Information Fusion 54 (2020) 99–118

Contents lists available at ScienceDirect

Information Fusion

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

IFCNN: A general image fusion framework based on convolutional neural network

Yu Zhang [a,*], Yu Liu [b], Peng Sun [c], Han Yan [a], Xiaolin Zhao [d], Li Zhang [a,*]

[a] Department of Electronic Engineering, Tsinghua University, Beijing 100084, China
[b] Department of Biomedical Engineering, Hefei University of Technology, Hefei 230009, China
[c] Beijing Aerospace Automatic Control Institute, Beijing 100854, China
[d] School of Aeronautics and Astronautics Engineering, Airforce Engineering University, Xi'an 710038, China

ARTICLE INFO

Keywords:
General image fusion framework
Convolutional neural network
Large-scale multi-focus image dataset
Better generalization ability

ABSTRACT

In this paper, we propose a general image fusion framework based on the convolutional neural network, named as IFCNN. Inspired by the transform-domain image fusion algorithms, we firstly utilize two convolutional layers to extract the salient image features from multiple input images. Afterwards, the convolutional features of multiple input images are fused by an appropriate fusion rule (elementwise-max, elementwise-min or elementwise-mean), which is selected according to the type of input images. Finally, the fused features are reconstructed by two convolutional layers to produce the informative fusion image. The proposed model is fully convolutional, so it could be trained in the end-to-end manner without any post-processing procedures. In order to fully train the model, we have generated a large-scale multi-focus image dataset based on the large-scale RGB-D dataset (i.e., NYU-D2), which owns ground-truth fusion images and contains more diverse and larger images than the existing datasets for image fusion. Without finetuning on other types of image datasets, the experimental results show that the proposed model demonstrates better generalization ability than the existing image fusion models for fusing various types of images, such as multi-focus, infrared-visual, multi-modal medical and multi-exposure images. Moreover, the results also verify that our model has achieved comparable or even better results compared to the state-of-the-art image fusion algorithms on four types of image datasets.

1. Introduction

The target of image fusion is to integrate the salient features of multiple input images into one comprehensive image [1–4]. Nowadays, image fusion has become more closely related to our daily lives and played more important role in the industrial field and military field. For instance, the mobile phones are often integrated with HDR (High Dynamic Range) [5–7] or refocusing algorithms [8–10] to enable us to capture satisfactory and informative pictures, where HDR and refocusing are essentially image fusion algorithms. In hospitals, the surgeons diagnose diseases of patients by inspecting multiple modalities of medical images (such as computed tomography (CT) image and magnetic resonance (MR) image), and especially they determine the precise boundaries of bone tumors according to the fused CT and MR images [11,12]. In the military or civil surveillance systems, fusion of the infrared and visual images could bring the observers great convenience to fully learn about the supervised environment [2,13–16].

In general, the image fusion algorithms could be divided in two categories [2,17], i.e., spatial-domain algorithms and transform-domain algorithms. The spatial-domain image fusion algorithms [9,17–19] firstly parse the input images into small blocks or regions according to some criterion, then measure the saliency of the corresponding regions, and finally combine the most salient regions to form the fusion image. This kind of algorithms are mainly suitable for fusing images of the same modality (such as multi-focus images), and probably suffer from block or region artifacts around the stitching positions. On the other hand, the transform-domain image fusion algorithms [3,20–25] firstly transform the source images into some feature domain through multi-scale geometric decomposition (such as multi-scale pyramids and multi-scale morphological operators), and then perform weighted fusion on the features of multiple input images. Afterwards, the fused features are inversely transformed to produce the fusion image. Since in the feature domain, even the images of different modalities would share the similar property, thus the transform-domain image fusion algorithms could be generally

* Corresponding authors.
E-mail addresses: uzeful@163.com (Y. Zhang), chinazhangli@mail.tsinghua.edu.cn (L. Zhang).
URL: https://uzeful.github.io (Y. Zhang)

https://doi.org/10.1016/j.inffus.2019.07.011
Received 6 August 2018; Received in revised form 3 March 2019; Accepted 24 July 2019
Available online 24 July 2019

used to fuse more types of images, such as infrared-visual images and CT-MR images. Moreover, this kind of algorithms have achieved great success in the past two decades. However, the fusion strategy or weight coefficients of the transform-domain algorithms are often hard to optimize for the fusion purpose, and thus probably could not achieve the optimal fusion results and suffer from low-contrast effect or blurring effect.

In recent years, the machine learning algorithms have been widely used in completing different kinds of image fusion tasks, and achieved great success in the image fusion field. At beginning, Yang et al. [26] employed the sparse representation technique to fuse multi-focus images, in which the image patches were represented with an overcomplete dictionary and corresponding sparse coefficients, and then the input images were fused through fusing the sparse coefficients of each pair or set of image patches. In the following, a sequence of sparse representation based algorithms [27–29] appeared to further improve the algorithms' performance and extend to fuse more types of images (such as multi-modal medical images and infrared-visual images).

More recently, deep learning techniques, especially convolutional neural network (CNN), have brought new evolution into the field of image fusion [30]. Firstly, Liu et al. [31] introduced CNN to fuse multi-focus images. They formulated multi-focus image fusion as a classification task and used CNN to predict the focus map, as each pair of image patches could be classified into two categories: (1) first patch was focused and second blurred and (2) first patch was blurred and second focused. In [32], Tang et al. proposed a CNN model to learn the effective focus-measure (i.e., metric for quantifying the sharpness degree of an image or image patch) and then compared the focus-measures of local image patch pairs of input images to determine the focus map. Afterwards, the above two algorithms both post-processed the focus maps and reconstructed the fusion images according to the refined focus maps. In [33], Song et al. applied two CNNs to fuse the spatiotemporal satellite images, i.e., large-resolution MODIS and low-resolution landsat images. Specifically, they respectively used two CNNs to perform super-resolution on the low-resolution landsat images and extract image features, and then adopted high-pass modulation and weighting strategy to reconstruct the fusion image from the extracted features similar to the transform-domain image fusion algorithms [15]. However, the above three algorithms were not designed in the end-to-end manner and all required post-processing procedures to produce fusion images, thus their models might have not been fully optimized for the image fusion task. In [34], Prabhakar et al. proposed an end-to-end multi-exposure fusion model. Specifically, they firstly used CNN to fuse the intensity channel (Y channel in the YCbCr color space) of the multiple input images, then leveraged the contrast-enhancement method to adjust the fused intensity channel, and afterwards employed the weighted-average strategy to respectively fuse the Cb and Cr channels. Finally, the fused channels (Y, Cb and Cr) were stacked together to produce the fusion image. Their model could be trained end-to-end and could be applied to fuse other types of images, such as multi-focus images. However, their results on the multi-focus image dataset appear to suffer from low-contrast effect.

Even though the CNN models have achieved some success in the field of image fusion, the current models lack the generalization ability and could only perform well on one specific type of images. This problem will however bring us great difficulty in developing the CNN based algorithms for fusing images without ground-truth images (such as infrared-visual images and CT-MR images). Moreover, most of the proposed CNN models are not designed in the end-to-end manner, and thus require additional procedures to complete the image fusion task. Overall, the CNN based models have not been fully exploited for the image fusion task, thus there is still much space to improve the architectures of the CNN based image fusion models, so as to increase their performance and generalization ability.

Through comparing the transform-domain image fusion algorithms and CNN based image generation models, we find there are several similar characteristics between these two kinds of algorithms. Firstly, the transform-domain algorithms usually extract the image features using several filters (such as Gaussian filters or morphological filters) at the beginning, and the CNN models also extract extensive features using large number of convolutional filters. Secondly, the transform-domain fusion algorithms usually fuse the features through the weighted-average strategy, and the CNN models also utilize the weighted-average strategy (weighted sum of the convolutional features) to generate the target image. Compared to the transform-domain image fusion algorithms, the CNN models have three advantages: (1) the number of convolutional filters is usually much greater than that of the filters in the conventional transform-domain algorithms, and thus the convolutional filters could extract more informative image features; (2) the proper parameters of convolutional filters can be learned to fit the image fusion task; (3) the parameters of the CNN models can be jointly optimized through training them in the end-to-end manner.

Inspired by the transform-domain algorithms, we proposed a general image fusion framework based on the convolutional neural network, architecture of which in the training phase has been shown in Fig. 1. Firstly, we used two convolutional layers to extract the informative low-level features from multiple input images. Secondly, the extracted convolutional features of each input image were elementwisely fused by the appropriate fusion strategy, such as elementwise-maximum and elementwise-mean. Finally, the integrated features were reconstructed by two convolutional layers to produce the fusion image. As the proposed model is fully convolutional, thus it could be trained in the end-to-end manner with any post-processing procedures, which is one superior advantage compared to most of the existing image fusion models. Furthermore, in order to fully train the proposed model, we have created a large-scale multi-focus image dataset by blurring portions of images from our prebuilt NYU-D2 dataset [35] according to the random depth ranges, which is more reasonable than blurring whole or certain portions of image patches in [31,32]. Compared to the previous datasets for training image fusion models, the resolution (224×224) of our dataset is much larger than those (16×16, 32×32 and 64×64) of datasets in [31,32,34], and the source RGB images in NYU-D2 dataset can be taken as the ground-truth fusion images of our dataset which is much better than no ground-truth fusion images in datasets of [31,32,34]. Thanks to the above merits, our high-resolution large-scale multi-focus image dataset can be used to finely train the image fusion models. During the training phase, we firstly adopted the mean square error (MSE) of the fusion image and the ground-truth fusion image to pretrain the model's parameters, and then equipped the perceptual loss (mean square error of the deep convolutional features of the predicted fusion image and ground-truth fusion image) with MSE to jointly optimize the model's parameters. The appropriate model archieteture, well generated multi-focus image dataset and superior loss function, together guarantee our algorithm to achieve good performance on the image fusion task. Moreover, the extensive experimental results show that the proposed model can well fuse multiple types of images without any finetuning procedure, and meanwhile achieve comparable or even better performance than the state-of-the-art image fusion algorithms.

To sum up, the contributions of this paper are fourfold:

- In this paper, the image fusion task is formulated as a fully convolutional neural network, thus the proposed image fusion model can be trained in the end-to-end fashion so that all parameters of the proposed model could be jointly optimized for the image fusion task without any post-processing procedures. Based on the proposed CNN based image fusion framework, the researchers can conveniently develop their own image fusion models for fusing various types of images.
- To fully train the model's parameters, we have generated a large-scale multi-focus image dataset. Instead of creating low-resolution pairs of fully focused and fully blurred image patches, we have generated high-resolution pairs of partially-focused images by blurring image portions of random depth ranges from the RGB and depth images in our prebuilt RGB-D dataset.

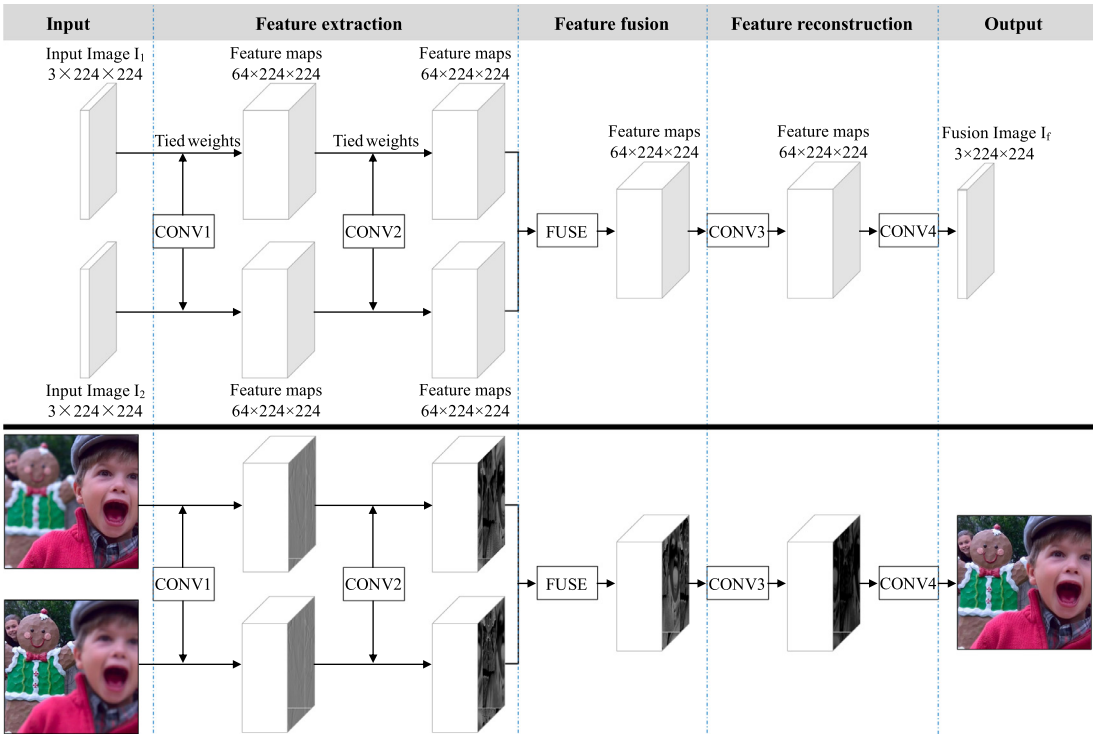

Fig. 1. The proposed general image fusion framework based on convolutional neural network. The above part illustrates the architecture of our image fusion model, and the below part shows a demonstration example for fusing multi-focus images. Please note that the spatial sizes marked in the figure just indicate the ones used in our training phase, and the inputs can be extended to more than two images.

Compared to the existing multi-focus image generation methods, our method is more close to the imaging principle of optical lens, therefore the multi-focus images generated by our method are more natural and diverse than the pairs of fully focused and fully blurred image patches. Moreover, the RGB source images can be naturally taken as the ground-truth fusion images of the generated multi-focus image dataset, which is of great importance for supervising the image fusion models (i.e., regression models) to transfer the salient details from multiple inputs into one fusion image. Owning to these merits, our multi-focus image dataset can be used to fully and finely train the image fusion models.

