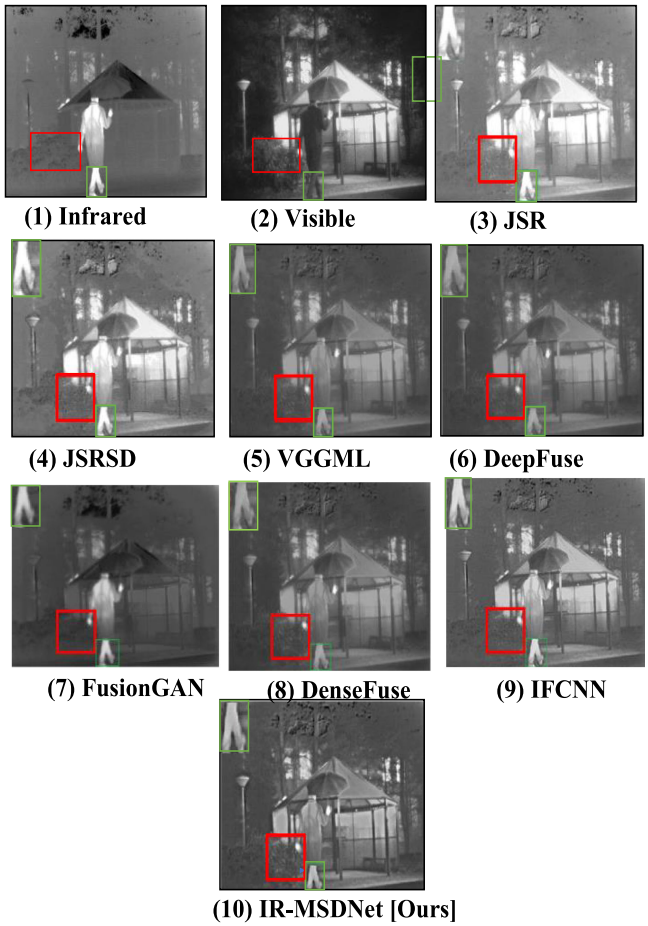


Fig. 11. Fusion results TNO dataset with different fusion methods. The detail fusion results are shown by red boxes.

TABLE II
OBJECTIVE EVALUATION INDICES FOR DIFFERENT FUSION METHODS ON TNO DATASETS

M*	En	SD	Q_{abf}	FMI_{dct}	FMI_w	$SSIM_a$	MS-SSIM
1	6.7226	74.1078	0.323	0.8846	0.185	0.5407	0.7552
2	6.7205	**79.1953**	0.3228	0.8642	0.1849	0.5412	0.7551
3	6.1826	48.1577	0.3681	0.9107	0.4168	**0.7779**	0.8747
4	6.6993	68.7931	0.4379	0.9047	0.4247	0.7288	0.9335
5	6.3628	67.5728	0.2183	0.8906	0.3708	0.6538	0.7318
6	6.6715	54.3575	0.44	0.9084	0.4276	0.7315	0.9289
7	6.5954	66.8757	0.5032	0.9009	0.4016	0.7318	0.9052
8	**6.8716**	77.2246	**0.5389**	0.9138	**0.4387**	0.7491	**0.9352**

*M. Method 1. GJSR [4], 2. JSRSD [7], 3. VggML [9], 4. DeepFuse [8], 5. FusionGAN [11], 6. DenseFuse [12], 7. IFCNN [10], 8. IR-MSDNet

