Frequency-Spatial Feature Fusion Network for Infrared and Visible Image Fusion

Wenhao Song[1], Mingliang Gao[1(✉)], Qilei Li[2], Gwanggil Jeon[3], and David Camacho[4(✉)]

[1] School of Electrical and Electronic Engineering, Shandong University of Technology, Zibo 255000, China
sdut_songwenhao@163.com, mlgao@sdut.edu.cn

[2] School of Electronic Engineering and Computer Science, Queen Mary University of London, London E1 4NS, United Kingdom
q.li@qmul.ac.uk

[3] Department of Embedded Systems Engineering, Incheon National University, Incheon 22012, South Korea
gjeon@inu.ac.kr

[4] Computer Systems Engineering Department, Universidad Politécnica de Madrid (UPM), Madrid, Spain
david.camacho@upm.es

Abstract. Infrared and visible image fusion seeks to retain complementary information from source images and generate a comprehensive image. Most fusion methods ignore the detailed information in the frequency domain. To address this problem, we propose a Frequency-spatial Feature Fusion Network (F3Net) in this work. The F3Net consists of three modules, namely Frequency-Spatial Feature Extraction Module (FSFEM), Feature Fusion Module (FFM), and Image Reconstruction Module (IRM). First, the FSFEM is built to extract complementary information from the source image separately in the frequency and spatial domains. Then, the FMM is introduced to fuse the features of the frequency and spatial domains. Finally, the fused image is reconstructed by IRM. Comprehensive experiments demonstrate that the F3Net outperforms the state-of-the-art (SOTA) methods subjectively and objectively.

Keywords: Image fusion · Frequency domain feature · Feature fusion · Deep learning

1 Introduction

Image fusion aims to integrate information from source images of the same scene, and produces a fused image with enhanced quality and details [6]. Image fusion has various applications in medical imaging [5], remote sensing [21], surveillance [15], and night vision [16]. In the realm of image fusion, infrared and

N. T. Nguyen et al. (Eds.): CITA 2024, LNNS 882, pp. 28–37, 2024.
https://doi.org/10.1007/978-3-031-74127-2_3

visible image fusion (IVIF) hold significance. This method combines thermal information from infrared (IR) images with texture information from visible (VIS) images and yields a more comprehensive and informative result [17].

Over the past few years, numerous methods for fusing infrared and visible images have emerged. These approaches can be broadly divided into two main groups, namely traditional methods and deep learning-based methods. Traditional methods usually operate in a specific transform domain, such as wavelet [27], curvelet [28], or contourlet [4], and fuse the features of the source images according to some predefined rules or criteria [13]. Nevertheless, the traditional method relies on complicated transforms or representations to improve the fusion quality, which increases the computational cost and time. Consequently, these traditional methods may not be well-suited for real-time applications [9].

To address this problem, deep learning-based methods have been developed and they can be divided into three categories based on different network architectures, namely autoencoder (AE)-based method [31], convolutional neural network (CNN)-based method [19], and generative adversarial network (GAN)-based method [3]. Given convolutional neural networks' potent feature extraction capabilities, deep learning methods can effectively extract rich and complementary information from source images. Nevertheless, many existing deep learning methods predominantly concentrate on spatial domain features and overlook the frequency domain features. This is because convolutional neural networks, commonly used in these methods, are particularly effective at extracting rich spatial features that are more directly related to visual aspects like texture and object details. Consequently, the rich edge detail and other high-frequency information in the frequency domain might be underutilized.

We propose a frequency-spatial feature fusion network (F3Net) for IVIF to tackle these challenges. The main contributions of this work are as follows,

- We design a frequency-spatial feature extraction module (FSFEM) to extract frequency and spatial domain features from the source images.
- We propose a feature fusion module (FFM) to fuse the frequency domain and spatial domain features, thereby obtaining more comprehensive and representative fusion features.
- Comprehensive experiments demonstrate that the proposed method can achieve superior performance over state-of-the-art (SOTA) methods, objective and subjective.

2 Proposed Method

2.1 Overview

The overall architecture of F3Net is illustrated in Fig. 1. It aims to integrate the thermal information from the IR image and the texture information from the VIS image. It includes the frequency-spatial feature extraction module (FSFEM), the Feature Fusion Module (FFM), and the Image Reconstruction Module (IRM). First, FSFEM extracts features from both the frequency and spatial domains of the source images. Subsequently, the FFM combines these frequency and spatial features. Finally, the IRM reconstructs the fused image using four convolutional layers.