- Owing to the similar structure with the transform-domain image fusion algorithms, our model owns better generalization ability than the existing CNN models for fusing various types of images. Although the proposed model has been trained only on the multi-focus image dataset, it has well learned the ability to fuse the convolutional features of multiple images of the same type or even different types. Therefore, our model could be directly applied to fuse other types of images (such as infrared-visual, CT-MR and multi-exposure images) without any finetuning procedures, and still achieve state-of-the-art results.
- To the best of our knowledge, it is the first time to introduce perceptual loss in training the CNN based image fusion model. The chief reason is that computation of perceptual loss requires ground-truth fusion images, which however have not been generated in the existing image datasets for training image fusion models. Through introducing the perceptual loss, the trained image fusion model could produce fusion images with more textural information than those without incorporating perceptual loss.

In our opinion, there are two major novelties in this paper. Firstly, our model's characteristics of fully convolutional neural network and good generalization ability together compose the first major novelty of this paper. Secondly, our high-resolution large-scale multi-focus image dataset (with ground-truth fusion images) is another major novelty of this paper. The reasons are as follows: (1) To the best of our knowledge, there is still no fully convolutional neural networks based image fusion model that can achieve state-of-the-art fusion images on multiple types of images without any finetuning procedure as our model does, and (2) in the field of deep learning, the quality of training dataset often directly determines the upper limit of the model's performance, thus our high-resolution large-scale multi-focus image dataset (with ground-truth fusion images) is more superior for fully training the image fusion models than the existing low-resolution large-scale image datasets (without ground-truth fusion images). Therefore, either of the two major novelties can make the proposed image fusion model stand out from the existing CNN based image fusion models. The rest of this paper is organized as follows. In Section 2, the proposed method including our image fusion model and training dataset is introduced in detail. The extensive experimental results and discussions are described in Section 3. Finally, the conclusions are drawn in Section 4.

2. Proposed method

2.1. Overview

In the field of computer vision, the convolutional layer plays the role of feature extraction, and usually could extract more extensive and informative features than the traditional handcrafted feature extractors [35,36]. In addition, the convolutional layer also plays the role of weighted average for producing the output image. These characteristics of the convolutional layer are quite similar to the transform-domain image fusion algorithms, thus the convolutional neural network has great potential to achieve success in the field of image fusion.

Inspired by the framework of the transform-domain image fusion algorithms, we have designed a general image fusion framework based on the convolutional neural network, which has been abbreviated as IFCNN hereinafter. IFCNN consists of three modules: feature extraction module, feature fusion module and image reconstruction module, as

shown in Fig. 1. Firstly, we adopt two convolutional layers to extract the informative image features. Secondly, the convolutional features of multiple input images are fused via the feature fusion module. Finally, the fused features are reconstructed by two convolutional layers to produce the fusion image.

In order to fully train our image fusion model, we have online generated a large-scale multi-focus image dataset based on our prebuilt NYU-D2 dataset [35,37], which consists of about 100,000 pairs of RGB and depth images. The focused and blurred portions of our multi-focus image pairs are determined according to their depth ranges, thus the generation of our multi-focus image dataset is intuitive and reasonable. In addition, the source RGB images can be directly taken as the ground-truth fusion images, which are important for fully training the regression models for image fusion. Moreover, in order to more effectively train the proposed model, we have introduced the perceptual loss to regularize the proposed model to generate fusion images with more similarity to the ground-truth fusion image. The details about the proposed image fusion model are introduced in the following subsections.

2.2. Image fusion model

In order to conveniently describe the proposed modules, we assume that there are N ($N \geq 2$) input images to fuse, denoted by I_k ($1 \leq k \leq N$). Then, the three modules of the proposed image fusion model can be respectively detailed as follows.

2.2.1. Feature extraction module

Firstly, we adopt two convolutional layers to extract the extensive low-level features from the input images. As feature extraction is the crucial procedure in the transform-domain image fusion algorithms, and is usually conducted by processing images with multi-scale DOG (Difference of Gaussian) [15], multi-scale morphological filters [3] and so on. As for CNN, training the regression model (image-to-image) is usually hard and not stable from the random initialized convolutional kernels, and thus a practical way is to transfer the parameters of a well-trained classification model to the regression model [35]. Thereby, in this paper, we adopt the first convolutional layer of the superior ResNet101 pretrained on ImageNet as our first convolutional layer (CONV1). As is known, CONV1 contains 64 convolutional kernels of size 7×7, which are sufficient enough to extract extensive image features, and CONV1 has been trained on the largest natural image dataset (i.e., ImageNet). Therefore, CONV1 can be used to extract the effective image features, and thus we have fixed the parameters of CONV1 during training the proposed model. However, the extracted features by CONV1 are originally used for the classification task, thus directly feeding them into the feature fusion module might be not appropriate for the image fusion task. Hence, we add the second convolutional layer (CONV2) to tune the convolutional features of CONV1 to suit for feature fusion.

2.2.2. Feature fusion module

In this paper, our target is to propose a general image fusion model based on CNN, which can fuse various types of input images and also can fuse various number of input images. In general, there are usually two methods to fuse the convolutional features of multiple inputs: (1) the convolutional features of multiple inputs are firstly concatenated along the channel dimension, and then the concatenated features are fused by the following convolutional layer, (2) the convolutional features of multiple inputs are directly fused by the elementwise fusion rules (such as elementwise-maximum, elementwise-sum and elementwise-mean). As the concatenation fusion method requires parameter number of the feature fusion module varies with the input number, thus the models with this fusion method can only fuse a specific number of images once the model architecture is fixed. While the feature fusion module with elementwise fusion method does not contain any parameter and can fuse various number of input images, and has ever been introduced in the image fusion models [34].

Therefore, in our feature fusion module, the elementwise fusion rules have been utilized to fuse the convolutional features of multiple inputs, which can be mathematically expressed as Eq. (1). As described above, there are three commonly used elementwise fusion rules, i.e., elementwise-maximum, elementwise-sum and elementwise-mean. In practice, the fusion rule should be selected according to the characteristics of the image dataset. For instance, the sharp features (maximum values) indicate the salient objects of the supervised scene, thus the elementwise-maximum fusion rule has been often used in the transform-domain image fusion algorithms to fuse the multi-focus images, infrared and visual images, and medical images. However, multi-exposure image fusion is to integrate the visually-pleasant middle-exposure portions of each input image, where most probably correspond to the mean features of multiple inputs. Thus, at this time, elementwise-mean fusion rule might be more suitable to fuse the multi-exposure images than elementwise-maximum fusion rule.

$$\widehat{f}^j(x,y) = \underset{i}{fuse}\left(f^j_{i,C_2}(x,y)\right),\ 1 \leq i \leq N, \tag{1}$$

where f^j_{i,C_2} denotes the jth feature map of the ith input image extracted by CONV2, $\widehat{f}^j$ denotes the jth channel of fused feature maps by our feature fusion module, and *fuse* denotes the elementwise fusion rule (such as elementwise-maximum, elementwise-sum and elementwise-mean).

Hence, in this paper, we have used the elementwise-mean fusion rule to fuse the multi-exposure images, and used the elementwise-maximum fusion rule to fuse other types of images, including multi-focus images, infrared and visual images, and multi-modal medical images.

2.2.3. Image reconstruction module

Because our feature extraction module only contains two convolutional layers, thus abstraction level of the extracted convolutional features is not high. Therefore, in the final stage of our proposed model, we also adopt two convolutional layers (CONV3 and CONV4) to reconstruct the fusion image from the fused convolutional features $\widehat{f}$.

2.2.4. Model details

As down-sampling feature maps will inevitably lose the source information of input images, which might affect the fusion image's quality. Therefore, in our image fusion model, we have not down-sampled the feature maps in any layer, and thus the size of feature maps has been kept same with that of input images throughout the model. To satisfy the above condition while generating good fusion images, the parameters of our image fusion model are set as follows.

Firstly, as the kernel size of CONV1 is 7×7, thus the stride and padding parameters of CONV1 are respectively set as 1 and 3. Secondly, because CONV2 is used to tune the convolutional features of CONV1, thus number of feature maps of CONV2 should be same with CONV1. Therefore, the kernel number and kernel size of CONV2 are respectively set as 64 and 3×3, and both stride and padding parameters of CONV2 are set as 1. Thirdly, CONV3 also plays the role of tuning the fused convolutional features after the feature fusion module, thus parameter settings are same with CONV2, i.e., kernel number and kernel size of CONV3 are respectively set as 64 and 3×3, and both stride and padding parameters of CONV2 equal to 1. Finally, CONV4 plays the role of reconstructing feature maps into the 3-channel output, which is often implemented by the elementwise weighted average [38]. Thereby, the kernel number and kernel size of CONV4 are respectively set as 3 and 1×1, and both stride and padding parameters of CONV4 are set as 0.

Moreover, in order to overcome the overfitting problem and boost the training process, both two middle convolutional layers (CONV2 and CONV3) have been equipped with ReLU activation layer [39] and batch-normalization layer [40]. Because CONV1 has been well trained on ImageNet and thus does not need retraining, and the last convolutional layer (CONV4) is usually not equipped with activation layer or batch-normalization layer, thus we have not appended RELU layer and batch-normalization layer after CONV1 and CONV4.

Overall, our model is designed to fuse multiple RGB images and produce one RGB fusion image. Nevertheless, the proposed model can be conveniently extended to fuse the single-channel images by stacking three same channels. Specifically, the RGB multi-focus images can be directly fused by the proposed model, the infrared and visual images or multi-modal medical images should be firstly extended to three channels and then could be fused by our model. Finally, fusion of the RGB multi-exposure images is performed referred to [34]: (1) converting the RGB input images to YCbCr color space, (2) for each input image, separating the YCbCr channels and stacking three Y channels as the input of our image fusion model, (3) using our model to fuse the three-channel Y images of all source images and convert the three-channel output to the single-channel Y' according to Eq. (2), (4) fusing Cb and Cr channels of all source images by the same weighted strategy with Prabhakar et al., (5) stacking Y', fused Cb and fused Cr together and converting it back to RGB color space to produce the fusion image. Note that the input and output of Prabhakar et al.'s method and those of ours are a little different, i.e., both input and output of their model are single-channel, while both input and output of our model are three-channel. Therefore, during fusing the multi-exposure images, Y channel of each source image is extended to three Y channels before inputting into our image fusion model.

$$Y' = 0.299 \times R + 0.587 \times G + 0.114 \times B, \tag{2}$$

where R, G and B respectively correspond to the three channels of the produced fusion image.

In the end of this subsection, we have demonstrated the performance of feature extraction module and feature fusion module on one pair of multi-focus images, as shown in Fig. 2. We can see from Figs. 2(d) and (e) that the feature extraction module has extracted extensive feature maps from Figs. 2(a) and (b), ranging from informative edge details to flat basic elements. Figs. 2(d)–(f) show that the sharp features extracted from Figs. 2(a) and (b) have been successfully integrated into Fig. 2(f) by the feature fusion module. For the clear observation, a set of feature maps bounded by red boxes in Figs. 2(d)–(f) have been separately shown in Figs. 2(g)–(i) and projected to HSV color space for better visualization effect. We can find that the edge details of Fig. 2(g) are concentrated on the near focused boy, the edge details of Fig. 2(h) are concentrated on the far focused background, and the edge details in both Figs. 2(g) and (h) have been successfully integrated by our feature fusion module into the feature map in Fig. 2(i). Finally, Fig. 2(c) demonstrates that the fusion image reconstructed from Fig. 2(i) by our feature reconstruction module has well combined the clear portions of Figs. 2(a) and (b). The results in Fig. 2 can validate the effectiveness of our feature extraction module and feature fusion module.

2.3. Training dataset

As is known, the CNN models are data-driven and a well-generated large-scale image dataset is the foundation of this kind of algorithms. In [31], Liu et al. assumed that each pair of multi-focus image patches mainly had two classes: (1) the first patch was focused and the second one blurred, and (2) the first patch was blurred and the second one focused. Therefore, they generated a large image dataset consisting of 2,000,000 pairs of image patches of size 16×16, by randomly cropping the focused patches from the ImageNet dataset and blurred ones by blurring the focused patches with random scale of Gaussian kernel. Tang et al. proposed p-CNN to learn effective focus-measure, where their p-CNN was formulated as a simple image classification task (three categories: each image patch is focused, blurred, or unknown). Based on this assumption, they generated a large-scale image dataset containing about 1,450,000 image patches of size 32×32. In this image dataset, there are 650,000 focused patches, 700,000 blurred patches and 100,000 unknown type of patches, which were rendered with 12 handcrafted blurring masks. In [34], since there is also no large-scale multi-exposure image dataset, Prabhakar et al. generated their multi-exposure image dataset by randomly cropping patches of size 64×64 from a small-scale of multi-exposure image set. Furthermore, due to lacking ground-truth fusion images, they trained the multi-exposure image fusion model through an unsupervised method, i.e., using the structural similarity loss (SSIM [41]).

As detailed above, the current datasets for training image fusion models mainly consist of small image patches (16×16, 32×32 and 64×64). Among the three existing datasets, the resolution of Liu et al.'s multi-focus image dataset is lowest, and only contains one type of image patch pairs (i.e., one patch is focused and another one blurred), which is not appropriate for fully training the end-to-end image fusion models. While the second multi-focus image dataset only consists of single blurred images and is designed to train the model's ability for identifying the focus types (i.e., focused, defocused, and unknown) of the image patches, which is not suitable for training the end-to-end image fusion network either. Finally, Prabhakar et al. took the randomly cropped patches of size 64×64 from a small-scale sets of multi-exposure images as their training dataset, and trained their image model on the multi-exposure image dataset with the unsupervised SSIM loss. However, low-resolution of image dataset and no ground-truth fusion images will absolutely limit performance of the trained image fusion models.

As reported in the previous literatures, the multi-focus image dataset could be more easily generated compared to other types of image dataset, and more importantly, the ground-truth fusion images of the multi-focus images could be obtained simultaneously while generating the dataset. Due to the characteristics of optical lens, focused and blurred portions of the naturally captured images are generally related to the scene depth. Thus, a reasonable way for generating the multi-focus image dataset is to create the partially-focused image pairs from the RGB-D image sets. As is known, NYU-D2 is a famous indoor dataset for depth estimation, which consists of thousands of RGB and depth image pairs. In our previous work [35], we have built a large-scale NYU-D2 training set by uniformly sampling the training sequences of the NYU-D2 raw dataset, and the training dataset consists of about 100,000 pairs of RGB and depth images all of which have been resized to 422×321. Therefore, in this paper, we have online generated a large-scale multi-focus image dataset based on our previously built NYU-D2 training dataset in a more reasonable and intuitive way.