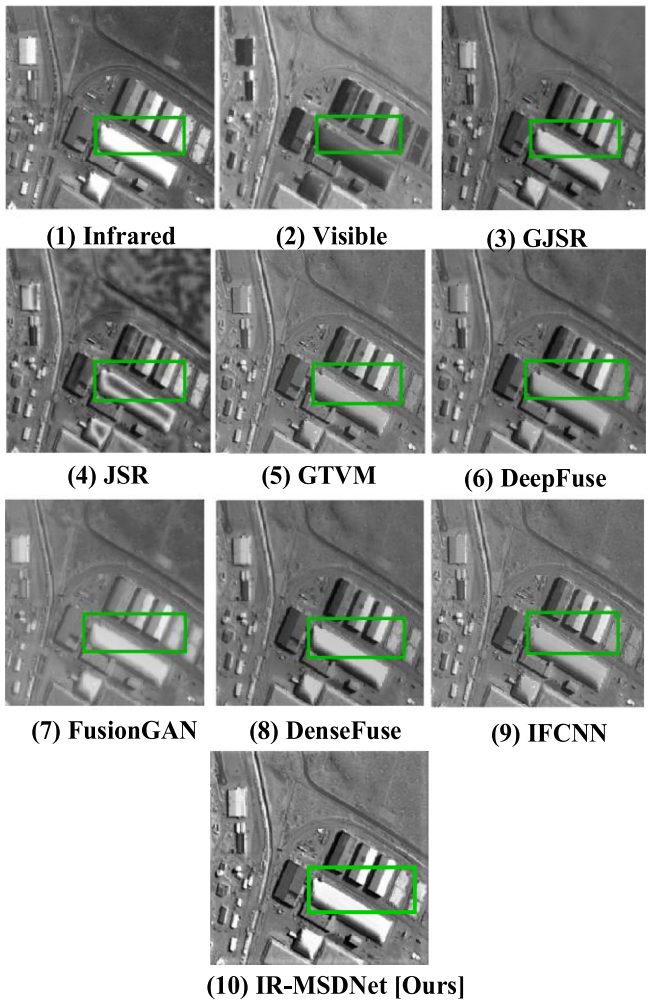


Fig. 12. Fusion results on Aerial Image dataset with different fusion methods. The different fusion results in green frame represent IR intensity regions.

TABLE III
OBJECTIVE EVALUATION INDICES FOR DIFFERENT FUSION METHODS ON AERIAL IMAGE DATASETS

M*	En	SD	Q_{abf}	FMI_{dct}	FMI_w	$SSIM_a$
1	6.8852	**59.4904**	0.0804	0.1048	0.1351	0.3217
2	6.6321	52.5197	0.1334	0.1147	0.1554	0.447
3	6.9121	49.9145	0.2158	0.2049	0.2828	0.5068
4	6.9336	51.9348	0.2204	0.2189	0.2912	0.5298
5	6.7526	48.0478	0.1578	**0.3656**	**0.3853**	0.4634
6	6.8494	54.9544	0.1804	0.125	0.2039	0.4884
7	6.898	49.4878	0.231	0.2124	0.2847	0.503
8	**6.9573**	57.1419	**0.2479**	0.1948	0.2821	**0.5334**

*M. Method 1. GJSR [4] 2. JSR [4], 3. GTVM [6], 4. DeepFuse [8], 5. FusionGAN [11], 6. DenseFuse [12], 7. IFCNN [10], 8. IR-MSDNet

TABLE IV
COMPARISON OF RESULTS OF OD BASED ON RPN FOR VISIBLE, INFRARED, AND FUSED IMAGES

Method *	Day	Night	All
1	31%	64%	47.70%
2	55%	**25%**	38%
3	29.10%	35.10%	32.10%
4	29.80%	35.90%	32.85%
5	29.30%	35.70%	32.50%
6	26.60%	33.80%	30.20%
7	27.40%	35.20%	31.30%
8	29.50%	34.90%	32.20%
9	26%	32%	29%
10	**24.30**%	30.60%	**27.45%**

*M. Method 1. Visible 2. IR 3. JSRSD [4], 4. VggML [9], 5. DeepFuse [8], 6. FusionGAN [11], 7. DenseFuse [12], 8. IFCNN [10], 9. Two streams (feature fusion from visible and IR image) [2] 10. IR-MSDNet

with the qualitative comparison, as it fully inherits IR target information and visible contents. In consideration of both the subjective and objective evaluations, it could be concluded that the proposed IR-MSDNet is superior in almost all aspects.

In addition, the time-complexity is another key metric to evaluate image fusion methods. Average time comparison is made on an Intel(R) Core (TM) i7-7700 U CPU (3.1 GHz), 16 GB RAM and GTX 1080ti GPU. Besides, the test time of traditional methods like GJSR, JSR, and JSRD on GPU version which show 49.2, 30, and 32.6 s, respectively. Nevertheless, deep methods like VggML, DeepFuse, FusionGAN, DenseFuse, IFCNN, and IR-MSDNet cost 2.19s, 0.91s, 0.41s, 3.64s, 0.31s, and 3.82s, respectively. Although IR-MSDNet is slightly slower than other deep methods but its performance is better than theirs.

