Optics and Lasers in Engineering 156 (2022) 107078

Contents lists available at ScienceDirect

Optics and Lasers in Engineering

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

Multimodal super-resolution reconstruction of infrared and visible images via deep learning

Bowen Wang[a,b], Yan Zou[c], Linfei Zhang[a,b], Yuhai Li[d], Qian Chen[a,b], Chao Zuo[a,b,*]

[a] Jiangsu Key Laboratory of Spectral Imaging and Intelligent Sense, Nanjing University of Science and Technology, Nanjing, Jiangsu Province 210094, China
[b] Smart Computational Imaging (SCI) Laboratory, Nanjing University of Science and Technology, Nanjing, Jiangsu Province 210094, China
[c] Military Representative Office of army equipment department in Nanjing, Nanjing, Jiangsu Province 210094, China
[d] Science and Technology on Electro-Optical Information Security Control Lab, Tianjin 300000, China

ARTICLE INFO

Keywords:
Super-resolution
Infrared image
Convolutional neural network
Multi-modal imaging
Image fusion

ABSTRACT

In this paper, we propose a deep-learning-based infrared-visible images fusion method based on encoder-decoder architecture. The image fusion task is reformulated as a problem of maintaining the structure and intensity ratio of the infrared-visible image. The corresponding loss function is designed to expand the weight difference between the thermal target and the background. In addition, a single image super-resolution reconstruction based on a regression network is introduced to address the issue that traditional network mapping functions are not suitable for natural scenes. The forward generation and reverse regression models are considered to reduce the irrelevant function mapping space and approach the ideal scene data through double mapping constraints. Compared with other state-of-the-art approaches, our experimental results achieve superior performance in terms of both visual effects and objective assessments. In addition, it can stably provide high-resolution reconstruction results consistent with human visual observation while bridging the resolution gap between the infrared-visible images.

1. Introduction

Image fusion techniques [1–3] aim to generate an informative image with specific algorithms from multiple source images. Thanks to the ability to recombine disparate information, infrared and visible image fusion technology plays a pivotal role in the detecting imaging systems. Hence, the fused result has a more distinct and complete depiction of the scene, which is beneficial to human perception and machine processing. The fusion image can synthesize a novel image with complementary information of the source images. Maximizing the integration of interest information is an essential bottleneck to reveal novel insights and fundamental scientific issues in biomedicine [4], forest fire fighting [5], and safe driving. For example, it is common to generate high dynamic range (HDR) images by applying the multiple exposure fusion (MEF) [6–8] approach. HDR imaging method can provide more prosperous image details, making reconstructed images more distinct and pleasing to human visual observation. Based on this approach, the infrared and visible fusion algorithm [9–11] can integrate the advantages of each information. Generally speaking, infrared images lack texture information and cannot effectively characterize the scene. Notwithstanding, it has been widely applied own to its inherent thermal radiation characteristics and the ability to realize cloud penetration imaging in long-wave infrared bands. In contrast, the visible image contains texture details with high spatial resolution, which is conducive to enhancing the ability of target recognition and conforms to the human visual system. However, the visible image also has a fatal disadvantage: it is impossible to obtain a high-quality image under low illumination conditions. Therefore, visible-infrared imaging is interdependent and jointly promoted.

Although the image fusion technology has made significant improvements, the pixel size of the long-wave infrared detector has approached the physical limit (17μm) due to limitations in software algorithms and hardware technology. Meanwhile, with the imaging resolution increasing, the manufacturing cost of the device will also dramatically expand. Therefore, the current dual-band image fusion technology is insufficient to stably realize all-weather high-resolution imaging. At this time, the traditional super-resolution (SR) models and algorithms are no longer suitable, and their computational complexity adds the pressure of massive calculation to the application. Recently, deep learning (DL) [12,13] has emerged as a powerful technique in the field of image fusion owing to its outstanding feature extraction, representation capability, strong robustness, and efficient reconstruction performance. From the artificial intelligence robot developed by Deepmind company to the powerful robot dog in Boston, promising news came one after another. Artificial intelligence [14–17] produces a familiar word around

* Corresponding author.
E-mail address: zuochao@njust.edu.cn (C. Zuo).

https://doi.org/10.1016/j.optlaseng.2022.107078
Received 27 December 2021; Received in revised form 31 March 2022; Accepted 10 April 2022

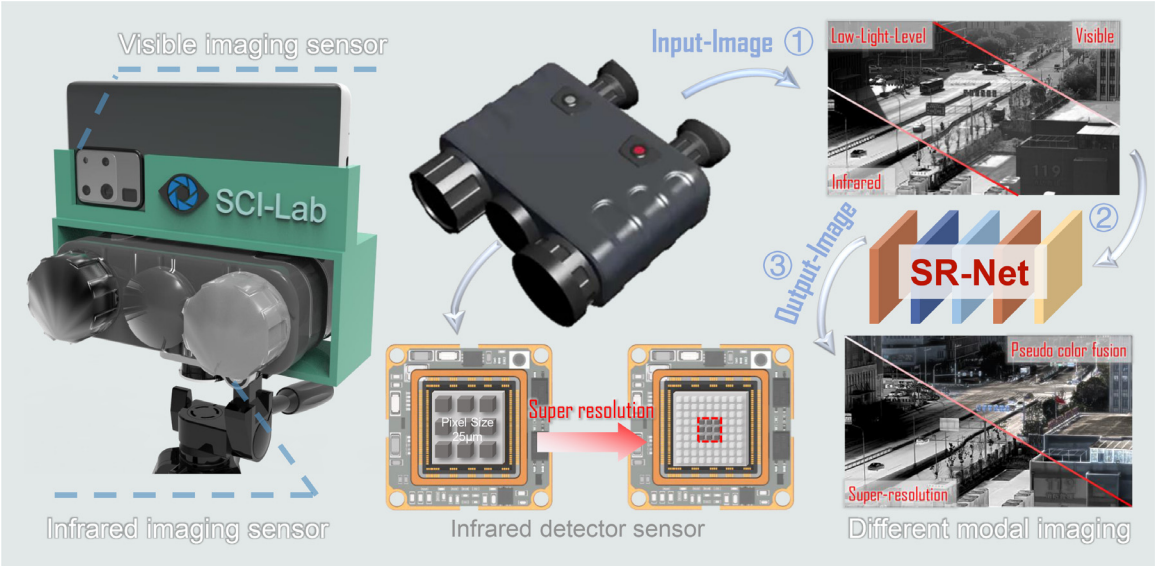

Fig. 1. Structural diagram and imaging reconstruction notion of the cross-modal fusion imaging system.

us. This is a remarkable manifestation of the gradual replacement of manual operation by intelligent machines. This trend is being driven by the increasing demand for the emergence of multi-dimensional sensors coupled with artificial intelligence computing technology. Over the past decades, deep learning technology has become a research hotspot in the era of massive data. Both academia and industry show strong interest to this technology, especially in computer vision [18,19]. As a "Data-Driven" technology that has emerged in recent years, it has achieved surpassing achievement in many applications such as image classification [20], object detection [21,22], and recognition [23,24]. And as shown in Fig. 1, overcoming the pixelation imaging problem caused by inadequate spatial sampling is also the novelty of Multi-image super-resolution fusion (Multi-SR-Fusion) technology.

The remaining structures of this paper are as follows. In Section 2, we briefly review related works on deep learning frameworks. Section 3 depicts the basic principle of our proposed method. Section 4 presents the details of the proposed Multi-SR-Fusion network for infrared and visible image fusion. Abundant experimental results and analysis are illustrated in Section 5. Finally, Section 6 provides a discussion and summarizes the paper.