To be specific, during training the model, our online multi-focus image dataset is simultaneously generated from pairs of RGB and depth images of NYU-D2 dataset as the following procedures:

(1) a complete blurry image I_b is generated by randomly blurring the source RGB image I_s with Gaussian filter, which can be expressed as

$$I_b = G * I_s, \tag{3}$$

where $*$ denotes the convolution operation, and G denotes the Gaussian kernel, which is generated with random kernel radius kr ranging from 1 pixel to 15 pixels according to Eq. (4).

$$G(x, y) = \frac{1}{\left(\sqrt{2\pi}\sigma\right)} e^{\frac{x^2+y^2}{2\sigma^2}}, \tag{4}$$

where σ denotes the standard deviation of the Gaussian filter and can be calculated as $\sigma = 0.3 \times (kr - 1) + 0.8$.[1]

(2) the focus map I_m for determining the focused and blurred portions of the multi-focus images is generated according to the random depth range. Specifically, the near portions and far portions of the scene are separated by the random depth threshold d_{th} ranging from 0.3 to 0.7 percent of the maximum scene depth. Then, the near portions (where depth is less than or equal to d_{th}) of I_m are set as 1, and the far portions (where depth is greater than d_{th}) of I_m are set as 0.

[1] https://docs.opencv.org/3.4.2/d4/d86/group_imgproc_filter.html#gac05a120c1ae92a6060dd0db190a61afa.

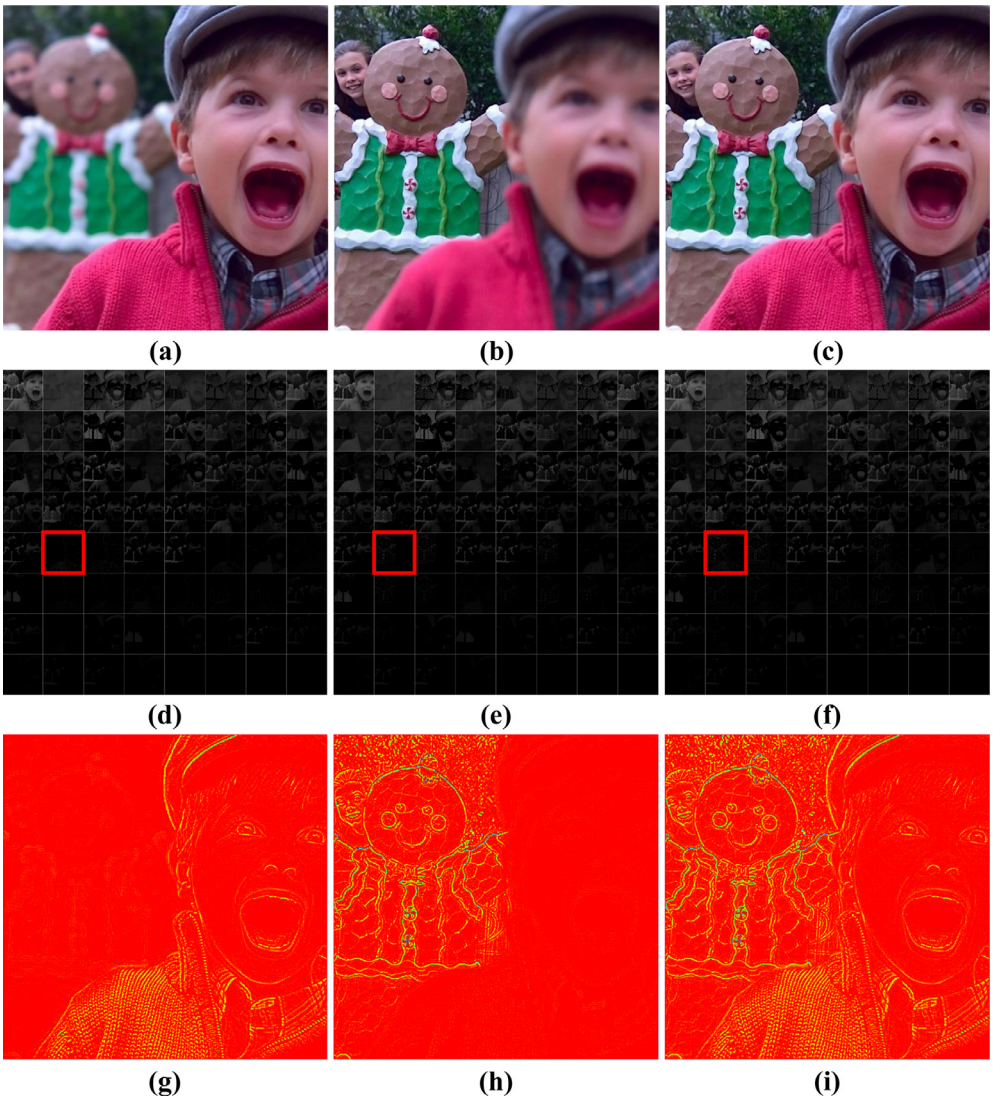

Fig. 2. Demonstration of feature extraction and feature fusion. (a) and (b) are a pair of multi-focus images. (c) is the fusion image of (a) and (b) produced by our image fusion model. (d) and (e) are respectively the 64 feature maps of (a) and (b) extracted by our feature extraction module (after CONV2). (f) shows the fused 64 feature maps of (a) and (b) by the feature fusion module (after FUSE). (g)–(i) indicate the closeups of feature maps bounded by red boxes in (d)–(f), which are projected to HSV color space for the clear observation of feature details. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

(3) a pair of multi-focus images are generated according to the RGB image I_s, blurry image I_b and focus map I_m as: the near focused image I_1 and far focused image I_2 can be generated according to Eq. (5). Naturally, I_1 and I_2 form a pair of multi-focus images, and I_s is their ground-truth fusion image I_g.

$$\begin{cases} I_1 = I_s \odot I_m + I_b \odot (\mathbf{1} - I_m) \\ I_2 = I_s \odot (\mathbf{1} - I_m) + I_b \odot I_m \end{cases}, \tag{5}$$

where $\mathbf{1}$ denotes the all one matrix of the same size with I_s, and $\odot$ denotes the elementwise product.

(4) in the training phase, the randomly generated multi-focus images and the corresponding source RGB images are further augmented by random resized-crop and random flips. Specifically, the final inputs and ground truth of our image fusion model are generated from each set of images (i.e., a pair of multi-focus images and one ground-truth fusion image) as follows. For each set of images, the two multi-focus images and the ground-truth fusion image are simultaneously processed by three procedures: firstly cropped by a random scale ranging from 0.5 to 1, then resized to 224×224, and finally randomly flipped in both vertical and horizontal directions. In the end, the processed multi-focus images are taken as the final inputs of the image fusion model, and the corresponding ground-truth fusion image is taken as the ground truth.

In this way, our multi-focus image dataset can be generated more naturally compared to the previous synthetic datasets, and the number of our multi-focus image pairs should be infinite owing to our random generation method. Moreover, the focus maps generated by our method vary with the random depth threshold, and thus our method could produce more diverse multi-focus image than the previous methods. Finally, the source RGB image can be taken as the ground-truth fusion image of the corresponding pair of multi-focus images, which is a great advantage over the previous dataset generation methods. Fig. 3 shows a demonstration example of our dataset generation method. It can be seen that the focus map in Fig. 3(d) has been reasonably generated by step (2) according to the depth range, and the multi-focus images in Figs. 3(e) and (f), generated by step (3) according to the RGB image, blurred image and focus map, are just like the naturally captured ones. As described in step (4), we have further augmented the multi-focus images and ground-truth images by random resized-crop, vertical flip and horizontal flip, thus we also show four sets of online generated multi-focus images and ground-truth fusion images for training our models in Fig. 4, from which we can see the blurring styles of the generated multi-focus images are related to the depth range and look natural.

Overall, compared to the previous datasets, our online generated multi-focus image dataset has four advantages: (1) having much larger image resolution, i.e., 224×224 compared to 16×16, 32×32 and

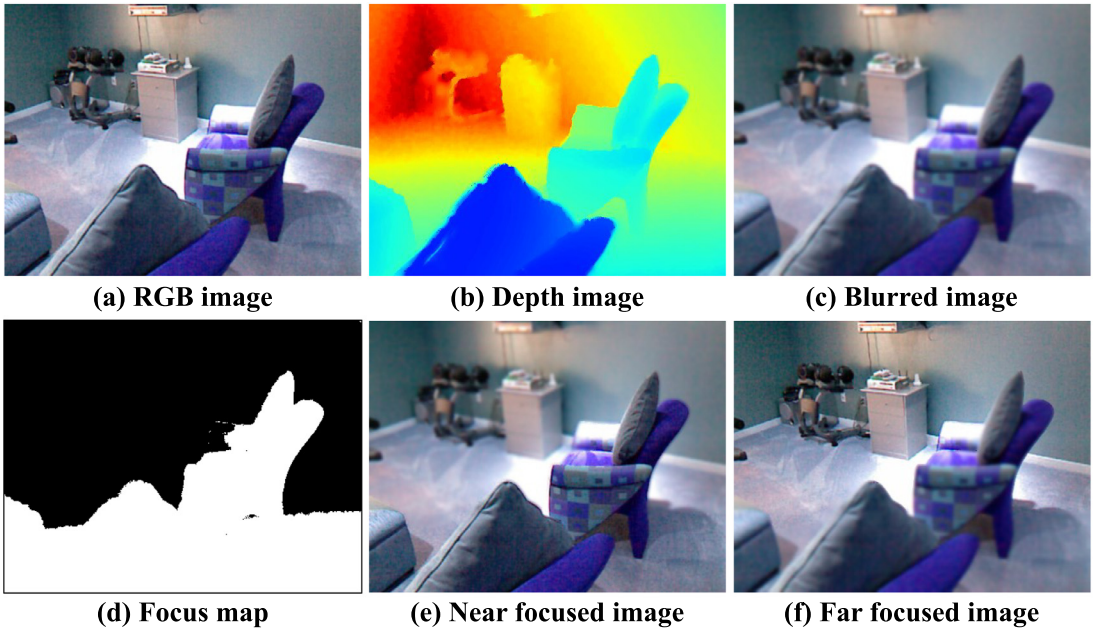

Fig. 3. Demonstration of generating multi-focus image dataset according to Section 2.3. (a) and (b) are a pair of RGB image and depth image in the NYU-D2 dataset. (c) is the randomly blurred image according to step (1). (d) is the randomly generated focus map according to step (2). (e) and (f) are a pair of multi-focus images generated from (a), (c) and (d) according to step (3).

64×64, (2) consisting of more multi-focus image pairs, (3) having more diverse blurring styles compared to fully focused-blurred style [31] and 12 handcrafted blurring styles [32], and (4) owning ground-truth fusion images. Owing to these advantages, our multi-focus image dataset can be used to fully and finely train the CNN based image fusion models.

2.4. Loss function

Prior to using the learning based algorithms, the model's parameters should be optimized with the appropriate loss function so as to obtain predictions similar to the ground truth. In this paper, the target of our image fusion model is to regress one informative fusion image from multiple input images. Mean square error (MSE) is a basic but often used loss function to regularize prediction of the model close to the ground-truth output. However, due to the characteristics of L_2-norm, only regularizing the model with MSE loss probably yields smooth fusion images. To solve this problem, the researchers usually introduce perceptual loss functions [42,43] to facilitate regularizing the network to produce images having more structural similarity with the ground-truth fusion images.

The perceptual loss functions are usually formulated as the mean square error of high-level (deep) convolutional features of the output image and ground-truth image. Since the high-level features are usually extracted by the CNNs pretrained for image classification, thus difference of the high-level features of the output image and ground-truth image could be used to discriminate whether they belong to the same category or they are the same object. In [43], the authors adopted the features of the fourth convolutional layer of the VGG16 [44] pretrained on ImageNet to abstract the high-level representations of input images. As is known, ResNet101 [45] could achieve better performance and extract deeper convolutional features than VGG16, and ideally the deeper convolutional features of ResNet101 could obtain better abstraction of images. Therefore, in this paper, we used the features of the last convolutional layer of the ResNet101 pretrained on ImageNet to construct our perceptual loss. To be specific, the proposed perceptual loss is formulated as the mean square error of feature maps of the predicted fusion image and ground-truth fusion image extracted by the last convolutional layer of ResNet101, as shown in Eq. (6).

$$P_{loss} = \frac{1}{C_f H_f W_f} \sum_{i,x,y} \left[f_p^i(x,y) - f_g^i(x,y) \right]^2, \tag{6}$$

where f_p and f_g respectively denote feature maps of the predicted fusion image and ground-truth fusion image. i denotes the channel index of feature maps. C_f, H_f and W_f denote the channel number, height and width of feature maps.

$$B_{loss} = \frac{1}{3 H_g W_g} \sum_{i,x,y} \left[I_p^i(x,y) - I_g^i(x,y) \right]^2, \tag{7}$$

where I_p and I_g respectively denote the predicted fusion image and ground-truth fusion image. i denotes the channel index of RGB images. H_g and W_g denote the height and width of the ground-truth fusion image.

During training the model, we firstly choose the mean square error (MSE) of the predicted fusion image and the ground-truth image as the basic loss (calculated as Eq. (7)) to pretrain the proposed model. Afterwards, we add the proposed perceptual loss to the basic loss to finely train the model, calculation of which can be expressed as Eq. (8).

$$T_{loss} = w_1 B_{loss} + w_2 P_{loss}, \tag{8}$$

where w_1 and w_2 respectively denote the weight coefficients of the basic loss and perceptual loss. In this paper, w_1 and w_2 are both set to 1, which is verified effective by the extensive experimental results.

Since all components of the proposed model and loss function are differentiable, thus the parameters of the proposed image fusion model can be learned through the error back-propagation method. Specially, in this paper, the parameters of our models are all updated through the stochastic gradient descent (SGD) back-propagation. More training details can be found in Section 2.5.

2.5. Training details

As reported in many literatures, training details are important for repeating the training process of the CNN models. Therefore, we have described our training methods in detail as follows.

First of all, the proposed model was pretrained on our online generated multi-focus image dataset with minibatch equal to 64 and

Fig. 4. Four sets of the generated multi-focus images and ground-truth fusion images. Each row shows a set of images, which respectively are the near focused image, far focused image and their ground-truth fusion image from left to right.

under the regularization of the basic loss (B_{loss}) for 5000 iterations, during which the momentum of the batch-normalization (BN) layers was linearly decreased from 0.99 to 0. Afterwards, we freezed the parameters of the BN layers and adopted our integrated loss (T_{loss}) to finely train the parameters of the convolutional layers for 60,000 iterations. Because T_{loss} requires relatively large computational resource when fine-training the model, thus minibatch was decreased from 64 to 32 in the fine-training procedure.