F. Objective Evaluations on OD

Furthermore, to emphasis the effectiveness of IR-MSDNet, it is further applied to OD of remote sensing. RPN [2] is chosen as the basic state-of-the-art detector on KAIST [42] Multispectral (IR, Visible) benchmark datasets. Log average Miss Rate (MR) metric [2] as the standard measure for OD is used for quantitative evaluation. Log average MR is computed by averaging the miss rates over different false positive per-image (FPPI) points sampled within the evenly spaced in log-space. Since MR specifies the rate of undetected objects (e.g., persons, vehicles, etc.) as false negatives, so a lower value represents a robust OD. In Table IV, RPN detector is used to detect different objects from visible images, IR images, fused images generated by different fusion methods, i.e., JSRSD [4], VggML [9], DeepFuse [8], FusionGAN [11], DenseFuse [12], IFCNN [10], two streams (feature fusion from visible and IR image) [2], and our proposed IR-MSDNet, respectively. It can be seen from Table IV, the lowest average value of 27.45% is achieved by the proposed IR-MSDNet, which means that the proposed IR-MSDNet has greatly enhanced the details in the fused images and makes the RPN detector obtain the best OD results. So, the proposed IR-MSDNet not only competes for the image fusion task, but also can be applied to OD tasks in multispectral images (IR and Visible) as well.

V. CONCLUSION

In this article, a novel and effective deep architecture named IR-MSDNet is exclusively proposed to learn robust and discriminative salient representation to perform IR and visible image fusion. IR-MSDNet is based on specially designed IR features and a multiscale dense network with attention. It mainly contains an encoder, a multiscale decoder and a fused unit. The dense and multiscale features fused by l1-norm strategy, and further enhanced by attention module could capture more scale-related features. Edges features are concatenated with the features output by CWA module to form more detailed features for fusing. The final fused image is reconstructed from both the decoded multiscale features and IR features extracted traditional methods. Taking advantage of both deep learning and traditional methods, IR-MSDNet fully inherits IR target information and visible rich background details in the fused image. In experiments, both subjective and objective quality metrics are utilized to evaluate the proposed IR-MSDNet with other fusion methods. The results show that the proposed IR-MSDNet has achieved the state-of-the-art fusion performance. A further experiment on multispectral (RGB-Infrared) OD further validates the proposed method has greatly enhanced the details in the fused images, which are crucial for obtaining the best OD results.

ACKNOWLEDGMENT

The authors would like to thank the Editors and anonymous reviewers for their qualified suggestion to improve the quality of article.

REFERENCES

[1] J. Ma, Y. Ma, and C. Li, "Infrared and visible image fusion methods and applications: A survey," *Inf. Fusion*, vol. 45, pp. 153–178, 2019.
[2] K. Fritz, D. König, U. Klauck, and M. Teutsch, "Generalization ability of region proposal networks for multispectral person detection," in *Autom. Target Recognit. XXIXInt. Soc. Opt. Photon.*, vol. 10988, 2019, Art. no. 109880Y.
[3] A. Vishwakarma and M. K. Bhuyan, "Image fusion using adjustable non-subsampled shearlet transform," *IEEE Trans. Instrum. Meas.*, vol. 68, no. 9, pp. 3367–3378, Sep. 2019.
[4] Q. Zhang, Y. Fu, H. Li, and J. Zou, "Dictionary learning method for joint sparse representation-based image fusion," *Opt. Eng.*, vol. 52, no. 5, 2013, Art. no. 057006.
[5] Z. Chen, X. J. Wu, H. F. Yin, and J. Kittler, "Noise-robust dictionary learning with slack block-diagonal structure for face recognition," *Pattern Recognit.*, vol. 100, 2020. Art. no. 107118.
[6] H. Pan, Z. Jing, L. Qiao, and M. Li, "Visible and infrared image fusion using l (0)-generalized total variation model," *Sci. China Inf. Sci*, vol. 61, no. 4, 2018, Art. no. 049103.
[7] C. H. Liu, Y. Qi, and W. R. Ding, "Infrared and visible image fusion method based on saliency detection in sparse domain," *Infrared Phys. Technol.*, vol. 83, pp. 94–102, 2017.
[8] 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. Int. Conf. Comput. Vis.*, vol. 1, no. 2, pp. 3, 2017.
[9] H. Li, X. J. Wu, and J. Kittler, "Infrared and visible image fusion using a deep learning framework," in *Proc. IEEE 24th Int. Conf. Pattern Recognit.*, 2018, pp. 2705–2710.