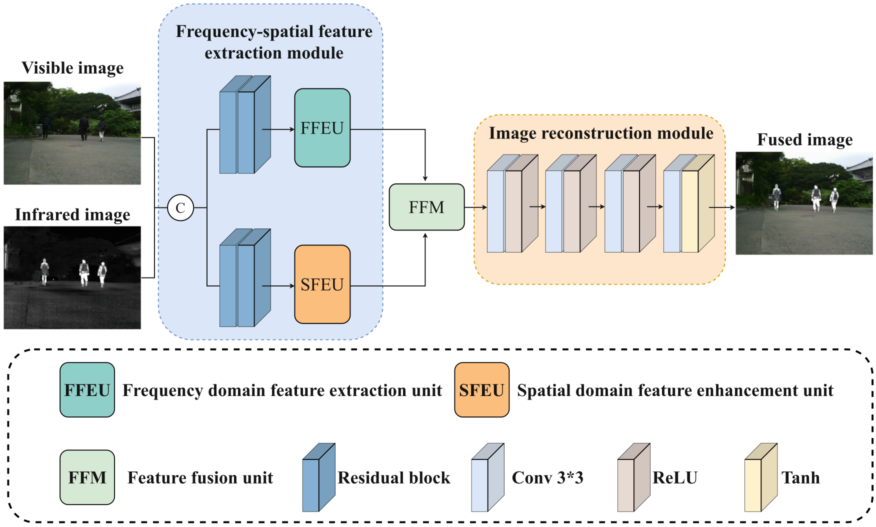

Fig. 1. The overall framework of the proposed method.

2.2 Network Architecture

Frequency-Spatial Feature Extraction Module The Frequency-Spatial Feature Extraction Module (FSFEM) is specifically designed to fuse features from both the frequency domain and spatial domain of the source images. It consists of three components, namely Residual block (Resblock), the Fcanet [14]-based Frequency domain Feature Extraction Unit (FFEU), and the CBAM [23]-based Spatial domain Feature Enhancement Unit (SFEU). The Residual Block (Resblock) serves as a fundamental component of FSFEM. It consists of two convolutional layers with a ReLU activation function and a skip connection. The Resblock is crucial in preserving low-level information from the source images and enhancing feature representation. The FFEU is designed to extract frequency-domain features from the source images. It captures global frequency information and enhances edge and texture details, contributing to a more comprehensive representation. The SFEU focuses on enhancing the spatial domain features of the source images. It dynamically adjusts the spatial and channel-wise importance of input features, highlighting salient regions.

Feature Fusion Module The feature fusion module (FFM) is designed to fuse the frequency domain and spatial domain features. It consists of a max pooling layer, two convolutional layers with 7×7 and 1×1 kernels, a sigmoid activation function, an element-wise summation operation, and an element-wise multiplication operation. The framework of FFM is illustrated in Fig. 2. The FFM initiates the process by applying max pooling on the spatial domain features. Subsequently, it utilizes a 7×7 convolution layer to reduce dimensionality and increase the receptive field of these features. Then, it integrates frequency domain features with spatial domain features, determining feature map weights through a Sigmoid activation function. Detailed information is enhanced by element-wise multiplication of feature map weights with frequency domain features, and original spatial domain information is supplemented through element-wise addition. Finally, it employs a channel attention mechanism to emphasize critical channel information.

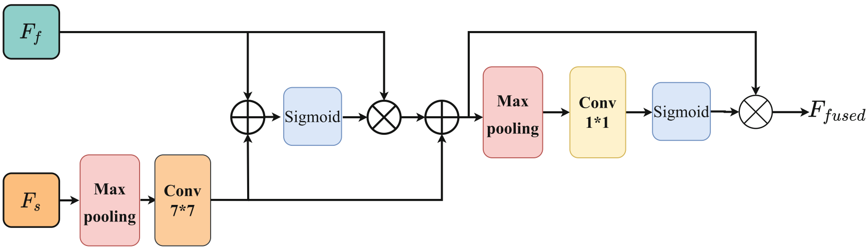

Fig. 2. Architecture of the Feature fusion module (FFM).