As for settings of learning rates, the basic learning rates during pretraining and fine-training are both set to 0.01, and are gradually decreased to 0 according to the 'poly' learning rate policy. That is, the learning rate lr_i in the ith iteration equals to the basic learning rate lr_0 multiplying $(1 - i/maxI)^{power}$, which can be expressed as Eq. (9). During the fine-training procedure, the online generated multi-focus dataset was further augmented by randomly tuning HSV channels of input images and ground-truth image, by multiplying each channel by a random ratio ranging from 0.8 to 1.2. Then, the color space of input images and ground-truth image were randomly transformed to grayscale at the probability of 0.5, which was helpful for the image fusion model to prevent color-shifting problem.

$$lr_i = lr_0 \times (1 - i/maxI)^{power}, \tag{9}$$

where *maxI* denotes the maximum permitted iteration number and *power* is used to tune the decreasing rate of lr_i. In this paper, *power* is set to 0.9 in all experiments.

Finally, all the proposed models are implemented in the PyTorch framework,[2] and trained and tested on a platform with Intel Core i7-3770k CPU and NVIDIA TITAN X GPU. The consumption amounts of computational resources during the pretraining, fine-training, and inference procedures are respectively detailed as follow. Pretraining our model occupies about 10 GB GPU memory and takes about 8.5 h, fine-training the model occupies about 9 GB GPU memory and takes about 44 h, and inferring one fusion image of size 520×520 from two input images occupies about 785 MB GPU memory and takes about 0.02 s. Therefore, our proposed IFCNN can be conveniently deployed in the realtime applications, without consuming much computational resources. The implementation code of our image fusion model will be available on the project page.[3]

3. Experimental results and discussions

In this section, we have done extensive experiments to validate the performance of the proposed image fusion model. Firstly, the experimental settings are briefly described, then qualitative and quantitative results are illustrated and discussed, and conclusions of this section are made in the end.

3.1. Experimental settings

In order to verify the advantages of our proposed model, we have compared it with four representative image fusion algorithms on four types of image datasets, and evaluated the algorithms in both the qualitative and quantitative ways. The comparison algorithms, image datasets, and evaluation methods are respectively introduced below.

3.1.1. Comparison algorithms

Since the proposal of our image fusion model is inspired by the framework of transform-domain image fusion algorithms, thus we compare our proposed model with one representative transform-domain algorithms to validate the effectiveness of our proposed model, i.e., multi-scale transform and sparse representation based image fusion algorithm (LPSR) [24]. Besides, we compare our model with the state-of-the-art guided filtering based image fusion algorithm (GFF) [4], which is also a general image fusion algorithm. Even though our model is only trained on the multi-focus image dataset, it is designed as a general image fusion model for fusing various types of images. Therefore, we further compare our model with two existing image fusion models (i.e., multi-focus image fusion model (MFCNN) [31] and multi-exposure image fusion model (MECNN) [34]), to validate that the proposed model could achieve comparable or even better performance than the current state-of-the-art image fusion models.

As for our own algorithms, we have compared the above four algorithms with three models implemented with elementwise-maximum, elementwise-mean and elementwise-sum fusion rules, respectively named as IFCNN-MAX, IFCNN-MEAN and IFCNN-SUM. As discussed in Section 2.2, we choose IFCNN-MAX as our chief model to fuse the multi-focus images, infrared and visual images, and multimodal medical images, and choose IFCNN-MAX trained only with B_{loss} (abbreviated as BASELINE-MAX) as the baseline model for fusing these three types of images. In addition, we choose IFCNN-MEAN as our chief model to fuse the multi-exposure images, and accordingly choose IFCNN-MEAN trained only with B_{loss} (abbreviated as BASELINE-MEAN) as the baseline model for fusing the multi-exposure images.

[2] https://pytorch.org.

[3] https://github.com/uzeful/IFCNN.

3.1.2. Image datasets

In order to fully demonstrate the effectiveness of the proposed image fusion model, we have evaluated the compared algorithms on four types of image datasets, including multi-focus image datasets [46], infrared and visual image dataset [15], multi-modal medical image dataset [47], and multi-exposure image dataset [48]. The four image datasets are respectively shown in Figs. 5–8.

3.1.3. Evaluation methods

During evaluating the image fusion algorithms, we have adopted both qualitative and quantitative methods to discriminate the performance of different image fusion algorithms. Firstly, qualitative evaluation is performed by judging the visual effects of their fusion images with respect to the source images. To be specific, whether the visual effects of fusion images are satisfying for each type of image dataset can be judged by the following criteria: (1) multi-focus image fusion should integrate as much clear and sharp features as possible from each source image into the fusion image, (2) infrared and visual image fusion should preserve as much visible appearance information as possible from the visual image and inject as much salient bright features as possible from the infrared image into the fusion image, (3) multi-modal medical image fusion should combine as much typical features as possible from source images of different modalities into the fusion image, and (4) multi-exposure image fusion should inject as much clear middle-exposure features as possible from each source image into the fusion image. Besides the above criteria, another important criterion is that the fusion images should look as natural as possible so that the human eyes can easily and accurately obtain the comprehensive information from the fusion image.

Since only comparing the visual quality probably cannot objectively and fairly discriminate the performance of different image fusion algorithms. Therefore, we further utilize five often used metrics to evaluate the quantitative performance of the algorithms on the multi-focus, infrared-visual and multi-modal medical image datasets. The five metrics are respectively the visual information fidelity (VIFF) [49], improved structural similarity (ISSIM) [50], normalized mutual information (NMI) [51], spatial frequency (SF) [52], and average gradient [53]. Among the five metrics, VIFF measures the visual information fidelity of the fusion image with respect to source images, ISSIM measures the structural similarity between the fusion image and source images, NMI measures the information amount of the fusion image that has been preserved from two source images, and SF and AG measure the textural information amount of the fusion image from two different statistical views. Evaluating fusion results with these five metrics can effectively reflect the algorithms' abilities on integrating the visual information, structural information, and on merging the image details, therefore the selection of these five metrics is appropriate.

Especially, there are more than two source images in each set of multi-exposure images (see Fig. 8), thus the metrics (VIFF, ISSIM and NMI) designed for two input images are not valid while evaluating algorithms on the multi-exposure image dataset. Therefore, we select SF, AG and another structural similarity metric MESSIM [7] (especially designed to evaluate the multi-exposure image fusion algorithms), to quantify the performance of the algorithms on fusing multi-exposure images. Finally, note that greater values of VIFF, ISSIM, NMI, SF, AG and MESSIM indicate better performance of the algorithms. In the next four subsections, the evaluation results on four types of image datasets are respectively described and discussed.

3.2. Multi-focus image fusion

Since our model has been only trained on the multi-focus image dataset, thus we firstly want to investigate the performance of the proposed model on fusing multi-focus images. Additionally, we also want to test the effectiveness of the perceptual loss and fusion rules on this dataset. In order to achieve the above two purposes, we

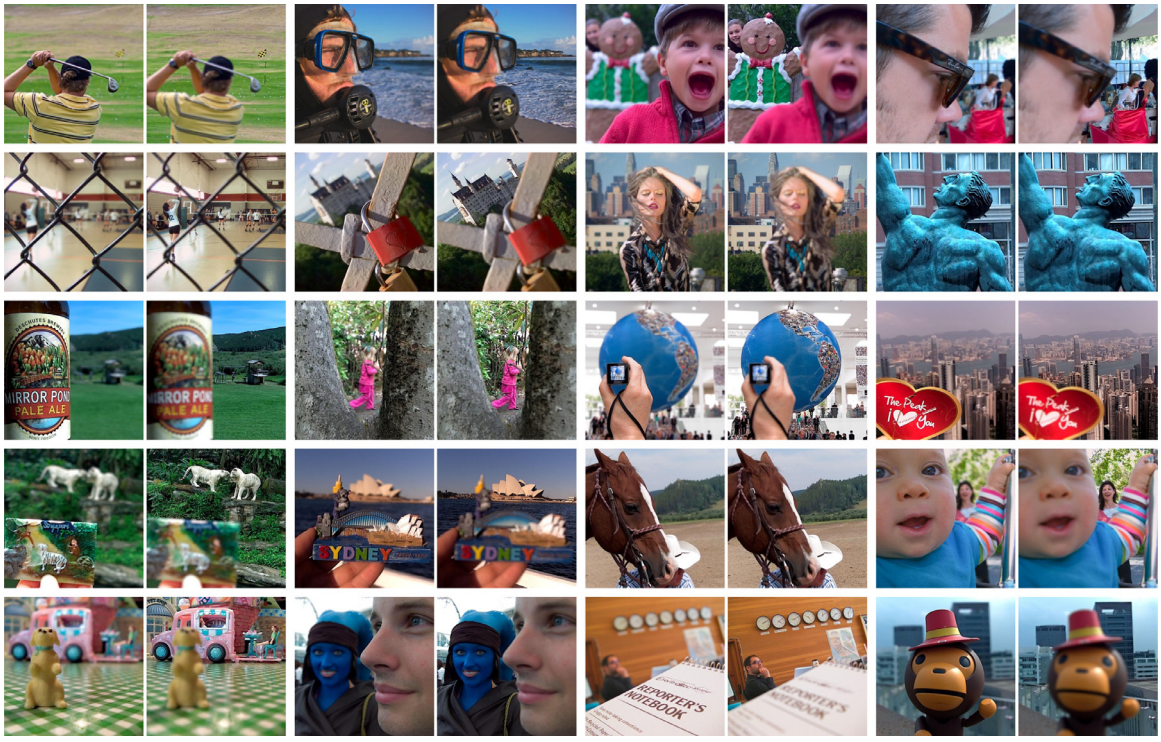

Fig. 5. The multi-focus image dataset. This dataset includes 20 pairs of near and far focused images. In each pair of images, the left one is the near focused image and the right one is the far focused image.

Fig. 6. The infrared and visual image dataset. This dataset includes 14 pairs of infrared and visual images. In each pair of images, the left one is the visual image and the right one is the infrared image.

have evaluated the algorithms on the multi-focus image dataset (see Fig. 5), and shown two comparison examples on this dataset in Figs. 9 and 10.

Fig. 9(c)–(j) show the fusion results of Fig. 9(a) and (b), which captured the volleyball court in front of the chain-link fence. The ideal fusion of Fig. 9(a) and (b) should directly combine the clear chain-link fence of the near focused Fig. 9(a) and the clear volleyball court of the far focused Fig. 9(b) together into the fusion image. Fig. 9(c) shows that the fusion image of GFF shows a little blurring effect around the fence (see the closeups) than other algorithms, which might be caused by its smoothing operation on weight map. It can been seen from Fig. 9(e) that MFCNN fails to fuse the clear court behind the fence corner (as

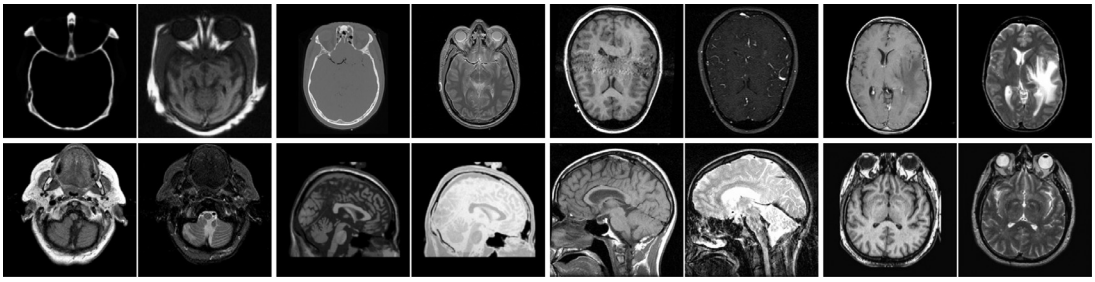

Fig. 7. The multi-modal medical image dataset. This dataset includes eight pairs of multi-modal medical images. In the first two pairs of images on the top row, the left one is CT image and the right one is MR image. In other pairs of images, the two images are MR images of different modalities.

Fig. 8. The multi-exposure image dataset. This dataset includes six sets of multiple images with different exposure degrees. Each row shows one set of multi-exposure images, in which exposure degrees of images from left to right are gradually ranging from under-exposure to over-exposure.

pointed by the red arrow in the closeup) into the fusion image due to its inaccurate focus map, while the other algorithms have all integrated this clear portion into their fusion images. Fig. 9(f) shows that MECNN has fused the salient features into the fusion image, contrast of which however has been degraded by a large margin compared to the source images. Finally, as shown in Figs. 9(d) and (g)–(j), the fusion images of LPSR and our IFCNN-SUM, IFCNN-MEAN, BASELINE-MAX and IFCNN-MAX all have well integrated the clear features of both source images, and showed better visual effects compared to those of other algorithms.

Fig. 10(c)–(j) shows the fusion results of Fig. 9(a) and (b), which captured a souvenir in front of Sydney Opera House. The ideal fusion image of Fig. 10(a) and (b) should directly combines the clear souvenir and fingers of the near focused Fig. 10(a) and the clear background of the far focused Fig. 10(b). It can be seen from Figs. 10(c) and (e) that the fusion images of GFF and MFCNN suffer from blurring effect around the koala's right ear (as pointed by the red arrows in the closeups). Figs. 10(f) shows that the fusion image of MECNN is still of low-contrast, which implies the generalization ability of MECNN is weak for fusing multi-focus images. In the end, Figs. 10(d) and (g)–(j) show that LPSR and our proposed four models have finely injected the sharp features into their fusion images, and achieve comparable performance.