2. Related works

At present, benefiting from the powerful feature extraction ability of DL convolution operation and learning mapping function parameters from massive data, the DL method has rapidly evolved the most potential direction in the field of image fusion. The traditional single-frame image SR [25,26] problem refers to the process of recovering from low-resolution (LR) images to high-resolution images, constantly pushing the limits to obtain higher real-world perception. In the field of computer vision, the introduction of convolutional neural networks (CNNs) [27] has extensively promoted the development of single image SR technology. The researchers continuously optimize the SR network model by introducing residual models, deep convolutional structures, and dense connectivity structures to enhance the reconstruction performance. However, due to the ill-posedness of the single image SR issue, most existing methods will generate artifacts and even lose the detailed texture under the condition of the sizeable scaling factor. Therefore, it is still a challenge to accurately reconstruct the high-frequency image details. Of the prominent DL-based methods, there are two mainstreams: convolutional neural network (CNN) [28–31] and generative adversarial network (GAN) [32–34]. A majority of representative works have been proposed on this challenging problem.

In ICCV 2017, a classical fusion method, termed as DeepFuse [35], was put forward to tackle the exposure image fusion task. On this basis, Li et al. replaced the convolution network in the previous part with dense-block for improvement [36]. The fusion network is composed of the encoder, fusion layer, and decoder structure. Considering the similarity between the fused features and the original image, Zhang et al. created the proposed method better focused on the effective extraction of image features [37] by the continuous feedback of feature information from each layer. With the rapid development of the GAN network, scholars have also applied it to the field of infrared and visible images. Ma et al. proposed a detail-preserving learning-based fusion model for infrared and visible images [38]. The dual loss functions of detail loss and target edge enhancement loss are designed to improve the quality of detail information and sharpen the edges of IR targets, respectively, in the adversarial network generation framework. Nonetheless, this method does not fully consider the characteristics of infrared and visible images, and the fused images are challenging to highlight the target information. According to the aspects of infrared-visible imaging, Li et al. proposed a GAN network with a multi-scale attention mechanism [39]. The multi-scale attention mechanism generator focuses on the target information of the infrared image and the background detail information of the visible image so that the fusion network can concentrate on the specific area of the source image to reconstruct the fusion image. Generally speaking, the method based on DL can produce satisfactory results without manually designed decomposition processing and fusion rules. However, they can not highlight important targets while retaining background information, resulting in low contrast of fusion results. Due to the limitations of the manufacturing process, power consumption, or the cost of the sensor, the pixel imaging of infrared images has not been sufficiently solved. Zou et al. successfully realized the SR reconstruction of infrared images by employing the encoder-decoder network and also verified the application potential in image SR and feature extraction [40]. Therefore, if the SR structure can be added to the network, the fusion result will be predictable improved.

Gatys et al. proposed the neural style transfer method [41] and first applied the DL method to the style transfer task. The network maintains the consistency of the basic information of the two images through content loss constraints and updates the style of the input image by back-propagation iterations. By continuous forward propagation calculation loss and backpropagation optimization loss and updating the pixel value

of the reconstructed image, the optimal reconstructed image is eventually obtained. The essence of image style migration is the fusion of two different style images. In a sense, infrared and visible images can also be regarded as two separate ”style” images. Therefore, this proposed method utilizes the notion of neural style transfer to alleviate the problem of infrared and visible image fusion.

As mentioned above, in recent years, infrared and visible image fusion technology based on the neural network has essential research prospects. In the task of infrared and visible image fusion, the following problems are still faced:

(1) End-to-end imaging datasets. DL reconstruction algorithms are based on multiple datasets, while fewer datasets are available for infrared and visible image fusion tasks. How to utilize the existing data to realize the network training model is one of the challenges. And the most critical point is that the current fusion networks do not consider the resolution of infrared images, and the quality of input infrared images is too poor, resulting in unsatisfactory reconstruction results.
(2) The resolution gap between the infrared-visible images. In the task of infrared-visible fusion, generally speaking, the resolution of the infrared detector will generally be much worse than the visible detector. Therefore, whether the infrared imaging quality can be improved through the mapping function to enhance the quality of fusion image is also one of the critical contents of this paper.
(3) Network structure. Image fusion is a low-level task in computer vision, and the network structure should be as lightweight as possible. And how to give full play to the network ability and trade-off the weight between two images is also one of the fundamental problems.
(4) Loss function. In the network training process, the network training parameter needs to be modified by the loss function, which puts forward more strict requirements for the loss function design.

3. Proposed methods

For the human visual system, the ”conspicuity area” that containing essential targets is more attractive. Based on the above analysis, the problem of infrared–visible image fusion is how to maintain the high-frequency detail information and the thermal radiation information so as to realize a multi-dimensional data fusion process. The primary task of the proposed method is to improve the resolution of the infrared image and then carry out the weighted fusion of the heterologous image while obtaining a high-quality image resolution. Therefore, efficiently extracting the feature information of each image and assigning fusion weight is the focus of our research. Based on the concept of U-net semantic segmentation and style transfer [42], the thermal radiation information of the infrared image can be effectively segmented, and then the thermal image and visible texture information are transferred by style transfer structure. In our workflow, the coding-decoding fusion structure is employed for end-to-end learning, as shown in Fig. 2, so that the network can not only center on the ”conspicuity area” information but also learn the image SR mapping function. The image merge problem is transformed into the issue of maintaining the structure and intensity ratio of infrared and visible images. The corresponding loss function is designed to expand the weight distinction between the thermal target and the background. Aiming at the shortage that the traditional network mapping function is ill-posed in the actual scene, the additional constraint of inverse regression is embedded to reduce the space of the possible mapping function. Lastly, the pseudo color SR reconstruction based on the scene is realized by expanding the number of channels. By doing so, the reconstructed image is more in line with the human visual effect.

Note that our method takes the infrared image and visible image as the input image and obtains the colorized fusion image through end-to-end supervised network. Multi-scale feature extraction is performed in infrared and visible images by applying the diverse dimensions kernels. Subsequently, the infrared and visible fusion image is generated through the fusion layer. The fusion structure contains multi-scale feature extraction and residual channel attention blocks (RCAB), which enables valuable feature mapping and suppresses unimportant feature mapping. The coding-decoding SR structure realizes the functions of feature extraction and reconstruction, respectively. Meanwhile, the introduction of the skip connection structure can transfer the image feature information from the encoding part to the decoding part of the network, solving the problem of gradient disappearance.

Table 1
The number of layers in the network structure.

Layer	Parameter	Numbers
Convolution layer	1 × 1, Strides = 1, padding = SAME	4
Convolution layer	3 × 3, Strides = 1, padding = SAME	21
Convolution layer	3 × 3, Strides = 2, padding = SAME	4
Convolution layer	5 × 5, Strides = 1, padding = SAME	4
ReLU layer	-	12
LReLU layer	alpha = 0.2	16
Concat layer	-	6
Deconvolution layer	3 × 3	4
Element-max later	-	1
Global average pooling layer	-	4
Fully connected layer	-	8
Sigmod layer	-	4
Pixelshuffle layer	2 × 2	2
Max-pooling layer	2 × 2	4

3.1. Problem formulation

To express the mapping relationship of the network more clear, the network model can be defined as:

$$I_{out}(x,y) = F_{\omega,\theta}\left[I_{LR1}(x,y), I_{LR2}(x,y)\right] \tag{1}$$

where, $F_{w,\theta}[.]$ represents the nonlinear mapping function of the network, ω and θ respectively describe the weight and deviation trainable parameters in the network, $I_{LR1}(x,y)$ describes the input long-wave infrared image, $I_{LR2}(x,y)$ describes the input visible image, and $I_{\text{out}}(x,y)$ is the HR image output by the network. Detailed network parameters are illustrated in Table 1.