Image Reconstruction Module The IRM is responsible for reconstructing the fused image from the integrated features obtained from the FFM. As shown in Fig. 1, IRM employs a sequence of four convolutional layers to perform the image reconstruction. The IRM is responsible for reconstructing the fused image from the integrated features obtained from the FFM. This module employs a sequence of four convolutional layers to perform the image reconstruction. Each layer applies 3×3 convolutional operations to transform the fused feature maps into higher resolution. Each convolutional layer employs ReLU activation functions to introduce non-linearity into the model, enabling it to learn more complex patterns. The final layer of the IRM might use a Tanh activation function to normalize the output pixels to the appropriate range. The primary purpose of the IRM is to reconstruct a coherent and visually enhanced output image using the feature maps that represent both frequency and spatial information of the source images, as processed and combined by the FSFEM and FFM.

2.3 Loss Function

To enhance the visual quality of the fused image, we develop a loss function comprising three distinct components, namely pixel loss $\mathcal{L}_{pixel}$, gradient loss $\mathcal{L}_{gradient}$, and structural loss $\mathcal{L}_{ssim}$. The total loss is expressed as,

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{pixel} + \lambda_2 \mathcal{L}_{gradient} + \lambda_3 \mathcal{L}_{ssim}, \tag{1}$$

where λ_1, λ_2, and λ_3 are the weights of the respective terms.

The pixel loss quantifies the pixel-wise discrepancy between the fused image and the source images. It is presented as,

$$\mathcal{L}_{\text{pixel}} = \frac{1}{HW}(\|I_f - I_{ir}\|_F^2 + \|I_f - I_{vis}\|_F^2), \tag{2}$$

where H and W are the height and width of the image, respectively. I_f is the fused image, I_{ir} is the infrared image, and I_{vis} is the visible image. $\|\cdot\|_F$ denotes the Frobenius norm of the matrix.

The gradient loss captures the gradient difference between the fused image and the source images. It is calculated as:

$$\mathcal{L}_{\text{gradient}} = \|\nabla I_f - \max\{\nabla I_{ir}, \nabla I_{vis}\}\|_2, \tag{3}$$

where $\|\cdot\|_2$ represents the ℓ_2-norm of the matrix. ∇ is the gradient operation, and $max\{,\}$ denotes maximum operator.

The structural loss evaluates the structural similarity between the fused and source images. It is defined as,

$$\mathcal{L}_{\text{ssim}} = 1 - \text{SSIM}\left(I_f, \max\{I_{ir}, I_{vis}\}\right), \tag{4}$$

where $SSIM$ is the structural similarity index [22].

3 Experimental Results and Analysis

3.1 Datasets and Experiment Setup

In this study, we employed the MSRS [20] dataset to assess the performance of the proposed model. The MSRS dataset comprises 1444 pairs of source images captured in diverse scenarios. The training set consists of 1083 pairs of images, while the test set comprises 361 pairs of images.

The learning rate of F3Net was initially set to 0.001. The Adam is adopted to optimize the model and set the batch size to 4. For the loss function, the hyperparameters λ_1, λ_2, and λ_3 were set to 1, 100, and 10, respectively. In this section, we compare the fusion results with the nine SOTA methods. These methods include CSF [26], CUFD [24], Densefuse [7], DIDFuse [32], FusionGAN [12], IFCNN [29], NestFuse [8], SuperFusion [18], and U2Fusion [25].

3.2 Comparison with SOTA Methods

Subjective Evaluation The subjective comparison is crucial for applications where human perception. Observers compare the fused images produced by different methods to assess clarity, detail retention, and overall visual appeal. When assessing images subjectively, observers anticipate that the fused image incorporates the optimal attributes of both IR and visible VIS images. Ideally, the fused image should accentuate critical thermal information from the IR image, while preserving the high-resolution detail from the VIS image. The results of the subjective comparison are shown in Fig. 3. It shows that the F3Net can effectively preserve the salient information of the IR image and the detailed texture of the VIS image. Compared with the competitors, the proposed method excels in retaining more details and salient information from the source images. Additionally, it avoids artifacts and distortions in some methods like CSF, CUFD, DIDFuse, and FusionGAN.