Besides the qualitative comparison, we have also calculated the five quantitative metrics of the fusion results as described in Section 3.1. The quantitative results are listed in Table 1. In this table and the tables in the following subsections, the value in **bold font** and value in *italic font* respectively denote the best result and second-best result in the corresponding metric row. Each value before the bracket denotes the mean metric value on the full dataset, and the three integers in each bracket respectively represent the overall rank, the number of fusion images ranking first and the number of fusion images ranking second of the current algorithm under the evaluation of the current metric among all algorithms. We can see from the results in Table 1 that LPSR

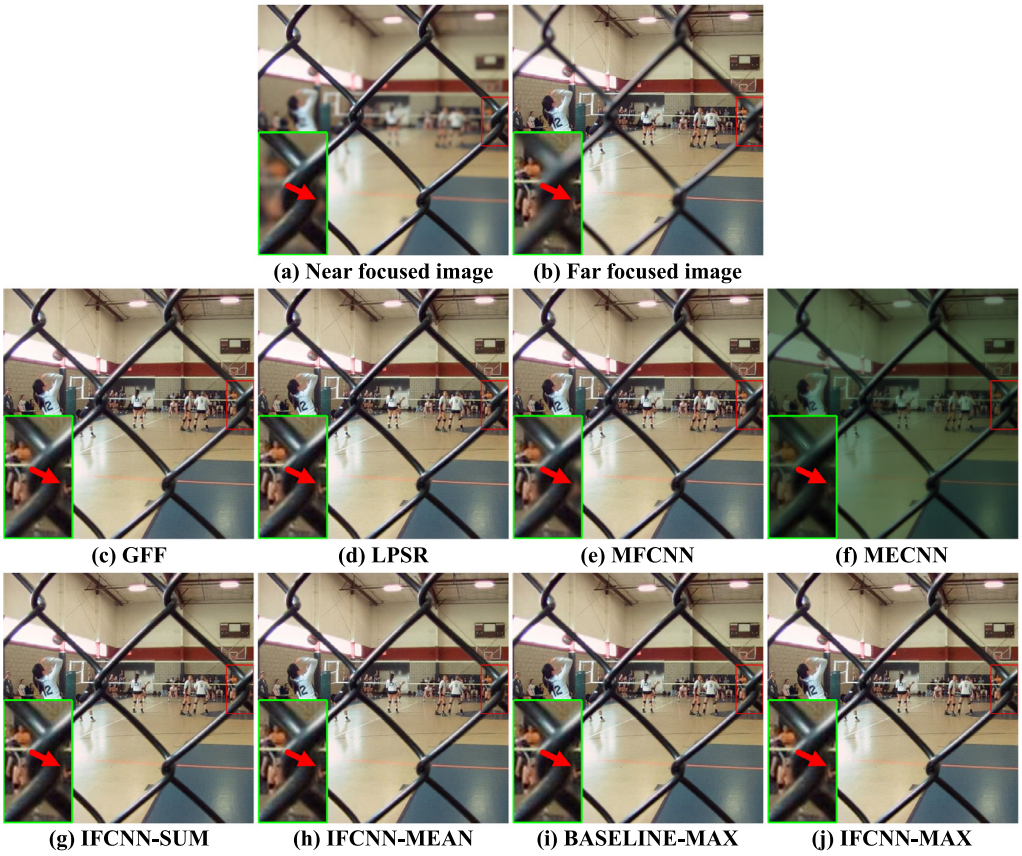

Fig. 9. The comparison example on the fifth pair of multi-focus images. (a) and (b) are fifth pair of multi-focus images. (c)-(j) are the fusion images of (a) and (b) respectively by GFF, LPSR, MFCNN, MECNN, IFCNN-SUM, IFCNN-MEAN, BASELINE-MAX and IFCNN-MAX. In each subfigure, the image patch bounded by green box indicates the closeup of the image patch bounded by red box. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 1
Quantitative evaluation results on multi-focus image dataset.

Metrics	GFF	LPSR	MFCNN	MECNN	IFCNN-SUM	IFCNN-MEAN	BASELINE-MAX	IFCNN-MAX
VIFF	0.9806(5,0,6)	**0.9930(1,9,2)**	0.9806(4,0,1)	0.5606(8,0,0)	0.9800(7,1,2)	0.9821(6,4,2)	0.9809(3,3,5)	*0.9823(2,3,2)*
ISSIM	0.6187(6,1,3)	0.6177(7,0,1)	0.6205(5,4,3)	0.5463(8,2,0)	0.6312(3,3,3)	**0.632(1,5,4)**	*0.6318(2,4,3)*	0.6293(4,1,3)
NMI	*1.044(2,0,16)*	0.9817(3,0,1)	**1.098(1,18,2)**	0.8436(8,2,1)	0.8576(6,0,0)	0.8535(7,0,0)	0.8664(5,0,0)	0.9034(4,0,0)
SF	19.29(4,0,3)	*19.41(2,8,7)*	19.21(5,0,0)	7.888(8,0,0)	18.93(7,1,0)	19.01(6,0,1)	19.33(3,6,1)	**19.42(1,5,8)**
AG	2.864(3,1,3)	**2.891(1,12,4)**	2.854(4,0,0)	1.194(8,0,0)	2.834(7,2,1)	2.844(5,1,2)	2.837(6,0,0)	*2.886(2,4,10)*

achieves the highest VIFF metric value and our chief model IFCNN-MAX ranks second, which implies these two algorithms have obtained higher visual information fidelity compared to other algorithms. As for the ISSIM metric, our proposed four models have achieved greater values than other algorithms, which indicates our proposed models have preserved more structural information than other algorithms. In addition, differences of ISSIM metric values within our four models are relatively small, which indicates our four models have preserved comparable structural information from input images to their fusion images. As is known, NMI relates to the joint distribution of gray values, and its value is greater if the distribution of gray values between input images and fusion image is more similar. Thus, direct combination or slight weighted addition of input images (MFCNN and GFF) could obtain higher NMI metric values compared to the general transform-domain image fusion algorithms (LPSR, MECNN and our IFCNNs). In comparing SF and AG, our IFCNN-MAX and LPSR obtain close metric values and respectively rank the best place and second place, which implies IFCNN-MAX and LPSR have produced fusion images with more textural details than other algorithms. Especially, due to the degradation of contrast information, MECNN has obtained the lowest values on all metrics. According to the comprehensive evaluation on the multi-focus image dataset, our IFCNN-MAX could retain relatively higher visual information fidelity, preserve more structural information, and produce informative fusion images compared to other algorithms. Therefore, our IFCNN-MAX has demonstrated comparable or even better performance on the multi-focus image dataset compared to other state-of-art algorithms.

As for the ablation study, we can see that our chief model IFCNN-MAX has obtained greater metric values than the baseline model BASELINE-MAX on most of the metrics except ISSIM, and also achieved better quantitative results than IFCNN-SUM and IFCNN-MEAN in most of cases. Therefore, in total, the quantitative results on the multi-focus image dataset indicate the elementwise-maximum fusion rule performs better than elementwise-sum and elementwise-mean for fusing multi-focus images, and the model trained with perceptual loss outperforms the model trained only with MSE loss.

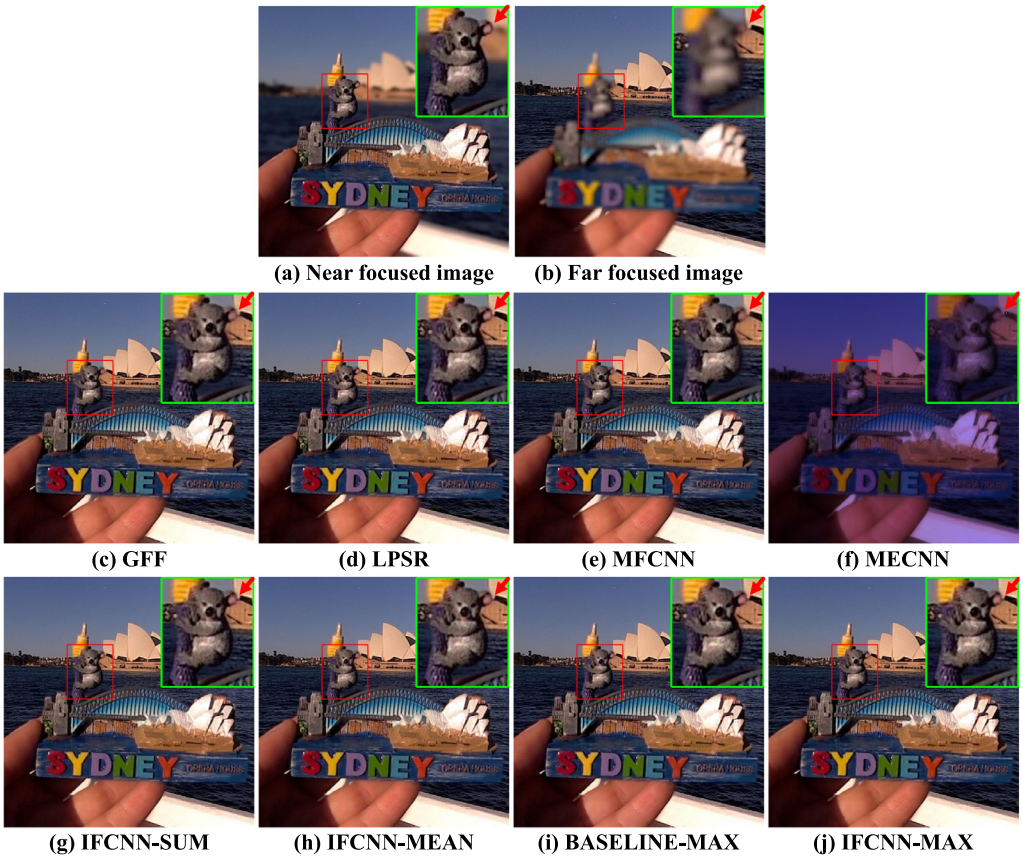

Fig. 10. The comparison example on the 14th pair of multi-focus images. (a) and (b) are 14th pair of multi-focus images. (c)-(j) are the fusion results of (a) and (b) respectively by GFF, LPSR, MFCNN, MECNN, IFCNN-SUM, IFCNN-MEAN, BASELINE-MAX and IFCNN-MAX. In each subfigure, the image patch bounded by green box indicates the closeup of the image patch bounded by red box. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3.3. Infrared and visual image fusion

In this subsection, we have compared the image fusion algorithms on the infrared and visual image dataset (see Fig. 6), and two comparison examples have been shown in Figs. 11 and 12.

Figs. 11(c)–(j) show the fusion results of Figs. 11(a) and (b), which captured the outdoor scene with a person standing in the mountain. A good fusion image of Figs. 11(a) and (b) should preserve as much as salient bright features (person and several spots in this case) and also maintain the visible appearance features (house, fence and trees in this case) from the visual image [16]. Fig. 11 shows that all algorithms have somewhat integrated the salient features of infrared and visual images into their fusion images. However, the fusion images of MECNN, IFCNN-SUM and IFCNN-MEAN (see Figs. 11(f)–(h)) have relatively low contrast compared to those of other algorithms. Figs. 11(e) shows that MFCNN fails to preserve much visible appearance features of fence and trees from the visual image and also fails to integrate one bright spot on top of the infrared image into the fusion image. While Figs. 11(c) and (d) show that GFF and LPSR have lost more appearance features of fence and trees around the person (pointed by the yellow arrows in the closeups) compared to IFCNN-MAX and BASELINE-MAX. Furthermore, we have observed their closeups around the person in Figs. 11(i) and (j) to closely compare IFCNN-MAX and BASELINE-MAX, and it can be seen that IFCNN-MAX has integrated more complete bright features of the person into the fusion image than BASELINE-MAX. However, the overall appearances of Figs. 11(i) and (j) show that BASELINE-MAX has preserved more visible features (see the trees in the bottom left corner) from the visual image than IFCNN-MAX.

Figs. 12(a) and (b) are the 10th pair of infrared and visual images, which captured the night scene of a street with several persons and two cars. Ideally, fusion of Figs. 12(a) and (b) should directly inject the salient bright features of persons, cars and traffic lights from the infrared image into the visual image, so that the fusion image can preserve most of visual appearance features of the visual image while integrating the salient bright features of the infrared image. Figs. 12(c) and (d) show that GFF and LPSR have integrated too much useless bright background features into their fusion images, which makes the local-contrast of their fusion images lower than that of other algorithms (except MECNN). The fusion image of MFCNN yields the region artifact around the person (pointed by the yellow arrow in the closeup of Fig. 12(c)) and also fails to integrate much useful bright features (such as the bright persons at top-right corner and two bright traffic lights beside road), due to inappropriate focus map generated by MFCNN. Fig. 12(f) shows that MECNN has integrated the salient features of infrared and visual image into the fusion image, which however is still under low-contrast. As for our four models, the fusion images of IFCNN-SUM and IFCNN-MEAN (see Figs. 12(g) and (h)) have lower contrast than those of BASELINE-MAX and IFCNN-MAX (see Fig. 12(i) and (j)). Finally, Figs. 12(i) and (j) show that both BASELINE-MAX and IFCNN-MAX have mainly injected the useful bright features from the infrared image into the fusion image and preserved most of the visible appearance features from the visual image to the fusion image, thus the fusion images of BASELINE-MAX and IFCNN-MAX are more suitable and comprehensive for visual perception than those of other algorithms. Compared with BASELINE-MAX, the fusion image of IFCNN-MAX has integrated more useful infrared features as pointed by the yellow arrows

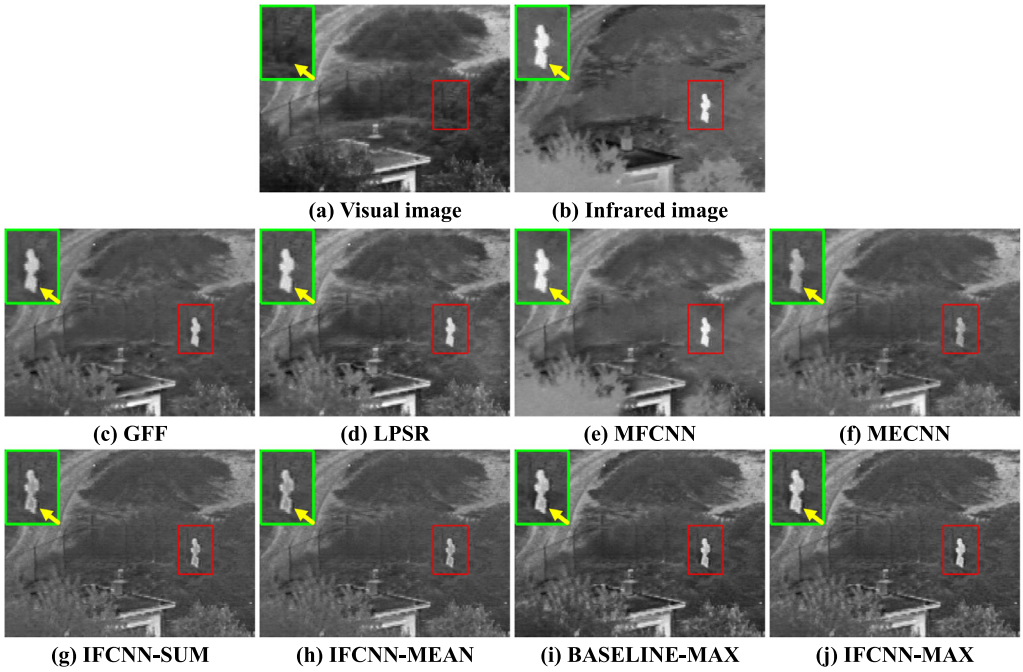

Fig. 11. The comparison example on the second pair of the infrared and visual images. (a) and (b) are respectively the visual image and infrared image, and (c)–(j) are the fusion results of (a) and (b) respectively by GFF, LPSR, MFCNN, MECNN, IFCNN-SUM, IFCNN-MEAN, BASELINE-MAX and IFCNN-MAX. In each subfigure, the image patch bounded by green box indicates the closeup of the image patch bounded by red box. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 2
Quantitative evaluation results on infrared and visual image dataset.