The network structures contain the convolution, deconvolution, element-addition or multiplication, channel-fusion, max-pooling, and element-max layers. The input image of the X_i layer is represented by i, and the convolution layer and deconvolution layer are represented as:

$$F(X_i) = \max\left(0, W_k * X_i + B_k\right) \tag{2}$$

where, W_k and B_k represent filter and deviation respectively. For convenience, $*$ represents convolution or deconvolution.

For the element-addition layer, the output is the addition of two inputs of the same size, followed by Leaky Rectified Linear Unit(LReLU) activation:

$$F(X_i, X_j) = \begin{cases} X_i + X_j, X_i + X_j \geq 0 \\ \alpha * (X_i + X_j), X_i + X_j < 0 \end{cases} \tag{3}$$

where, X_i and X_j represent layer $i+1$ and layer $j+1$ respectively, and $\alpha = 0.01$.

For the element multiplication layer, the output is the multiplication of two elements of the same size, followed by LReLU activation:

$$F(X_i, X_j) = \begin{cases} X_i \cdot X_j, X_i \cdot X_j \geq 0 \\ \alpha * (X_i \cdot X_j), X_i \cdot X_j < 0 \end{cases} \tag{4}$$

For the channel fusion layer, the output is the sum of two input channels of the same size:

$$F(X_i, X_j) = X_i \oplus X_j \tag{5}$$

For the max-pooling layer, the output image size is half of the input image, which is expressed by the following formula:

$$F(X_i) = down(X_i) \tag{6}$$

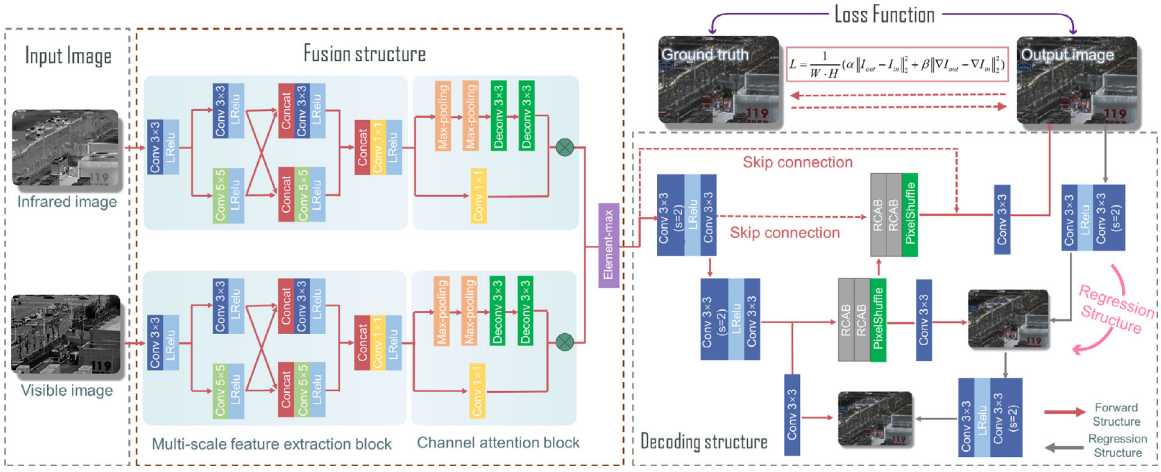

Fig. 2. Super-resolution fusion network structure of heterogeneous images based on encoding-decoding structure.

where $down$ represents pooling function, and this paper adopts max-pooling.

For the element-max layer, the size of the output image is the same as the input image, which is expressed by the following formula:

$$F(X_i, X_j) = \max(X_i, X_j) \tag{7}$$

For the sub-pixel convolution layer, the output image size is twice of the input image, which is expressed by the following formula:

$$F(X_i) = pixelshuffle(X_i) \tag{8}$$

3.2. Loss function

Weight distribution is the core problem of image fusion, which directly determines the quality of fused image. To perform network training, we need to accurately evaluate the information similarity between the fused image and the input image pair to minimize information loss, thus effectively preserving the thermal radiation information from the infrared image and the textural detail information from the visible image. Therefore, in this paper, the image fusion problem is transformed into the issue of maintaining the structure and intensity ratio of infrared-visible images. The intensity distribution and gradient information can characterize the thermal radiation and structural information, respectively. In order to preserve the representative features of the source image to the greatest extent, a hybrid loss function is designed to retain valuable feature information. Thus, the loss function of our proposed model is set to:

$$Loss = \sum_{i=1}^{N} Loss_1(F(x_i), y_i) + \lambda Loss_2(D(y_i), x_i) \tag{9}$$

where x_i and y_i respectively represent the input LR and output HR images. $Loss_1(F(x_i), y_i)$ and $Loss_2(D(y_i), x_i)$ describe the loss functions of forward regression and inverse regression tasks, respectively. During the training process, the reconstructed images $F(x_i)$ continuously converge to the corresponding HR images. Similarly, the similarity between the predicted image $D(y_i)$ and the forward input LR image is continuously approached in the regression process. Here we set λ to 0.1 for the weight distribution of the hybrid loss function.

If $F(x_i)$ is the accurate HR image, the image $D(y_i)$ in the inverse regression model should be dramatically similar to the LR image. With this constraint, we can reduce the possible mapping function so as to achieve robust image reconstruction.

$$Loss_1 = \alpha \left\| y_i - F(x_i) \right\|_2^2 + \beta \left\| \nabla y_i - \nabla F(x_i) \right\|_2^2 \tag{10}$$

where, $\| \bullet \|_2$ defines the L_2 norm, ∇ represents the gradient operator. α and β are two factors that balance these two terms, $\alpha = \beta = 0.5$ in this experiment.

$$Loss_2 = \left\| D(y_i) - x_i \right\|_2^2 \tag{11}$$

This formulation is an improved fusion method by taking SR into account. The forward generation process and reverse regression process of the input-output image are simultaneously constrained, and the dual-loss functions compensate each other to produce the whole loss function balance. The mixed loss between the input-out images is computed to update network parameters. By minimizing the loss, the network performs accurate reconstruction of the input data in the training phase, emphasizes the valuable information, and suppresses the irrelevant information.

4. Network architecture

4.1. Multi-scale feature extraction (encoding) module

An essential part of SR reconstruction is how to extract the features of the input image. Suppose the different dimensions information can be obtained. In that case, it will conduce for signal restoration. On the other hand, the image feature information is generally extracted by a convolution kernel. Therefore, the idea of extracting the image with large convolution to obtain a more extensive receptive field has been sprouting. A larger receptive field will facilitate the reception of feature information. However, if the convolution kernel is too large, the amount of calculation will increase sharply, which is not conducive to the boost of model depth. Therefore, we can decompose the large-scale convolution into several small-scale convolutions so as to reduce the amount of calculation. Although multi-scale convolutional blocks can extract adequate features, it is also crucial to selectively focus on the essential elements and ignore the less important ones. This means that not all features are beneficial for reconstruction. Intermediate features contain valuable information, such as primary structure and details, or even irrelevant information, such as noise. Therefore, We adopt a multiscale layer with different kernel sizes, such as 3×3 and 5×5, to acquire low-frequency and high-frequency features with various receptive fields. By doing so, comprehensive image information at different scales is fetched and reused with each other. The feature fusion convolution layer virtually reduces the computational complexity and improves the convergence speed of the network. Consequently, introducing a multi-scale extraction module is profitable to obtain higher-level robust semantic features, retain more underlying details, and enrich the image feature information.