[10] 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, 2020.
[11] 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, 2019.
[12] 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 2018.
[13] X. Bai, Y. Zhang, F. Zhou, and B. Xue, "Quadtree-based multi-focus image fusion using a weighted focus-measure," *Inf. Fusion*, vol. 22, pp. 105–118, 2015.
[14] L. Zhang, "In situ image segmentation using the convexity of illumination distribution of the light sources," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 10, pp. 1786–1799, Oct. 2008.
[15] P. Burt and E. Adelson, "The Laplacian pyramid as a compact image code," *IEEE Trans. Commun.*, vol. 31, no. 4, pp. 532–540, Apr. 1983.
[16] H. Li, B. S. Manjunath, and S. K. Mitra, "Multisensor image fusion using the wavelet transform," *Graphical Models Image Process.*, vol. 57, no. 3, pp. 235–245, 1995.
[17] J. J. Lewis, R. J. O'Callaghan, S. G. Nikolov, D. R. Bull, and N. Canagarajah, "Pixel-and region-based image fusion with complex wavelets," *Inf. fusion*, vol. 8, no. 2, pp. 119–130, 2007.
[18] F. Nencini, A. Garzelli, S. Baronti, and L. Alparone, "Remote sensing image fusion using the curvelet transform," *Inf. Fusion*, vol. 8, no. 2, pp. 143–156, 2007.
[19] Y. Liu, S. Liu, and Z. Wang, "Multi-focus image fusion with dense SIFT," *Inf. Fusion*, vol. 23, pp. 139–155, 2015.
[20] S. Li, X. Kang, J. Hu, and B. Yang, "Image matting for fusion of multi-focus images in dynamic scenes," *Inf. Fusion*, vol. 14, no. 2, pp. 147–162, 2013.
[21] 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.
[22] V. Aslantas and R. Kurban, "Fusion of multi-focus images using differential evolution algorithm," *Expert Syst. Appl.*, vol. 37, no. 12, pp. 8861–8870, 2010.
[23] I. De and B. Chanda, "Multi-focus image fusion using a morphology-based focus measure in a quad-tree structure," *Inf. Fusion*, vol. 14, no. 2, pp. 136–146, 2013.
[24] M. Li, W. Cai, and Z. Tan, "A region-based multi-sensor image fusion scheme using pulse-coupled neural network," *Pattern Recognit. Lett.*, vol. 27, no. 16, pp. 1948–1956, 2006.
[25] S. Li and B. Yang, "Multifocus image fusion using region segmentation and spatial frequency," *Image Vis. Comput.*, vol. 26, no. 7, pp. 971–979, 2008.
[26] A. Raza, H. Huo, and T. Fang, "PFAF-Net: Pyramid feature network for multimodal fusion," *IEEE Sensors Lett.*, vol. 4, no. 12, Dec. 2020, Art. no. 5501704.
[27] Y. Liu, X. Chen, H. Peng, and Z. Wang, "Multi-focus image fusion with a deep convolutional neural network," *Inf. Fusion*, vol. 36, pp. 191–207, 2017.
[28] A. Raza, H. Huo, S. Sirajuddin, and T. Fang, "Diverse capsules network combining multiconvolutional layers for remote sensing image scene classification," in *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 5297–5313, Sep. 2020. doi: 10.1109/JSTARS.2020.3021045.
[29] H. Ji, Z. Gao, T. Mei, and Y. Li, "Improved faster R-CNN with multiscale feature fusion and homography augmentation for vehicle detection in remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 11, pp. 1761–1765, Nov. 2019.
[30] Y. Liu, X. Chen, R. K. Ward, and Z. J. Wang, "Image fusion with convolutional sparse representation," *IEEE Signal Process. Lett.*, vol. 23, no. 12, pp. 1882–1886, Dec. 2016.
[31] 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, 2016.
[32] A. Toet, "TNO Image fusion dataset," *Figshare. Data*, 2014. [Online]. Available: https://figshare.com /articles/TN_Image_Fusion_Dataset/1008029
[33] N. Cvejic, T. Seppanen, and S. J. Godsill, "A nonreference image fusion metric based on the regional importance measure," *IEEE J. Sel. Top. Signal Process.*, vol. 3, no. 2, pp. 212–221, Apr. 2009.
[34] J. W. Roberts, J. A. van Aardt, and F. B. Ahmed, "Assessment of image fusion procedures using entropy, image quality, and multispectral classification," *J. Appl. Remote Sens.*, vol. 2, no. 1, 2008, Art. no. 023522.
[35] C. A. Xydeas and V. Petrovic, "Objective image fusion performance measure," *Electron. Lett.*, vol. 36, no. 4, pp. 308–309, 2000.
[36] M. Haghighat, and M. A., and Razian, "Fast-FMI: Non-reference image fusion metric," in *Proc. IEEE 8th Int. Conf. Application Inf. Commun. Technol.*, 2014, pp. 1–3.
[37] H. Li, X. J. Wu, and J. Kittler, "Infrared and visible image fusion using a deep learning framework," in *Proc. 24th Int. Conf. Pattern Recognit.*, 2018, pp. 2705–2710.
[38] 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.
[39] Y. J. Rao, "In-fibre Bragg grating sensors," *Meas. Sci. Technol.*, vol. 8, no. 4, pp. 355, 1997.
[40] T. Y. Lin, "Microsoft coco: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.
[41] Y. Zhang, L. Zhang, X. Bai, and L. Zhang, "Infrared and visual image fusion through infrared feature extraction and visual information preservation," *Infrared Phys. Technol.*, vol. 83, pp. 227–237, 2017.
[42] S. Hwang, J. Park, N. Kim, Y. Choi, and I. So Kweon, "Multispectral pedestrian detection: Benchmark dataset and baseline," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1037–1045.