Metrics	GFF	LPSR	MFCNN	MECNN	IFCNN-SUM	IFCNN-MEAN	BASELINE-MAX	IFCNN-MAX
VIFF	0.5507(5,0,3)	*0.7327(2,6,2)*	**0.7757(1,6,0)**	0.3167(8,0,0)	0.4506(6,0,0)	0.4495(7,0,0)	0.6621(3,2,4)	0.6405(4,0,5)
ISSIM	0.4272(8,2,1)	0.4364(7,0,3)	**0.6084(1,7,0)**	0.4395(6,0,1)	0.5039(3,0,4)	*0.5041(2,2,1)*	0.5002(4,3,4)	0.4956(5,0,0)
NMI	0.3614(6,0,4)	0.365(5,1,1)	**1.047(1,12,1)**	0.3783(3,1,1)	0.3384(8,0,0)	0.3402(7,0,0)	*0.3945(2,0,7)*	0.3699(4,0,0)
SF	9.623(6,0,0)	10.48(3,0,4)	9.503(7,1,0)	6.242(8,0,0)	9.936(5,0,0)	9.955(4,0,0)	*11.08(2,4,6)*	**11.31(1,9,4)**
AG	1.505(6,0,0)	1.678(3,1,3)	1.358(7,0,0)	0.9449(8,0,0)	1.600(4,0,0)	1.598(5,0,0)	*1.781(2,1,9)*	**1.865(1,12,2)**

in the closeups of Figs. 12(i) and (j), and has preserved much visible appearance features. Thereby, IFCNN-MAX outperforms BASELINE-MAX by a bit for fusing infrared and visual images.

Besides the above two comparison examples, the fusion results on another four pairs of infrared and visual images have been briefly illustrated in Fig. 13 to further validate our previous conclusions. In each column of Fig. 13, the top two rows show a pair of visual and infrared images, and the bottom five rows from top to bottom show the fusion images of the visual image and infrared image respectively by GFF, LPSR, MFCNN, MECNN and IFCNN-MAX. As shown in third and fourth rows of Fig. 13, GFF and LPSR have only integrate very few visible appearance features from the visual images to their fusion images in most of cases, which thus will impact the visual perception of the supervised scene. The fifth row of Fig. 13 shows that the fusion images of MFCNN suffer from severe region artifacts in all four cases, and the fusion image in second column has even lost the visible head features of the visual image and the important bright gun features of infrared image. As for MECNN, the sixth row of Fig. 13 shows that all fusion images of MECNN have much lower contrast compared to those of other algorithms. Finally, the bottom row of Fig. 13 shows that the fusion images of IFCNN-MAX have appropriately combined the visible appearance features of visual image and the salient bright features of infrared image and show the best visual effect in most of cases.

Moreover, we have evaluated the quantitative performance of the image fusion algorithms on the infrared and visual image dataset, and the metric values are listed in Table 2. Different from multi-focus images, correlation of the infrared image and visual image is usually low and thus their features are often directly complementary. Therefore, directly combining the salient regions of the infrared image and visual image would usually yield additive effect (i.e., achieving high values) on metrics that measure the mutual feature or information correlation between input images and fusion image, as the VIFF, ISSIM and NMI values of MFCNN shown in Table 2. This is the reason why although most fusion images of MFCNN yield inappropriate region effects as shown in Fig. 13, MFCNN could still obtain such high values on VIFF, ISSIM and NMI. Except MFCNN, our BASELINE-MAX and IFCNN-MAX could achieve relatively high metric values on VIFF, ISSIM and NMI metrics, thus our two chief models could preserve much visual information and structural information and maintain much original gray-value distribution. Moreover, the metric values of SF and AG rank the algorithms in consistence to our visual judgement. Especially, our chief models IFCNN-MAX and BASELINE-MAX respectively rank the first place and second place, which means these two models could produce fusion images with more textural details compared to other algorithms.

Overall, the qualitative and quantitative evaluation results on the infrared and visual image dataset imply that the fusion images of IFCNN-MAX have showed best visual effects and preserved much textural information from input images, and IFCNN-MAX has demonstrated better generalization ability for fusing various types of images than MFCNN and MECNN.

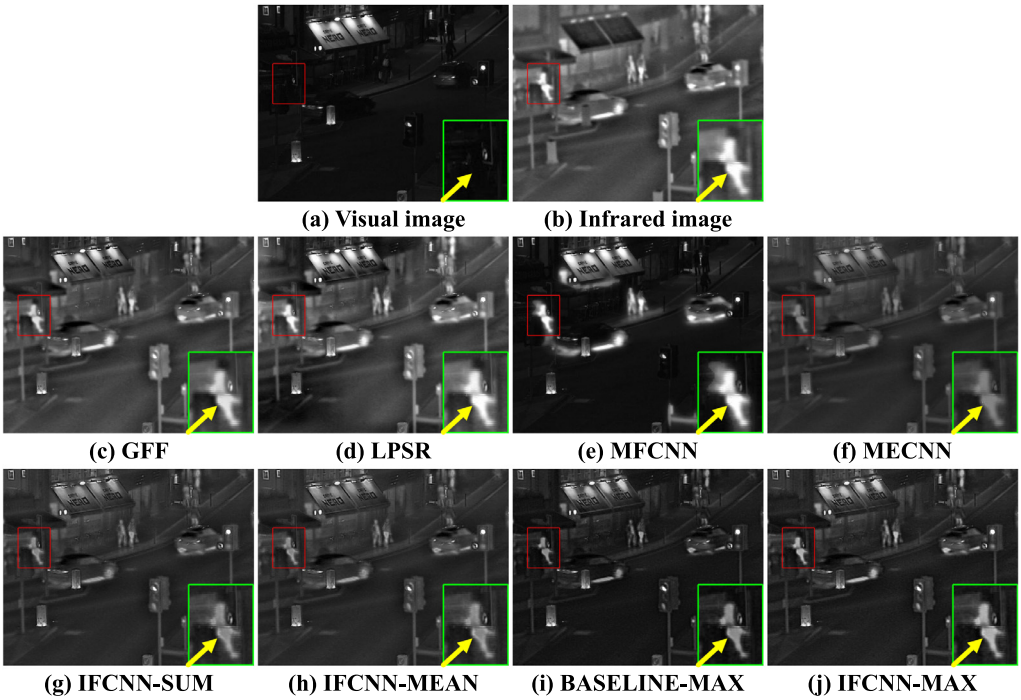

Fig. 12. The comparison example on the 10th pair of the infrared and visual images. (a) and (b) are respectively the visual image and infrared image, and (c)-(j) are the fusion results of (a) and (b) respectively by GFF, LPSR, MFCNN, MECNN, IFCNN-SUM, IFCNN-MEAN, BASELINE-MAX and IFCNN-MAX. In each subfigure, the image patch bounded by green box indicates the closeup of the image patch bounded by red box. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3.4. Medical image fusion

In this subsection, we have evaluated the image fusion algorithms on the multi-modal medical image dataset (see Fig. 7), and one comparison example is shown in Fig. 14.

Figs. 14(c)–(j) show the fusion results of Figs. 14(a) and (b), which are a pair of CT and MR slices scanned along the axial plane of human brain. Ideally, fusion of Figs. 14(a) and (b) should integrate the bright skull features of the CT image and also the textural tissue features of the MR image into the fusion image. Figs. 14(c)–(e) show that GFF, LPSR and MFCNN have failed to inject some portions of bright skull features (pointed by the yellow arrows in the closeups) from the CT image into the fusion image, and especially the fusion image of MFCNN lacks the largest portion of skull features. It can be seen from Fig. 14(f) that MECNN has successfully integrated the salient features of the CT and MR images, but its fusion image suffers from blurring effect. Finally, Figs. 14(g)–(j) show that our proposed four models have fused most of the bright skull and textural tissue features of the CT and MR images into their fusion images, among which the fusion image of IFCNN-MAX has integrated most bright skull features. Therefore, in this comparison example, IFCNN-MAX has generated the fusion image with better visual quality than other algorithms.

As discussed in Section 3.3, if the input images have low correlation between each other, MFCNN could obtain higher values of VIFF, ISSIM and NMI. Since the medical images of different modalities are also lowly correlated due to their different imaging principles, thus the similar phenomenon also occurs on the multi-modal medical image dataset, i.e., it can be seen from Table 3 that MFCNN has ranked the first place in all VIFF, ISSIM and NMI metrics. Except MFCNN, LPSR have obtained higher VIFF, ISSIM and NMI values than other algorithms, and our BASELINE-MAX and IFCNN-MAX just rank after LPSR. While our IFCNN-MAX and BASELINE-MAX still rank the first and second place on the metric values of SF and AG, which indicates IFCNN-MAX could integrate much more textural information from input images into the fusion image than other algorithms. In addition, IFCNN-MAX has beat BASELINE-MAX in almost all metrics except VIFF, thus the perceptual loss is effective for boosting the performance of our image fusion model. Overall, according to the qualitative and quantitative evaluations, IFCNN-MAX could inject more useful features of input images into the fusion image, and perform comparably or even better than other algorithms for fusing multi-modal medical images.

3.5. Multi-exposure image fusion

Finally, the image fusion algorithms are evaluated on the multi-exposure image dataset (see Fig. 8), and one comparison example is illustrated in Fig. 15.

Figs. 15(a)–(e) show that this image set contains five source images ranging from low-exposure degree to high-exposure degree, which gradually capture the outdoor scene to the indoor layout. The ideal fusion image of this image set should integrate the clear portions of each source image, which usually correspond to the image portions under appropriate exposure degree (i.e., middle exposure degree). To be specific, the ideal fusion image should should integrate outdoor scene including blue sky, table and house from Fig. 15(a) and (b), the bottom indoor layout from Figs. 15(c) and (d), and the top indoor layout from Figs. 15(d) and (e). It can be seen from Figs. 15(f)–(m) that except MECNN all the other algorithms have integrated the visible outdoor scene and indoor layout together into their fusion images. The reason that MECNN fails to fuse this set of multi-exposure images might be because MECNN is originally trained to fuse two images and cannot be directly applied to well fuse more than two images, which indicates the weak generalization of MECNN for changing the number of inputs. Even though GFF, LPSR, IFCNN-SUM and IFCNN-MAX have integrated the visible outdoor and indoor features, their fusion images have somewhat inappropriate visual effects compared to those of BASELINE-MEAN and IFCNN-MEAN. For instance, it can be seen from the closeups in Figs. 15(f) and (g) that the fusion images of GFF and LPSR show lower contrast on indoor layout than those of our four models. Fig. 15(h) shows that MFCNN mainly combines the outdoor scene of Fig. 15(b)

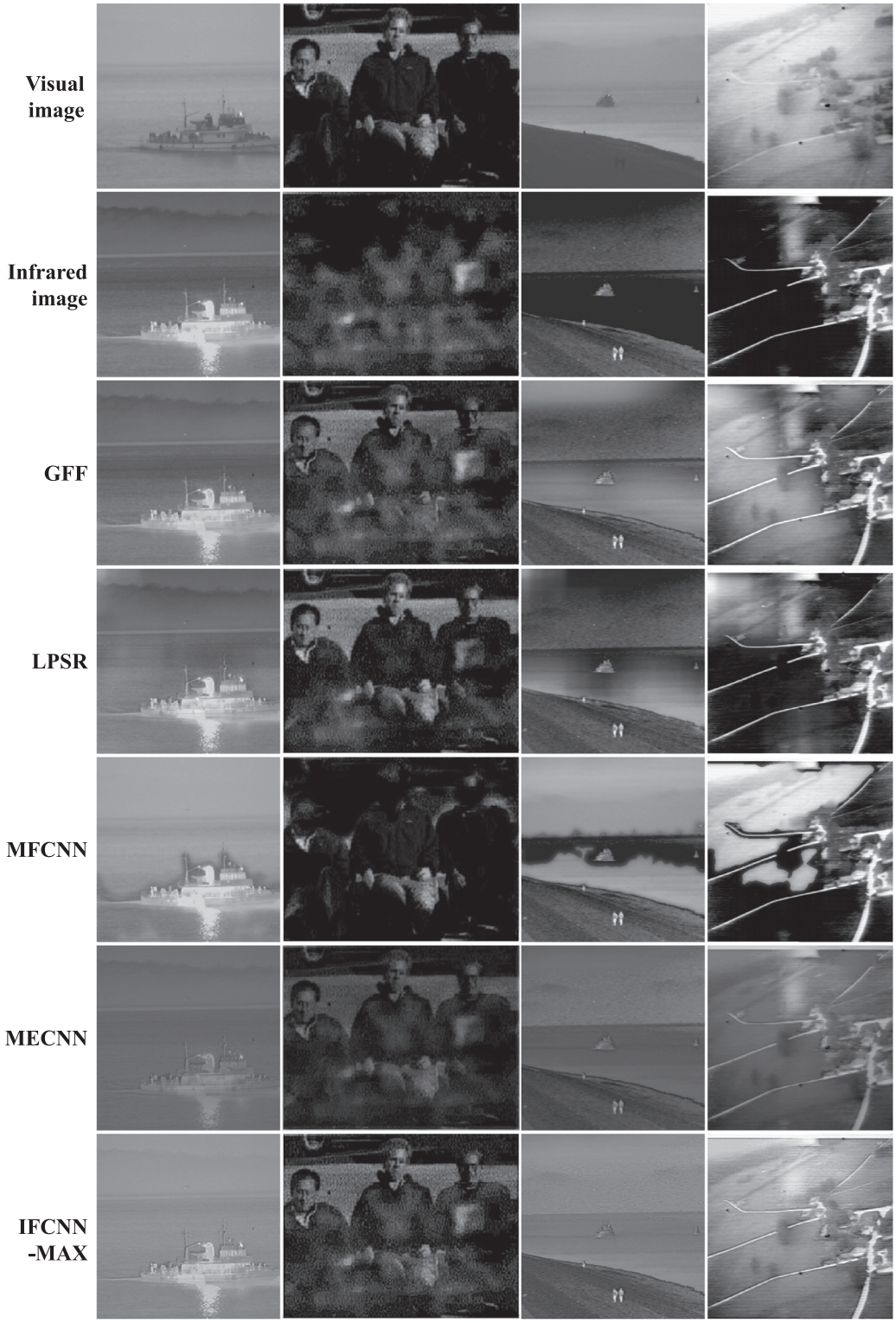

Fig. 13. In each column, the top two images are respective the visual image and visual image, and images in the bottom five rows are respectively the fusion images of top two source images by GFF, LPSR, MFCNN, MECNN and IFCNN-MAX.