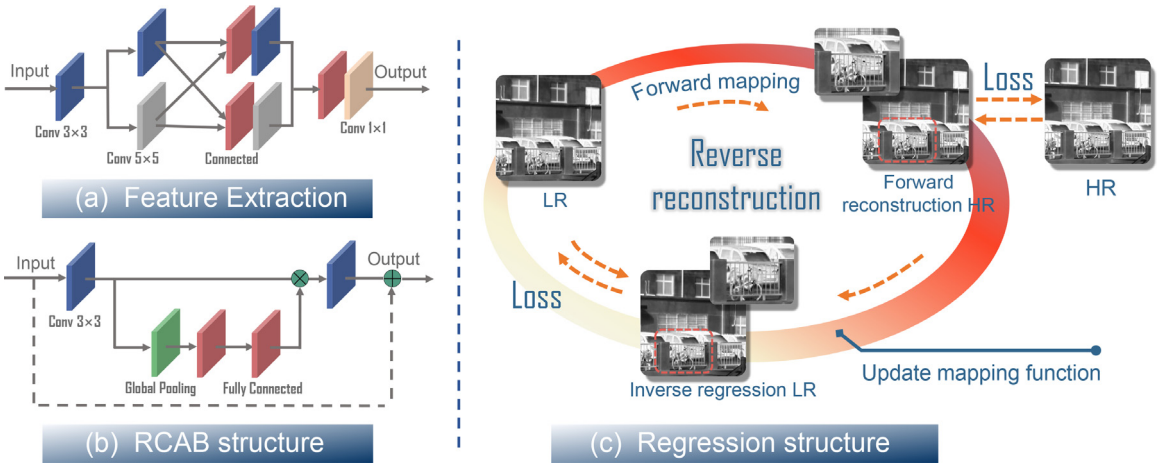

Fig. 3. Schematic diagram of the critical network modules. (a) Multi-scale feature extraction structure. (b) Residual channel attention blocks. (c) Dual-regression mapping structure.

4.2. Super-resolution (decoding) module

The SR network adopts an encoder-decoder architecture. In the decoding layer, the pixel-shuffle method is operated to enlarge the feature map size corresponding to the convolution layer in the coding layer, and the different dimensional information is transmitted by skip connection. Skip connection can not only transfer image feature information but also alleviate the problem of gradient disappearance. We introduce the residual channel attention module to adjust the channel feature information, which is conducive to reconstructing HR images. The global average pooling layer encodes all spatial features into a whole feature on one channel. After receiving the global features, the nonlinear relationship between each channel is learned through the full connection layer. The whole operation can be regarded as learning the weight coefficients of each channel to make the model more discriminative about the features of each channel.

Currently, the mainstream network architecture model is moving in a deeper direction. A deeper network model means that it has better nonlinear expression ability. Thereby, it can learn more complex transformations and fit more complex feature inputs. However, a common accompaniment problem is that the information extracted by the middle layers is not employed thoughtfully. Therefore, the skip connection in the residual structure is worthwhile to enhance the gradient propagation and alleviate the problem of gradient disappearance caused by network deepening. In addition, the existing methods only focus on the mapping from the LR image to the HR image. However, the under-determined possible mapping space is volatile and challenging during the training process. In order to ameliorate this problem, we propose a dual regression project in the SR structure, as shown in Fig. 3(c). Through the restriction of double constraints, the robustness of the network model and its applicability to natural scenes can be promoted.

5. Experiment and results

5.1. Network and dataset settings

In the network, the batch size is 4, and the epoch is set to 200. Empirically, we use Adam optimizer to optimize the network structure, and the initial learning rate is set to 10^{-4}. The network is conducted on hardware platform with an Intel Core ^{TM} i7-9700K CPU @ 3.60GHz×8, and RTX2080Ti. The software platform is running under Ubuntu 16.04 operating system. The total training time of our network is 11.20 hours, and the average test time for each image is 1.31 seconds.

The long-wave infrared (self-developed, 800×600, 25μm) and visible images are collected by the cross-modal image acquisition equipment and transmitted to the network for training after registration. The corresponding images are cut into 128×128 pieces and sent to the network for training. The infrared dataset contains 1300 images, of which 800 images are employed as the training set, and 200 images are utilized as the validation sets. The fusion dataset includes the lake, jungle, and urban imaging environments (https://figshare.com/s/0d35b35c18c70cd3bba1).

It is worth noting that the HR infrared images are acquired at long focal lengths, and conversely, the LR infrared images are obtained at short focal lengths (large field-of-view imaging). The pixel mapping of the HR image is yielded by partially recording the central region of the LR image, as shown in Fig. 1. Instead of creating the training dataset through simulations (bicubic down-sampling or an approximate model of the point spread function), in the presented technique, the desired target 3× super-resolved images are accordingly obtained by tripling the focal length (25mm-75mm).

We utilized the visual saliency map (VSM) and weighted least square (WLS) to realize heterogeneous image fusion. The original image can be decomposed into the bottom and several detail layers by multi-scale decomposition (MSD). The bottom layer mainly contains low-frequency information, which determines the overall appearance and the fused image contrast. In this paper, VSM is used to merge the bottom layer to effectively extract the salient structure so as to avoid the blurring of low-frequency information. Detail layers are merged according to the traditional ”maximization” rule. The absolute value of the detail layer coefficient is considerable, which corresponds to more significant features.

The monochrome display has constantly perplexed low-light-level night vision and infrared imaging systems. Therefore, it is also an essential task to employ the visible color component information to achieve pseudo-color of fused images. We map the RGB color components to the HSV color space in the color migration task. The grayscale fused image is created as the V component of the predicted image, and the chromaticity H and saturation S are kept constant to achieve the final color image output.

5.2. Experimental results analysis

A majority of visible texture information plays a significant role in restoring and reconstructing HR color fusion images. However, in the night vision imaging environment, the visible detector can not provide enough detailed information, so improving the SR reconstruction ability of the infrared image is also an essential research direction. In order to verify this concept, we partially modify the network structure and remove the visible image from the input structure. Fig. 4 depicts