Asif Raza received the B.S. degree in computer system engineering, from Bahauddin Zakariya University, Multan, Pakistan, in 2012 and the M.S. degree in control engineering From Beihang University of Aeronautics and Aerospace, Beijing, China, in 2017, and he is a currently working toward the Ph.D. degree with Remote Sensing Group, State Key Laboratory of Controls Engineering, the School of Information Engineering and Electrical Engineering (SEIEE), Shanghai Jiaotong University, Shanghai, China, with major in control engineering.

His research concentrates on the area of information fusion, deep learning, multimodal image classification with applications to remote sensing

Jingdong Liu received the B.E. degree in electrical engineering and automation from Donghua University, Shanghai, China, in 2019. He is currently working toward the M.S. degree in control science and engineering with Shanghai Jiaotong University, Shanghai, China.

His research concentrates on the area of research direction: knowledge graph, deep learning, and computer vision with applications to remote sensing.

Yifan Liu received the B.S. degree in electronic information engineering from Department of electronic information engineering, School of Information Science and Technology, Hainan University, Haikou, China, in 2018. He is currently working toward the M.D. degree with the Department of Automation, School of Electronic Information and Electrical Engineering, Shanghai Jiaotong University, Shanghai, China.

His research interests include semantic segmentation and change detection of remote sensing images.

Jian Liu received the bachelor's degree in automation from Harbin Institute of Technology, Harbin, China, in 2015, the master's degree in control science and engineering from Harbin Institute of Technology, in 2017. He is currently working toward the Ph.D. degree in control science and engineering at Shanghai Jiaotong University, Shanghai, China.

He majors in control science and engineering. His research concentrates on the area of control theory and complex network with applications to neural network.

Zeng Li received the B.E. degree in electrical engineering from the Electrical Engineering and Automation Department, Zhejiang University, Zhejiang, China, in 2019. He is currently working toward the master's degree with the Department of Automation, School of Electronic Information and Electrical Engineering, Shanghai Jiaotong University, Shanghai, China.

His research interests include remote sensing image process and knowledge graph.

Xi Chen received the B.Eng. degree in automation from Chongqing University, Chongqing, China, in 2002, the M.Eng. degree in computer aided design from Shantou University, Shantou, China, in 2006, and the Ph.D. degree in automation from Shanghai Jiao Tong University, Shanghai, China, in 2011.

He is currently a Researcher with the School of Geographical Sciences, East China Normal University, Shanghai, China. His research interests include pattern recognition and image processing, as well as their applications to remote sensing images**.**

Hong Huo received the B.S. and M.S. degrees in computer application from the Jilin University of Technology, Jilin, China, in 1995 and1998, respectively, and the Ph.D. degree in pattern recognition and intelligent systems from Shanghai Jiao Tong University, Shanghai, China.

Since 2000, she has been an Associate Professor with the Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University. Her research interests include remote sensing and image classification, machine learning, and datamining with applications to remote sensing.

Tao Fang received the B.S. and M.S. degrees in geology and survey from the Xi'an University of Science and Technology, Xi'an, China, in 1988 and1991, respectively, and the Ph.D. degree in remote sensing and geographical information system from the China University of Mining and Technology, Xuzhou, China, in 1996.

He was a Post-Doctoral Research Fellow with the Remote Sensing and Geographical Information System, State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, Wuhan, China, from 1996 to 1998. He was an Assistant Professor with the Department of Electronic Engineering and a Post-Doctoral Research Fellow of electronics with the HDTV Laboratory, University of Electronic Science and Technology of China, Chengdu, China, from 1999 to2000. He is currently a Professor with the Institute of Image Processing and Pattern Recognition of Shanghai Jiao Tong University, Shanghai, China. His current research interests include computer vision, statistical and structural pattern recognition, machine learning, and data mining with applications to remote sensing classification, object recognition, change detection, and land cover and land use.