and the indoor layout of Fig. 15(e) into the fusion image, which shows inappropriate region effect around the stitching area. In addition, MFCNN has not fully utilized the clear features of all source images, thus the visual quality of its fusion image is far from perfect for visual perception. Fig. 15(j) shows that the fused outdoor scene of IFCNN-SUM seems over-exposed and thus loses much outdoor details, which might be caused by adding too much salient features of more than two images. While Fig. 15(k) indicates that IFCNN-MAX have addressed textural details so much that the fusion image shows edging effect, which might impact the perception of human eyes. Through observing the closeups of Figs. 15(l) and (m), we can see the fusion image of IFCNN-MEAN is a little brighter than that of BASELINE-MEAN, which is helpful for

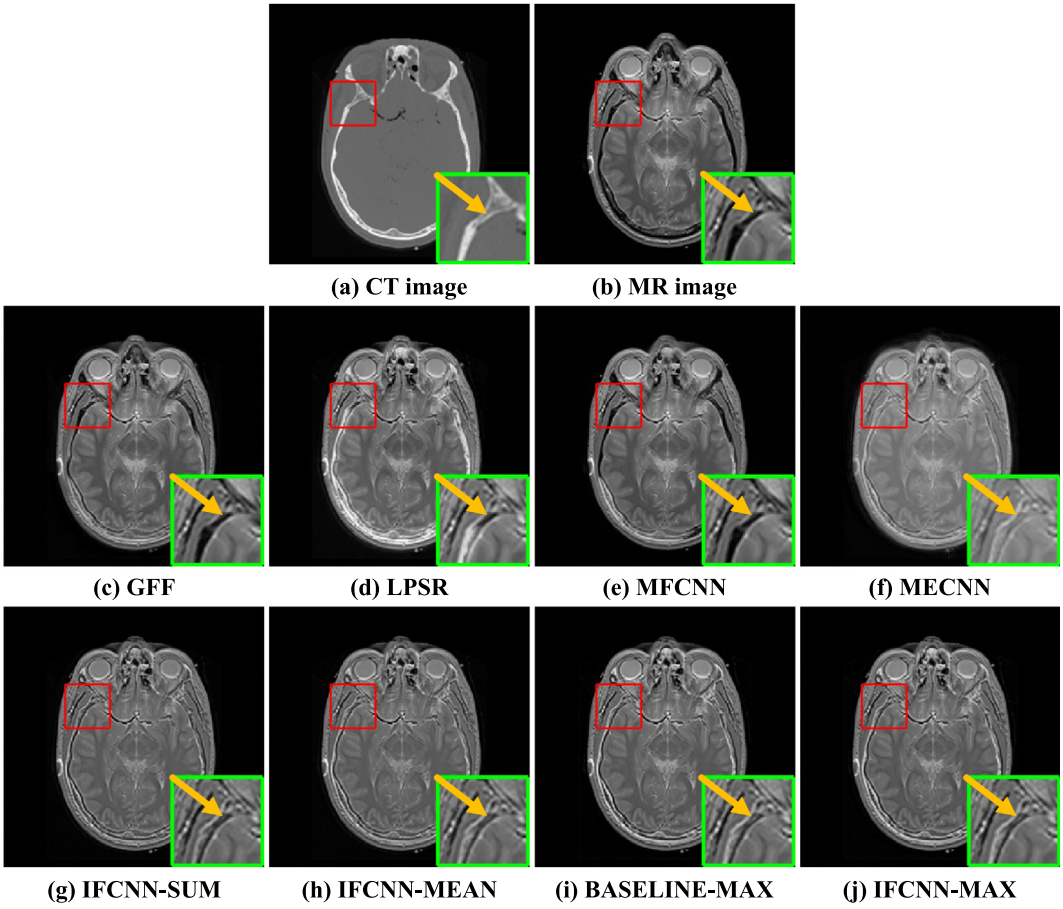

Fig. 14. The comparison example on the second pair of multi-modal medical images. (a) and (b) are respectively the CT image and MR image, and (c)-(j) are the fusion results of (a) and (b) respectively by GFF, LPSR, MFCNN, MECNN, IFCNN-SUM, IFCNN-MEAN, BASELINE-MAX and IFCNN-MAX. In each subfigure, the image patch bounded by green box indicates the closeup of the image patch bounded by red box. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 3
Quantitative evaluation results on medical image dataset.

Metrics	GFF	LPSR	MFCNN	MECNN	IFCNN-SUM	IFCNN-MEAN	BASELINE-MAX	IFCNN-MAX
VIFF	0.6714(5,0,1)	*0.7708(2,3,3)*	**0.7919(1,4,1)**	0.5223(8,0,1)	0.6147(7,0,0)	0.6200(6,0,0)	0.6911(3,1,2)	0.6799(4,0,0)
ISSIM	0.4150(7,0,1)	0.4641(6,1,1)	**0.5365(1,4,0)**	0.2975(8,1,0)	0.4792(3,0,2)	*0.4877(2,0,3)*	0.4702(5,1,1)	0.4739(4,1,0)
NMI	0.6063(7,0,1)	*0.6883(2,1,2)*	**0.9144(1,5,0)**	0.5826(8,1,0)	0.6207(3,0,0)	0.6202(4,0,3)	0.6123(6,1,2)	0.6144(5,0,0)
SF	21.73(7,0,0)	24.13(3,1,1)	22.40(6,0,0)	14.69(8,0,0)	23.05(4,0,0)	22.78(5,0,0)	*24.86(2,3,4)*	**24.98(1,4,3)**
AG	2.892(7,0,0)	3.312(3,2,0)	2.898(6,0,0)	2.123(8,0,0)	3.164(4,0,0)	3.130(5,0,0)	*3.340(2,3,4)*	**3.402(1,3,4)**

Table 4
Quantitative evaluation results on multi-exposure dataset.

Metrics	GFF	LPSR	MFCNN	MECNN	IFCNN-SUM	IFCNN-MAX	BASELINE-MAX	IFCNN-MEAN
MESSIM	**0.9204(1,6,0)**	0.8043(5,0,0)	0.8041(6,0,0)	0.5947(8,0,1)	0.6396(7,0,0)	0.8151(4,0,0)	0.8713(3,0,1)	*0.8786(2,0,4)*
SF	25.93(7,0,0)	25.98(6,0,2)	26.21(5,0,0)	17.16(8,0,0)	*32.48(2,2,2)*	**38.76(1,4,2)**	29.49(4,0,0)	30.6458(3,0,0)
AG	3.645(5,0,0)	3.495(6,0,0)	3.494(7,0,0)	1.896(8,0,0)	3.815(4,1,2)	**6.224(1,5,1)**	4.623(3,0,2)	*4.701(2,0,1)*

the human eyes to grasp scene details. Therefore, our chief model IFCNN-MEAN shows the best visual effect on this comparison example.

Afterwards, the quantitative evaluation is performed on the multi-exposure image dataset according to the experimental settings, and the evaluation results are listed in Table 4. Table 4 shows that GFF gets the largest MESSIM value, which indicates that GFF has preserved the most structural information from input images. While our chief model IFCNN-MEAN ranks second in the MESSIM metric, and LPSR, IFCNN-MAX, MFCNN, MECNN and IFCNN-SUM have obtained relatively lower values. As for the SF and AG metrics, the quantitative results show that our proposed four models rank the top four places. Specifically, IFCNN-MAX ranks first in both the SF and AG metrics and IFCNN-MEAN obtains the third place and second place respectively in the SF metric and AG metric, which means IFCNN-MAX and IFCNN-MEAN have produced fusion images with more textural information than other algorithms. Compared to GFF, our chief model (IFCNN-MEAN) could not only preserve comparable amount of structural information, but also retain more textural information from input images. Overall, the qualitative and quantitative evaluation results on the multi-exposure image dataset imply that IFCNN-MEAN could integrate the visible features of suitable

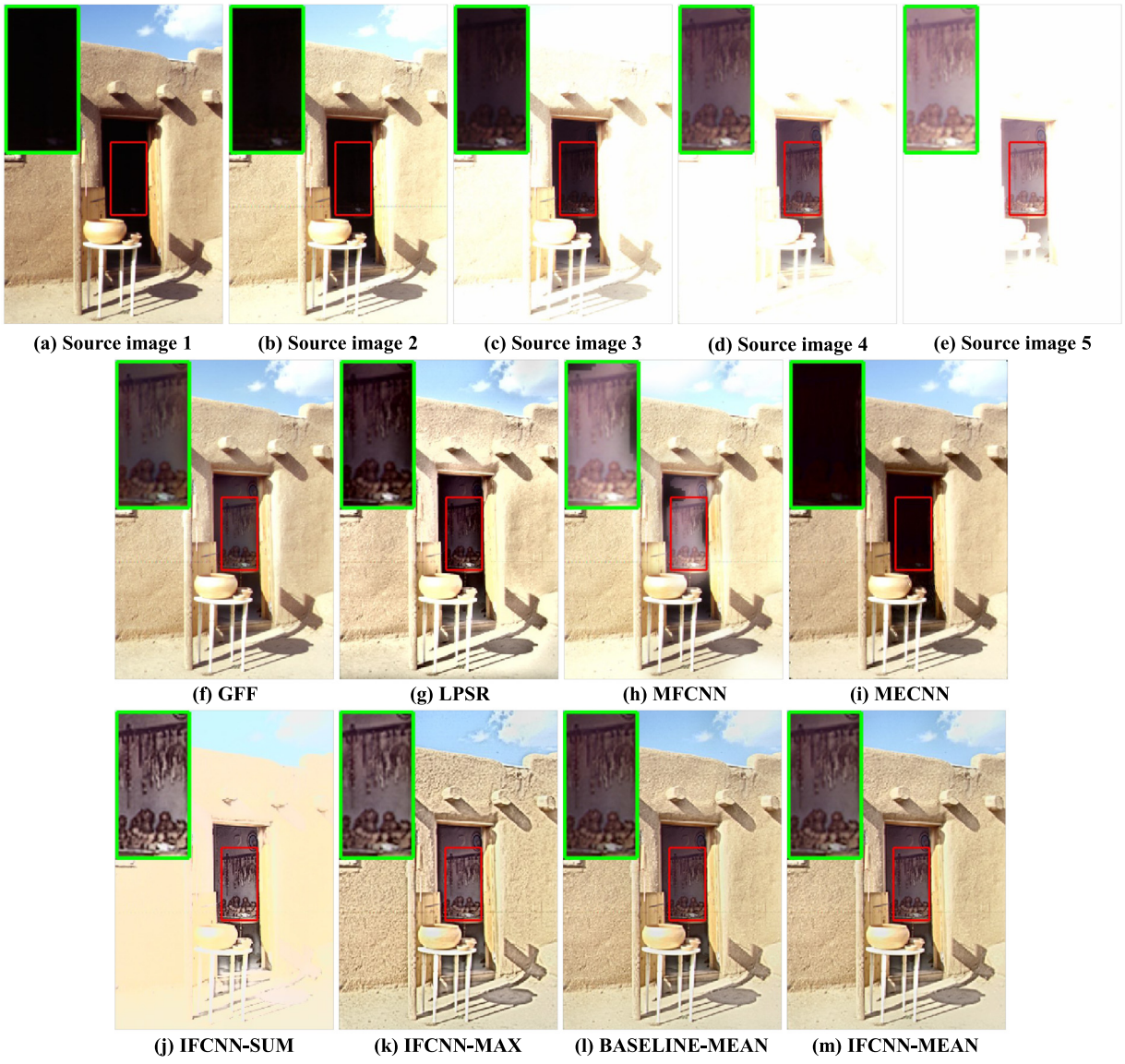

Fig. 15. The comparison example on the fifth set of the multi-exposure images. (a)–(e) are respectively the source images of different exposure degrees, and (f)–(m) are the fusion results of (a)–(e) respectively by GFF, LPSR, MFCNN, MECNN, IFCNN-SUM, IFCNN-MEAN, BASELINE-MAX and IFCNN-MAX. In each subfigure, the image patch bounded by green box indicates the closeup of the image patch bounded by red box. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

exposure-degree from input images into the fusion image, and could perform comparably or even better than other algorithms.

3.6. Time cost comparison

As introduced in Section 2.5, our algorithm takes about 0.02 s to produce one fusion image of size 520 × 520 from two input images on the platform with Intel Core i7-3770k CPU and NVIDIA TITAN X GPU. Since we have only tested the CPU version of MFCNN and MECNN, thus the time costs of MFCNN and MECNN are referred to the reports in Liu et al. [31] and Prabhakar et al. [34]. MFCNN (the slight model) takes about 0.33 s to fuse two input images of size 520 × 520 on the platform with Intel Core i7-4790k CPU and NVIDIA TITAN Black GPU. MECNN takes about 0.07 s to fuse two input images of size 512 × 384 on the platform with Intel Xeon @3.5 GHz CPU and NVIDIA Tesla K20c GPU. At last, GFF and LSPR respectively cost 0.33 s and 0.20 s to fuse two input images of size 520 × 520 on our platform with Intel Core i7-3770k CPU.