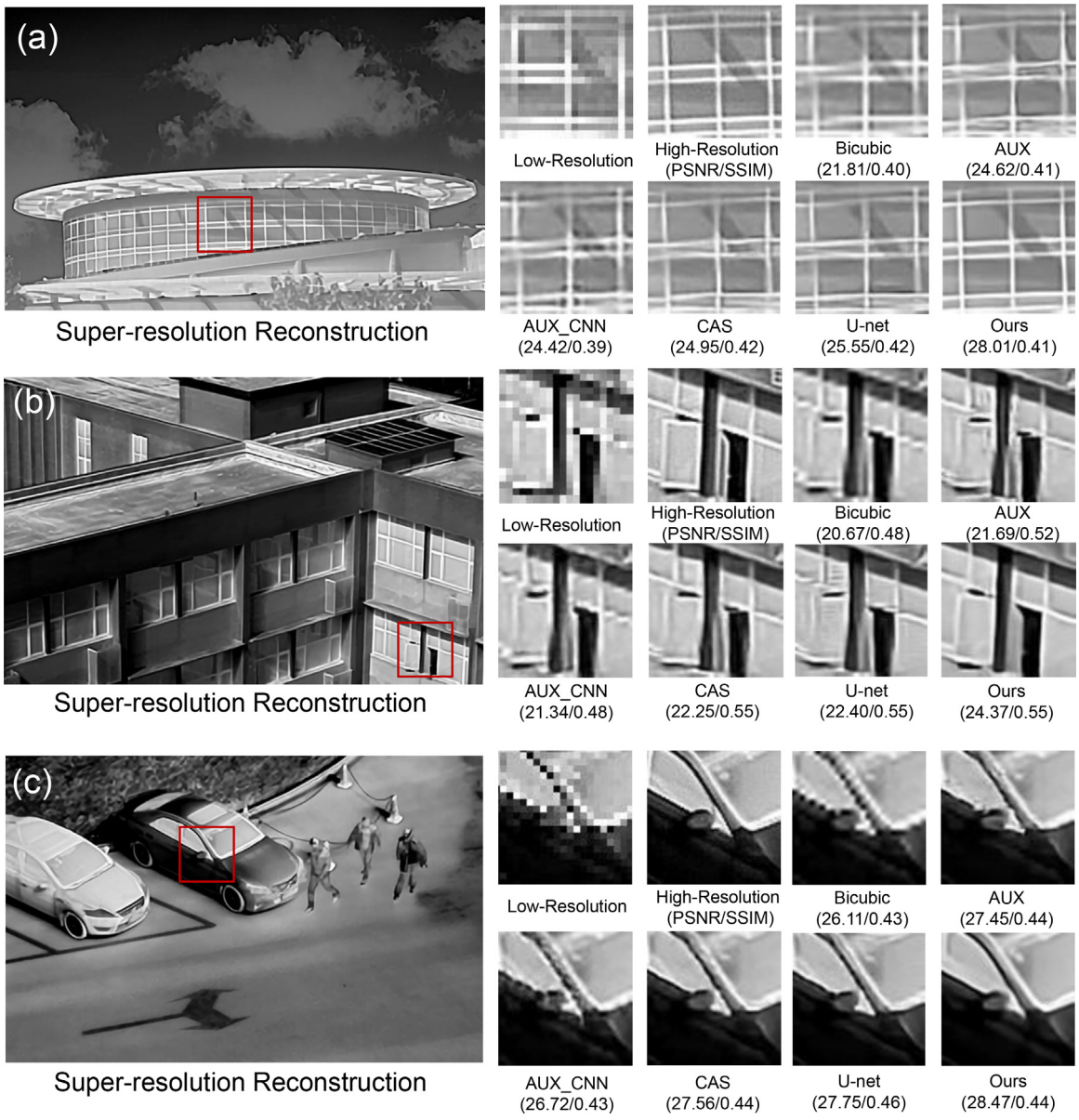

Fig. 4. The comparison of super-resolution imaging results with different scenes (Scale = 3).

the comparison of SR reconstruction results in three different scenes. It can be seen that our method has been sufficiently enhanced in the reconstructed image, whether in edge details or the recovery of spatial frequency components. Compared with bicubic interpolation, auxiliary neural network (AUX) [43], infrared image super-resolution imaging algorithm based on the auxiliary convolutional neural network (AUX-CNN) [44], cascade super-resolution(CAS) [45], and skip connected super-resolution (U-net) approach [40], our method improves the peak signal-to-noise ratio by 4.08dB, 2.36dB, 2.79dB, 2.03dB and 1.71dB, respectively. In addition, from the visual imaging performance, our results are consistent with the HR truth image and avoid the artifact phenomenon in the SR reconstruction result. Therefore, from a comprehensive point of view, the SR image obtained by the proposed method is more prominent. At the same time, it also verifies the feasibility of applying a dual-regression network to improve the SR reconstruction performance.

After verifying the feasibility of the network, we employed the network for heterogenous image fusion processing and made comparisons with the anisotropic diffusion and Karhunen Loeve transform (ADF) [46], fourth order partial differential equations (FPDE) [47], multi scale guided (MGFF) [48], multi singular value decomposition (MSVD) [49] and two scale image fusion using saliency detection (TIF) [50] methods, respectively, and the corresponding reconstruction results are shown in Fig. 5. Although it is difficult to accurately evaluate the visual quality of these methods, we can perceive apparent differences between them. As shown in Fig. 5, all the fusion methods have accomplished the task of merging the information of infrared and visible images to some extent. Overall, our method embraces more textual details while highlighting the important targets.

The reconstruction results are suitable for human eye perception due to the advantages of the high signal-to-noise ratio of the output image and complementary fusion information. From the objective data in Fig. 6, the evaluation indexes of the fused image in spatial frequency, edge intensity, and average gradient were improved over the existing imaging algorithm by 3.35, 8.97, and 0.94, respectively. The comparative data in Table 2 also verify the feasibility of introducing super-resolution networks to improve the reconstruction performance. The image fusion task is reformulated as a problem of maintaining the structure

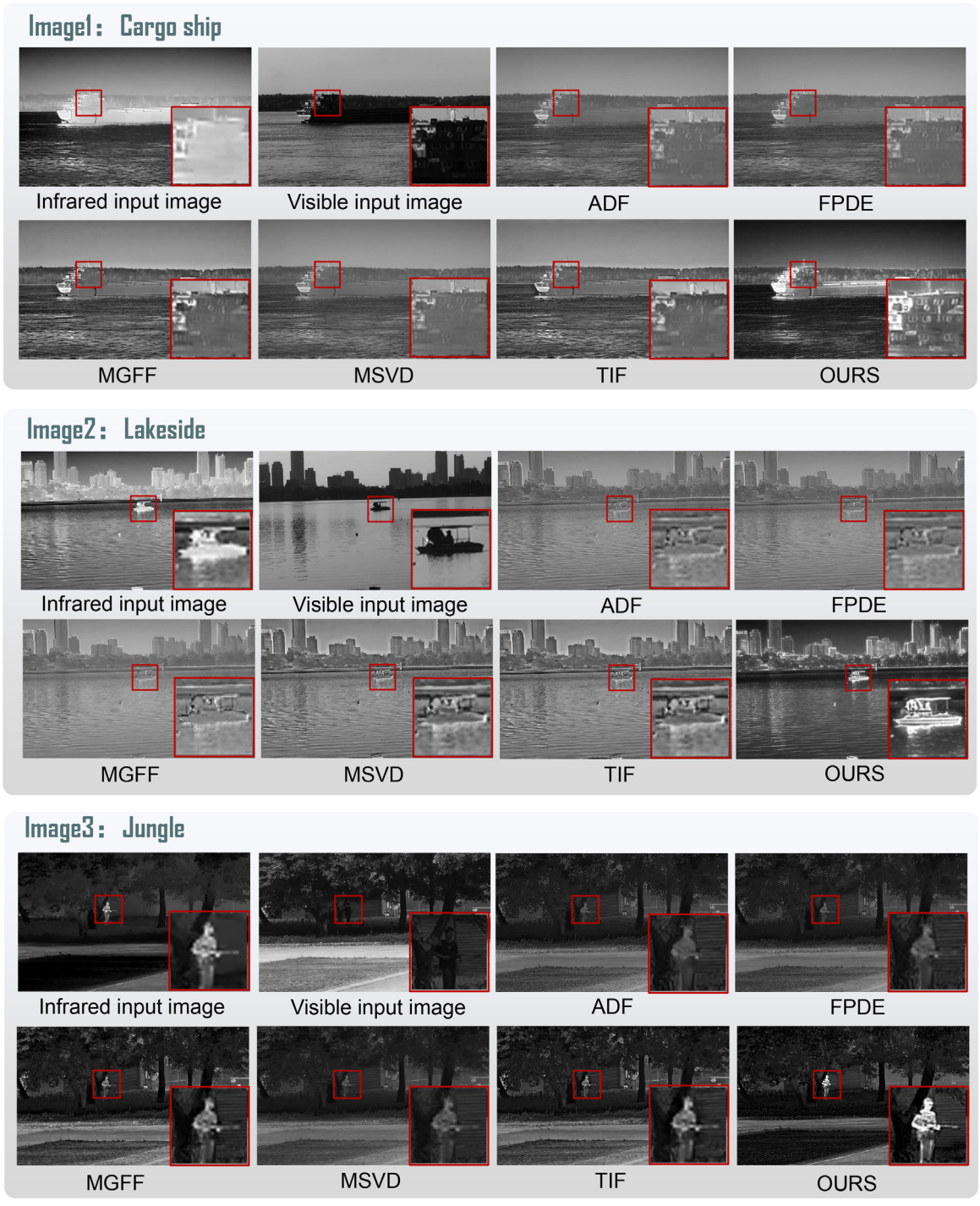

Fig. 5. The comparison of imaging fusion results with different scenes.

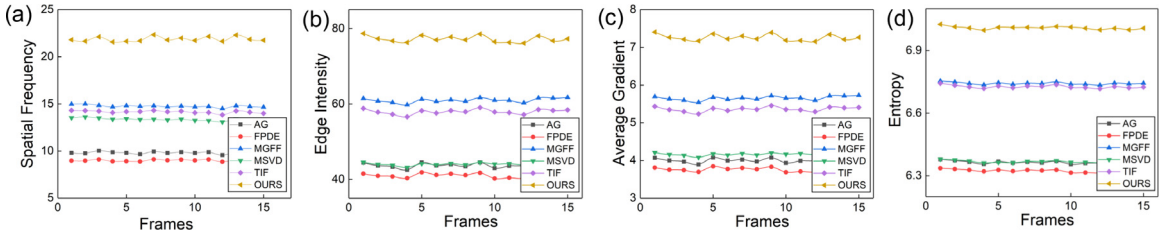

Fig. 6. Index evaluation curve under continuous frames of the same scene.

Table 2
The comparison of imaging fusion evaluation index with different scenes. The bold text indicates the best result.

Number	Methods	AG	Edge intensity	Entropy	Mutinf	Qcv	Rmse	SF
Image 1	ADF	6.2074	59.8872	6.8278	1.8177	0.1008e+03	0.0625	15.6246
Image 1	FPDE	5.7650	55.6909	6.7977	1.8299	0.1142e+03	0.0622	14.0359
Image 1	MGFF	6.9055	68.4516	**7.0109**	1.6050	0.2067e+03	0.0644	17.3653
Image 1	MSVD	5.3488	50.8112	6.7623	**1.8511**	0.1082e+03	0.0622	14.5423
Image 1	TIF	6.0770	60.4062	6.9641	1.6900	0.0927e+03	0.0634	15.6507
Image 1	Ours	**7.3762**	**71.4749**	6.6295	1.3591	**0.6739e+03**	**0.0698**	**18.3819**
Image 2	ADF	4.4591	45.9656	6.8876	2.1563	0.4278e+03	0.0705	12.7356
Image 2	FPDE	4.4130	45.5108	6.8797	2.1499	0.4231e+03	0.0704	12.1605
Image 2	MGFF	5.3440	60.8598	**7.2126**	1.9414	0.4549e+03	0.0725	17.3130
Image 2	MSVD	4.5975	46.9185	6.8986	2.1679	0.4229e+03	0.0705	13.9639
Image 2	TIF	5.2618	55.5220	7.1010	2.0178	0.2774e+03	0.0717	14.5801
Image 2	Ours	**5.9050**	**66.7448**	7.0556	**2.2175**	**0.9902e+03**	**0.1070**	**18.6244**
Image 3	ADF	4.0698	44.4168	6.2849	0.9410	1.1340e+03	0.0650	9.8957
Image 3	FPDE	3.8069	41.4784	6.2352	0.9521	1.1079e+03	0.0647	9.0497
Image 3	MGFF	5.6396	60.8485	6.6629	0.9208	1.0847e+03	0.0661	14.7737
Image 3	MSVD	4.1813	44.2260	6.2795	0.9723	1.1131e+03	0.0648	13.9639
Image 3	TIF	5.4249	58.7020	6.6533	0.8981	1.0184e+03	0.0665	14.1778
Image 3	Ours	**7.4351**	**78.8509**	**7.0185**	**1.2953**	**2.3615e+03**	**0.1062**	**22.4927**

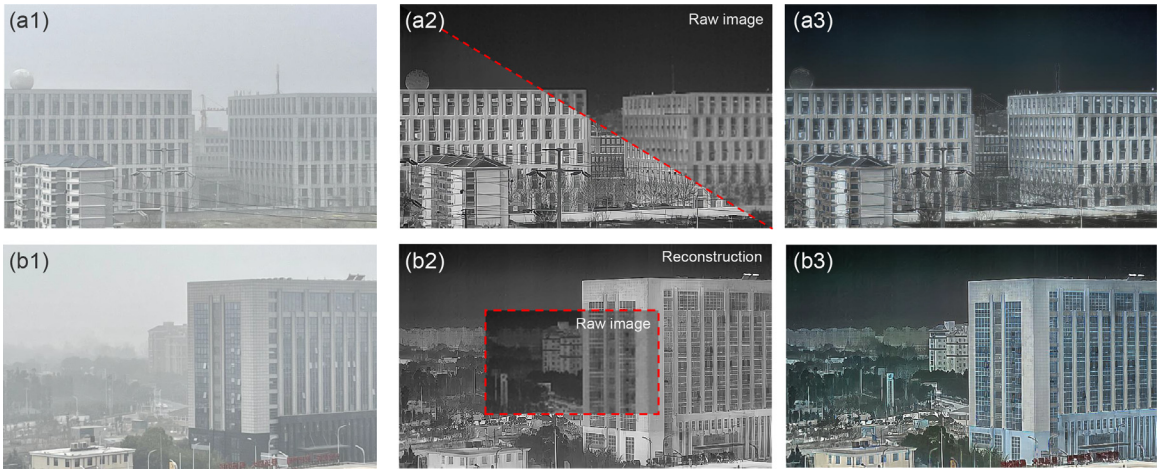

Fig. 7. The imaging results of the proposed algorithm in severe weather (foggy days). (a1, b1) Visible image. (a2, b2) Infrared image. (a3, b3) Fusion image.

and intensity ratio of the infrared-visible image, solving the problem of poor quality fusion performance and thermal information blurring due to the low resolution of the infrared image in conventional fusion imaging.

For the fusion imaging problem under severe weather (foggy days), we have also explored it accordingly. As shown in Fig. 7, under a foggy sky, the scene captured by the visible detector is muddy and contains an amount of interference information. On the contrary, long-wave infrared detectors capture unique signals by virtue of the characteristics of penetrating smoke imaging and thermal radiation sensing. The multi-scale feature extraction network effectively realizes the high-frequency information fusion of different detectors in the proposed method. An excellent color fusion image can be achieved with the help of color information from the visible detector, as depicted in Fig. 7(a3, b3). However, the infrared image also has the imaging problem of poor contrast due to less thermal radiation information on foggy days. By regressing the output of the super-resolution network, the corresponding high-frequency detail information is basically restored, as shown in Fig. 7(a2, b2).

In addition, the recovery of image color information is also an uncertainty problem. Deep learning-based color image reconstruction is mainly established on specific scenes and cannot recover color information that does not appear in the training set. Therefore, this imposes strict requirements on the training set, which should contain as much color information as possible for various scenes. Fig. 8 portrays the multi-modal imaging results of heterologous images based on the regression network. Various modes of reconstruction such as pseudo-color, SR reconstruction, and edge extraction are realized. See supplementary visualization materials 1, 2, and 3 for specific imaging videos. The experimental results indicate that the network is able to perform fused images containing thermal information of infrared images and high-frequency information of visible images, which comprehensively enhances the resolution of detailed textures of infrared images. At the same time, the obtained colored image is consistent with the visual perception effect of the human eyes. With the guidance of thermal radiation signals, the contour markings of moving objects can be unambiguous marked to further facilitate the information perception ability. For instance, the image resolution and thermal information of toy guns held by pedestrians are significantly improved in infrared images. From the reconstruction results in Fig. 8(c), we can clearly observe that the fused target images are effectively highlighted in the low light environment, which will be conducive to the subsequent target recognition and tracking. In general, our method can offer robust adaptability in different imaging environments, and it will also provide a promising way to improve the quality of infrared-visible fusion.