Table 5
Time cost comparison (Size unit: pixel, Time unit: second).

Algorithms	GFF	LPSR	MFCNN	MECNN	IFCNN
Image Size	520 × 520	520 × 520	520 × 520	512 × 384	520 × 520
Time Cost	0.33	0.20	0.33	*0.07*	**0.02**

In pursuit of the clear comparison, time costs of the compared algorithms are listed in Table 5, in which the shortest and second-shortest time costs are respectively highlighted as red and blue. Even though the evaluation platforms of IFCNN, MFCNN and MECNN are different, the impact of this difference on running times should be not that significant. Thus, this comparison could still reflect that our algorithm is faster than the current CNN models or even the classical transform domain algorithms. Moreover, our IFCNN only occupies about 785 MB GPU memory for fusing two input images of size 520 × 520. Therefore, it is very

convenient to deploy our image fusion model into the real-time surveillance systems without consuming too much computational resources.

3.7. Conclusions on experimental results

According to both qualitative and quantitative evaluation results on four types of image datasets, we can arrive at the following five conclusions:

- Our proposed chief models could achieve comparable or even better performance than the state-of-the-art image fusion algorithms.
- IFCNN-MAX outperforms IFCNN-SUM and IFCNN-MEAN on three types of image dataset (multi-focus, infrared-visual and multi-modal medical image datasets), thus IFCNN-MAX demonstrates better generalization ability than IFCNN-MEAN and IFCNN-SUM. Overall, the experiments verify that our chief models (IFCNN-MAX and IFCNN-MEAN) own better generalization ability than the existing models through comparing the evaluation results on all the four image datasets.
- Due to the wide range of exposure degrees of the multi-exposure images, IFCNN-MEAN is more suitable to fuse the multi-exposure images than IFCNN-MAX.
- Almost all results indicate that our chief models (IFCNN-MAX and IFCNN-MEAN) outperform baseline models (BASELINE-MAX and BASELINE-MEAN), which implies the perceptual loss can boost the image fusion models to produce more informative fusion images.
- The proposed models are light-weight and efficient, thus our models could be conveniently deployed in the real-time surveillance systems.

4. Conclusions

In this paper, we have proposed a general image fusion framework based on the convolutional neural network, which mainly has four advantages over the existing image fusion models: (1) Our model is fully convolutional and thus can be trained in the end-to-end manner without any post-processing procedures. (2) To finely train our model, we have reasonably generated a large-scale multi-focus image dataset by rendering the partially-focused images varied with random depth range from the RGB and depth images of NUY-D2 dataset. Moreover, rather than no ground truth or using focus maps as ground truth in the existing datasets, the source RGB images of NYU-D2 dataset naturally become the ground-truth fusion images of our multi-focus image dataset, which is of great importance for optimizing the essentially regressed image fusion models. (3) As our model is constructed similarly to the structure of the transform-domain image fusion algorithm, thus our model generally owns better generalization ability for fusing various types of images without any finetuning procedures than the existing image fusion models. (4) Owing to the existence of ground-truth fusion images, it is the first time to introduce perceptual loss to optimize the image fusion models, which can boost models to produce fusion images with more textural details. Without finetuning the image fusion models on other image datasets, the extensive experimental results on four types of image datasets validate that the proposed model demonstrates better generalization ability for fusing various types of images than the existing models, and achieves comparable or even better fusion images than the state-of-the-art image fusion algorithms.

This work sets pioneer foundation for the applications of the convolutional neural network in the field of image fusion. However, even though the extensive experimental results have validated the proposed model's advantages, there are still several points that should be further addressed in order to obtain image fusion models with better performance. Firstly, our multi-focus image dataset only contains indoor images, thus extending the dataset with the outdoor images such as KITTI dataset [54] could probably increase the model's performance. Secondly, the proposed model only consists of four convolutional layers, therefore using deeper convolutional neural network has great potential to further improve the model's performance. Thirdly, the proposed model is designed to fuse the registered images, thus adding an image alignment module might enable the image fusion model to deal with the unregistered cases. Fourthly, in this paper, we have only utilized linear elementwise fusion rules to fuse the convolutional features of multiple input images, thus incorporating more complex and powerful feature fusion module can also boost the model's performance. Finally, our proposed model is designed as a general image fusion framework, thus its performance might be limited for fusing a specific type of images. Therefore, one practical way, to improve performance of the CNN based image fusion models, is to design the architecture according to the specific characteristics of the target image dataset.

Acknowledgments

The authors are grateful to the anonymous reviewers and editors for their valuable comments on improving this paper's quality. The authors also would like to thank Bo Wang for sharing the infrared-visual image dataset, and thank K. Ram Prabhakar for providing their source code. This work is partly supported by the National Natural Science Foundation of China under Grant 61871248, Grant 61701160, Grant 61503405 and Grant U1533132.

References

[1] H. Li, B. Manjunath, S.K. Mitra, Multisensor image fusion using the wavelet transform, Graphical Models Image Process. 57 (3) (1995) 235–245.
[2] A.A. Goshtasby, S. Nikolov, Image fusion: advances in the state of the art, Inf. Fusion 8 (2) (2007) 114–118.
[3] X. Bai, F. Zhou, B. Xue, Fusion of infrared and visual images through region extraction by using multi scale center-surround top-hat transform, Opt. Express 19 (9) (2011) 8444–8457.
[4] S. Li, X. Kang, J. Hu, Image fusion with guided filtering, IEEE Trans. Image Process. 22 (7) (2013) 2864–2875.
[5] A.A. Goshtasby, Fusion of multi-exposure images, Image Vision Comput. 23 (6) (2005) 611–618.
[6] R. Shen, I. Cheng, J. Shi, A. Basu, Generalized random walks for fusion of multi-exposure images, IEEE Trans. Image Process. 20 (12) (2011) 3634–3646.
[7] K. Ma, K. Zeng, Z. Wang, Perceptual quality assessment for multi-exposure image fusion, IEEE Trans. Image Process. 24 (11) (2015) 3345–3356.
[8] A. Saha, G. Bhatnagar, Q.J. Wu, Mutual spectral residual approach for multifocus image fusion, Digital Signal Process. 23 (4) (2013) 1121–1135.
[9] X. Bai, Y. Zhang, F. Zhou, B. Xue, Quadtree-based multi-focus image fusion using a weighted focus-measure, Inf. Fusion 22 (2015) 105–118.
[10] Q. Zhang, M.D. Levine, Robust multi-mocus image fusion using multi-task sparse representation and spatial context, IEEE Trans. Image Process. 25 (5) (2016) 2045–2058.
[11] G. Bhatnagar, Q.M.J. Wu, Z. Liu, Directive contrast based multimodal medical image fusion in nsct domain, IEEE Trans. Multimedia 15 (5) (2013) 1014–1024.
[12] Z. Xu, Medical image fusion using multi-level local extrema, Inf. Fusion 19 (2014) 38–48, doi:10.1016/j.inffus.2013.01.001.
[13] Z. Xue, R.S. Blum, Concealed weapon detection using color image fusion, in: Proceedings of the 6th International Conference on Information Fusion, Vol. 1, IEEE, 2003, pp. 622–627.
[14] T. Wan, N. Canagarajah, A. Achim, Segmentation-driven image fusion based on alpha-stable modeling of wavelet coefficients, IEEE Trans. Multimedia 11 (4) (2009) 624–633.
[15] Z. Zhou, B. Wang, S. Li, M. Dong, Perceptual fusion of infrared and visible images through a hybrid multi-scale decomposition with Gaussian and bilateral filters, Inf. Fusion 30 (2016) 15–26.
[16] Y. Zhang, L. Zhang, X. Bai, L. Zhang, Infrared and visual image fusion through infrared feature extraction and visual information preservation, Infrared Phys. Technol. 83 (2017) 227–237.
[17] Y. Zhang, X. Bai, T. Wang, Boundary finding based multi-focus image fusion through multi-scale morphological focus-measure, Inf. Fusion 35 (2017) 81–101.
[18] W. Huang, Z. Jing, Evaluation of focus measures in multi-focus image fusion, Pattern Recognit. Lett. 28 (4) (2007) 493–500.
[19] Z. Zhou, S. Li, B. Wang, Multi-scale weighted gradient-based fusion for multi-focus images, Inf. Fusion 20 (2014) 60–72.
[20] P.J. Burt, E.H. Adelson, The laplacian pyramid as a compact image code, IEEE Trans. Commun. 31 (4) (1983) 532–540.
[21] A. Toet, Image fusion by a ratio of low-pass pyramid, Pattern Recognit. Lett. 9 (4) (1989) 245–253.

[22] J.J. Lewis, R.J. O'Callaghan, S.G. Nikolov, D.R. Bull, N. Canagarajah, Pixel- and region-based image fusion with complex wavelets, Inf. Fusion 8 (2 SPEC. ISS.) (2007) 119–130.
[23] F. Nencini, A. Garzelli, S. Baronti, L. Alparone, Remote sensing image fusion using the curvelet transform, Inf. Fusion 8 (2 SPEC. ISS.) (2007) 143–156.
[24] Y. Liu, S. Liu, Z. Wang, A general framework for image fusion based on multi-scale transform and sparse representation, Inf. Fusion 24 (2015) 147–164.
[25] X. Bai, Infrared and visual image fusion through feature extraction by morphological sequential toggle operator, Infrared Phys. Technol. 71 (2015) 77–86.
[26] B. Yang, S. Li, Multifocus image fusion and restoration with sparse representation, IEEE Trans. Instrumen.Meas. 59 (4) (2010) 884–892.
[27] N. Yu, T. Qiu, F. Bi, A. Wang, Image features extraction and fusion based on joint sparse representation, IEEE J. Sel. Top. Signal Process. 5 (5) (2011) 1074–1082.
[28] S. Li, H. Yin, L. Fang, Group-sparse representation with dictionary learning for medical image denoising and fusion, IEEE Trans. Biomed. Eng. 59 (12) (2012) 3450–3459.
[29] B. Yang, S. Li, Pixel-level image fusion with simultaneous orthogonal matching pursuit, Inf. Fusion 13 (1) (2012) 10–19.
[30] Y. Liu, X. Chen, Z. Wang, Z.J. Wang, R.K. Ward, X. Wang, Deep learning for pixel-level image fusion: recent advances and future prospects, Inf. Fusion 42 (2018) 158–173.
[31] Y. Liu, X. Chen, H. Peng, Z. Wang, Multi-focus image fusion with a deep convolutional neural network, Inf. Fusion 36 (2017) 191–207.
[32] H. Tang, B. Xiao, W. Li, G. Wang, Pixel convolutional neural network for multi-focus image fusion, Inf. Sci. (2017).
[33] H. Song, Q. Liu, G. Wang, R. Hang, B. Huang, Spatiotemporal satellite image fusion using deep convolutional neural networks, IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 11 (3) (2018) 821–829.
[34] K.R. Prabhakar, V.S. Srikar, R.V. Babu, Deepfuse: a deep unsupervised approach for exposure fusion with extreme exposure image pairs, in: Proceedings of the IEEE International Conference on Computer Vision, IEEE, 2017, pp. 4724–4732.
[35] H. Yan, X. Yu, Y. Zhang, S. Zhang, X. Zhao, L. Zhang, Single image depth estimation with normal guided scale invariant deep convolutional fields, IEEE Trans. Circuits Syst. Video Technol. (2017).
[36] L. Li, S. Zhang, X. Yu, L. Zhang, Pmsc: patchmatch-based superpixel cut for accurate stereo matching, IEEE Trans. Circuits Syst. Video Technol. (2016).
[37] N. Silberman, D. Hoiem, P. Kohli, R. Fergus, Indoor segmentation and support inference from rgbd images, in: European Conference on Computer Vision, Springer, 2012, pp. 746–760.
[38] H. Zhao, J. Shi, X. Qi, X. Wang, J. Jia, Pyramid scene parsing network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2881–2890.
[39] V. Nair, G.E. Hinton, Rectified linear units improve restricted boltzmann machines, in: Proceedings of the 27th International Conference on Machine Learning, Omnipress, 2010, pp. 807–814.
[40] S. Ioffe, C. Szegedy, Batch normalization: accelerating deep network training by reducing internal covariate shift, in: Proceedings of the 32nd International Conference on Machine Learning, JMLR.org, 2015, pp. 448–456.
[41] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Trans. Image Process. 13 (4) (2004) 600–612.
[42] J. Johnson, A. Alahi, L. Fei-Fei, Perceptual losses for real-time style transfer and super-resolution, in: European Conference on Computer Vision, Springer, 2016, pp. 694–711.
[43] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, et al., Photo-realistic single image super-resolution using a generative adversarial network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2017, pp. 4681–4690.
[44] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv:1409.1556v1(2014).
[45] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2016, pp. 770–778.
[46] M. Nejati, S. Samavi, S. Shirani, Multi-focus image fusion using dictionary-based sparse representation, Inf. Fusion 25 (2015) 72–84.
[47] Y. Liu, X. Chen, J. Cheng, H. Peng, A medical image fusion method based on convolutional neural networks, in: 2017 20th International Conference on Information Fusion, IEEE, 2017, pp. 1–7.
[48] Y. Liu, Z. Wang, Dense sift for ghost-free multi-exposure fusion, J. Visual Commun. Image Represent. 31 (2015) 208–224.
[49] Y. Han, Y. Cai, Y. Cao, X. Xu, A new image fusion performance metric based on visual information fidelity, Inf. Fusion 14 (2) (2013) 127–135.
[50] C. Yang, J.-Q. Zhang, X.-R. Wang, X. Liu, A novel similarity based quality metric for image fusion, Inf. Fusion 9 (2) (2008) 156–160.
[51] M. Hossny, S. Nahavandi, D. Creighton, Comments on 'information measure for performance of image fusion', Electron. Lett. 44 (18) (2008) 1066–1067.
[52] S. Li, B. Yang, Multifocus image fusion using region segmentation and spatial frequency, Image Vision Comput. 26 (7) (2008) 971–979.
[53] W. Zhao, D. Wang, H. Lu, Multi-focus image fusion with a natural enhancement via joint multi-level deeply supervised convolutional neural network, IEEE Trans. Circuits Syst. Video Technol. (2018).
[54] A. Geiger, P. Lenz, C. Stiller, R. Urtasun, Vision meets robotics: the kitti dataset, Int. J. Rob. Res. 32 (11) (2013) 1231–1237.