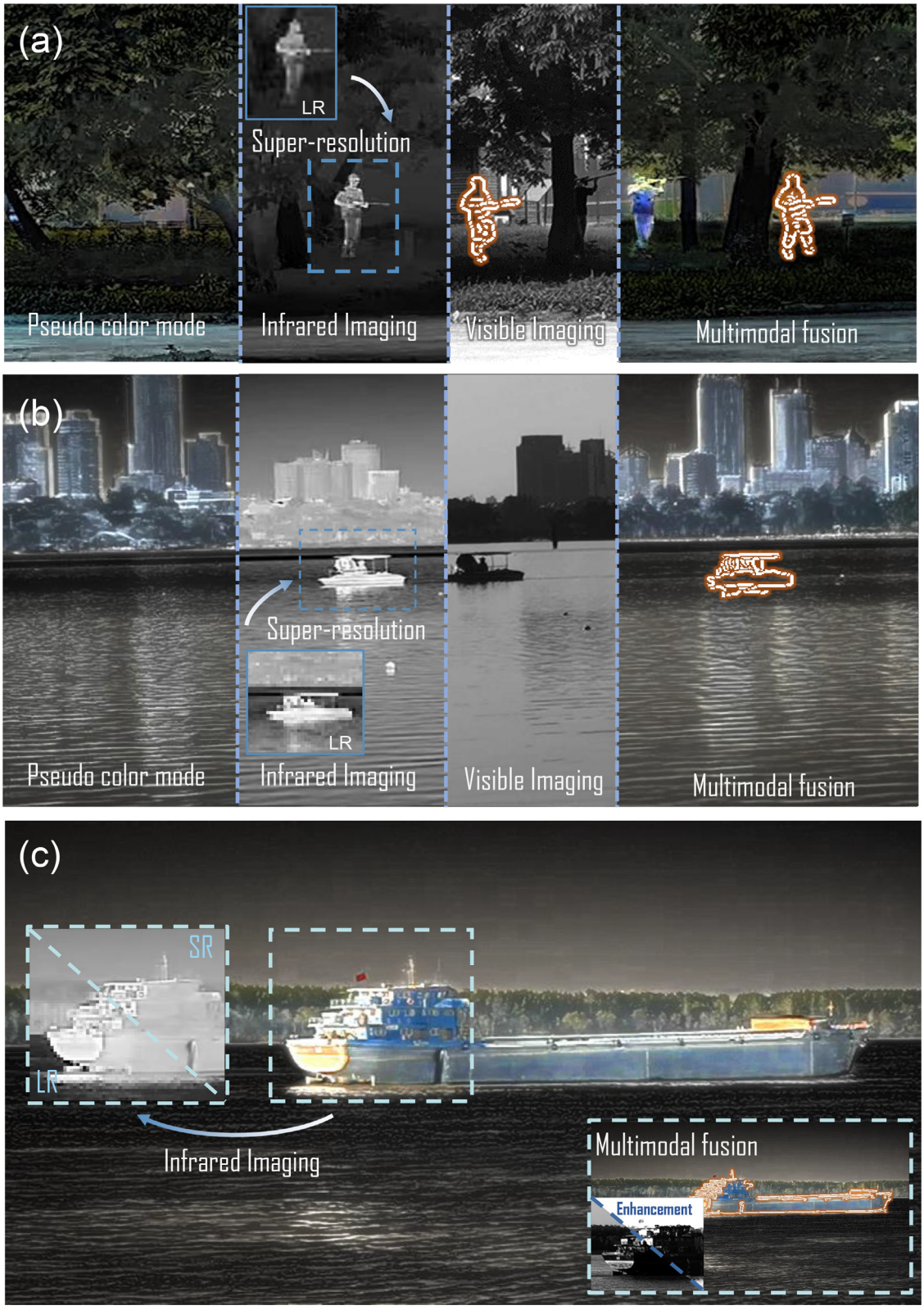

Fig. 8. Cross modal reconstruction results in different scenes.

6. Conclusion

To address the bottleneck of low-quality fusion imaging caused by different imaging mechanisms and mismatched spatial resolution of heterogeneous detectors, an infrared-visible cross-modal color fusion network based on DL is proposed. Affording the conception of semantic segmentation and style transfer, the encoding-decoding fusion network is adopted for end-to-end learning to improve the feature expression ability and suppress the interference of useless information. The corresponding dual-loss function is designed to expand the weight difference between thermal target and background. Experimental results prove the superiority in terms of visual quality and quantitative criteria compared to five representative methods. The evaluation indexes of spatial frequency, edge intensity, and average gradient were improved by 3.35, 8.97, and 0.94, respectively, which significantly improved the imaging quality of the fused images and verified the application potential of the network. On this basis, the imaging output of HR infrared reconstruction, heterologous image pseudo color fusion, edge feature extraction and other modes are realized, which opens new avenues for subsequent HR reconnaissance and identification tasks.

Declaration of Competing Interest

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

CRediT authorship contribution statement

Bowen Wang: Conceptualization, Methodology, Visualization, Writing – original draft. **Yan Zou:** Visualization, Investigation. **Linfei Zhang:** Investigation. **Yuhai Li:** Formal analysis. **Qian Chen:** Supervision, Funding acquisition. **Chao Zuo:** Funding acquisition, Supervision, Writing – review & editing, Project administration.

Acknowledgement

This work was supported by the National Natural Science Foundation of China (61905115, 62105151, 62175109, U21B2033), Leading Technology of Jiangsu Basic Research Plan (BK20192003), Youth Foundation of Jiangsu Province (BK20190445, BK20210338), Fundamental Research Funds for the Central Universities (30920032101), and Open Research Fund of Jiangsu Key Laboratory of Spectral Imaging & Intelligent Sense (JSGP202105).

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.optlaseng.2022.107078.

References

[1] Stathaki T. Image fusion: algorithms and applications. Elsevier; 2011.
[2] Sahu DK, Parsai M. Different image fusion techniques–a critical review. Int J Mod Eng Res (IJMER) 2012;2(5):4298–301.
[3] Chen Y, Cheng L, Wu H, Mo F, Chen Z. Infrared and visible image fusion based on iterative differential thermal information filter. Opt Lasers Eng 2022;148:106776.
[4] James AP, Dasarathy BV. Medical image fusion: a survey of the state of the art. Inform Fus 2014;19:4–19.
[5] Nemalidinne SM, Gupta D. Nonsubsampled contourlet domain visible and infrared image fusion framework for fire detection using pulse coupled neural network and spatial fuzzy clustering. Fire Saf J 2018;101:84–101.
[6] Shen J, Zhao Y, Yan S, Li X, et al. Exposure fusion using boosting laplacian pyramid. IEEE Trans Cybern 2014;44(9):1579–90.
[7] Mertens T, Kautz J, Van Reeth F. Exposure fusion: A simple and practical alternative to high dynamic range photography. In: Computer graphics forum, vol. 28. Wiley Online Library; 2009. p. 161–71.
[8] Li ZG, Zheng JH, Rahardja S. Detail-enhanced exposure fusion. IEEE Trans Image Process 2012;21(11):4672–6.
[9] Xiang T, Yan L, Gao R. A fusion algorithm for infrared and visible images based on adaptive dual-channel unit-linking pcnn in nsct domain. Infrared Phys Technol 2015;69:53–61.
[10] Ma J, Chen C, Li C, Huang J. Infrared and visible image fusion via gradient transfer and total variation minimization. Inform Fus 2016;31:100–9.
[11] Ma J, Ma Y, Li C. Infrared and visible image fusion methods and applications: a survey. Inform Fus 2019;45:153–78.
[12] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521(7553):436–44.
[13] Deng L, Yu D. Deep learning: methods and applications. Found Trend Signal Process 2014;7(3–4):197–387.
[14] Russell S, Norvig P. Artificial intelligence: a modern approach; 2002.
[15] Ertel W. Introduction to artificial intelligence. Springer; 2018.
[16] Feng S, Chen Q, Gu G, Tao T, Zhang L, Hu Y, Yin W, Zuo C. Fringe pattern analysis using deep learning. Adv Photon 2019;1(2):025001.
[17] Feng S, Zuo C, Hu Y, Li Y, Chen Q. Deep-learning-based fringe-pattern analysis with uncertainty estimation. Optica 2021;8(12):1507–10.
[18] Voulodimos A, Doulamis N, Doulamis A, Protopapadakis E. Deep learning for computer vision: a brief review. Comput Intell Neurosci 2018;2018.
[19] O'Mahony N, Campbell S, Carvalho A, Harapanahalli S, Hernandez GV, Krpalkova L, Riordan D, Walsh J. Deep learning vs. traditional computer vision. In: Science and Information Conference. Springer; 2019. p. 128–44.
[20] Perez L, Wang J. The effectiveness of data augmentation in image classification using deep learning. arXiv preprint arXiv:171204621 2017.
[21] Zhao Z-Q, Zheng P, Xu S-t, Wu X. Object detection with deep learning: a review. IEEE Trans Neural Netw Learn Syst 2019;30(11):3212–32.
[22] Liu L, Ouyang W, Wang X, Fieguth P, Chen J, Liu X, Pietikäinen M. Deep learning for generic object detection: asurvey. Int J Comput Vis 2020;128(2):261–318.
[23] Uçar A, Demir Y, Güzeliş C. Object recognition and detection with deep learning for autonomous driving applications. Simulation 2017;93(9):759–69.
[24] Eitel A, Springenberg JT, Spinello L, Riedmiller M, Burgard W. Multimodal deep learning for robust rgb-d object recognition. In: 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE; 2015. p. 681–7.
[25] Van Ouwerkerk J. Image super-resolution survey. Image Vis Comput 2006;24(10):1039–52.
[26] Wang L, Zhao S. Super resolution ghost imaging based on fourier spectrum acquisition. Opt Lasers Eng 2021;139:106473.
[27] Gu J, Wang Z, Kuen J, Ma L, Shahroudy A, Shuai B, Liu T, Wang X, Wang G, Cai J, et al. Recent advances in convolutional neural networks. Pattern Recognit 2018;77:354–77.
[28] Li H, Wu X-J, Kittler J. Infrared and visible image fusion using a deep learning framework. In: 2018 24th international conference on pattern recognition (ICPR). IEEE; 2018. p. 2705–10.
[29] Wang B, Zou Y, Zhang L, Hu Y, Yan H, Zuo C, Chen Q. Low-light-level image super-resolution reconstruction based on a multi-scale features extraction network. In: Photonics, vol. 8. Multidisciplinary Digital Publishing Institute; 2021. p. 321.
[30] Gurrola-Ramos J, Dalmau O, Alarcón T. U-Net based neural network for fringe pattern denoising. Opt Laser Eng 2022;149:106829.
[31] Qian J, Feng S, Tao T, Hu Y, Li Y, Chen Q, Zuo C. Deep-learning-enabled geometric constraints and phase unwrapping for single-shot absolute 3d shape measurement. APL Photonics 2020;5(4):046105.
[32] Ma J, Yu W, Chen C, Liang P, Guo X, Jiang J. Pan-gan: an unsupervised pan-sharpening method for remote sensing image fusion. Inform Fus 2020a;62:110–20.
[33] Bell-Kligler S, Shocher A, Irani M. Blind super-resolution kernel estimation using an internal-gan. arXiv preprint arXiv:190906581 2019.
[34] Yang X, Jiang P, Jiang M, Xu L, Wu L, Yang C, Zhang W, Zhang J, Zhang Y. High imaging quality of fourier single pixel imaging based on generative adversarial networks at low sampling rate. Opt Lasers Eng 2021;140:106533.
[35] Ram Prabhakar K, Sai Srikar V, Venkatesh Babu R. Deepfuse: a deep unsupervised approach for exposure fusion with extreme exposure image pairs. In: Proceedings of the IEEE international conference on computer vision; 2017. p. 4714–22.
[36] Li H, Wu X-J. Densefuse: a fusion approach to infrared and visible images. IEEE Trans Image Process 2018;28(5):2614–23.
[37] Zhang C, Hu H, Tai Y, Yun L, Zhang J. Trustworthy image fusion with deep learning for wireless applications. Wirel Commun Mob Comput 2021;2021.
[38] Ma J, Liang P, Yu W, Chen C, Guo X, Wu J, Jiang J. Infrared and visible image fusion via detail preserving adversarial learning. Inform Fus 2020b;54:85–98.
[39] Li J, Huo H, Li C, Wang R, Feng Q. Attentionfgan: infrared and visible image fusion using attention-based generative adversarial networks. IEEE Trans Multimedia 2020;23:1383–96.
[40] Zou Y, Zhang L, Liu C, Wang B, Hu Y, Chen Q. Super-resolution reconstruction of infrared images based on a convolutional neural network with skip connections. Opt Lasers Eng 2021;146:106717.
[41] Gatys LA, Ecker AS, Bethge M, Hertzmann A, Shechtman E. Controlling perceptual factors in neural style transfer. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2017. p. 3985–93.
[42] Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. Springer; 2015. p. 234–41.
[43] Han TY, Kim DH, Lee SH, Song BC. Infrared image super-resolution using auxiliary convolutional neural network and visible image under low-light conditions. J Vis Commun Image Represent 2018;51:191–200.
[44] Zou Y, Zhang L, Chen Q, Wang B, Hu Y, Zhang Y. An infrared image super-resolution imaging algorithm based on auxiliary convolution neural network. In: Optics Frontier Online 2020: Optics Imaging and Display, vol. 11571. International Society for Optics and Photonics; 2020. 115711B.

[45] He Z, Tang S, Yang J, Cao Y, Yang MY, Cao Y. Cascaded deep networks with multiple receptive fields for infrared image super-resolution. IEEE Trans Circuits Syst Video Technol 2018;29(8):2310–22.
[46] Bavirisetti DP, Dhuli R. Fusion of infrared and visible sensor images based on anisotropic diffusion and karhunen-loeve transform. IEEE Sens J 2015;16(1):203–9.
[47] Bavirisetti DP, Xiao G, Liu G. Multi-sensor image fusion based on fourth order partial differential equations. In: 2017 20th International conference on information fusion (Fusion). IEEE; 2017. p. 1–9.
[48] Bavirisetti DP, Xiao G, Zhao J, Dhuli R, Liu G. Multi-scale guided image and video fusion: a fast and efficient approach. Circuits, Systems, and Signal Processing 2019;38(12):5576–605.
[49] Naidu V. Image fusion technique using multi-resolution singular value decomposition. Def Sci J 2011;61(5):479.
[50] Bavirisetti DP, Dhuli R. Two-scale image fusion of visible and infrared images using saliency detection. Infrared Phys Technol 2016;76:52–64.