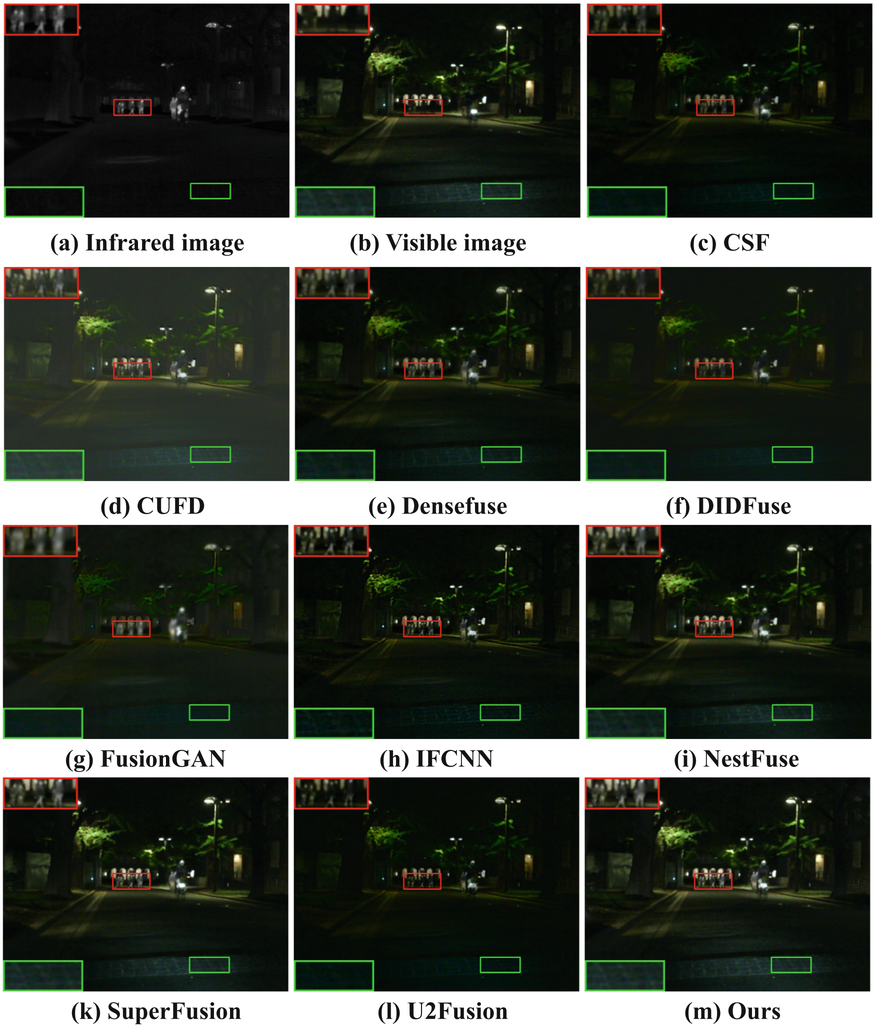

Fig. 3. Subjective results of the proposed F3Net and the competitors.

Objective Evaluation To evaluate objective performance, we use four metrics: entropy (EN), mutual information (MI), visual information fidelity (VIF), and Qabf. EN quantifies the information content in the fused image, MI gauges the preservation of information from source images to the fused image, VIF evaluates the perceptual quality of the fused image, and Qabf assesses fusion quality based on spatial and spectral information. Higher values for these metrics indicate superior performance.

The results of the objective comparison are presented in Table 1. F3Net outperforms state-of-the-art (SOTA) methods across all four metrics, confirming its superior performance. Specifically, F3Net attains the highest score in EN, indicating optimal retention of information content within the fused image. Moreover, its leading score in MI suggests that F3Net preserves information from the source images most effectively. F3Net also ranks first in VIF, signifying its superior perceptual quality. Additionally, achieving the highest score in Qabf underscores the excellent fusion quality of the F3Net, considering both spatial and spectral information. The proposed method effectively combines the complementary information from the source images to produce a high-quality fused image enriched with details.

Table 1. Objective evaluation of the proposed F3Net and the nine competitors in EN, MI, VIF, and Qabf metrics. (The best results are marked in bold)

Method	EN	MI	VIF	Qabf
CSF [26]	5.836	2.344	0.666	0.368
CUFD [24]	6.056	2.989	0.644	0.433
Densefuse [7]	6.217	2.642	0.773	0.485
DIDFuse [32]	5.303	2.536	0.487	0.260
FusionGAN [12]	5.440	1.853	0.500	0.139
IFCNN [29]	5.975	1.857	0.712	0.519
NestFuse [8]	6.501	3.573	0.926	0.627
SuperFusion [18]	6.587	4.216	0.960	0.631
U2Fusion [25]	5.561	2.246	0.422	0.419
Ours	**6.655**	**4.952**	**0.998**	**0.669**

3.3 Computational Complexity Analysis

To verify the adaptability of deep learning-based methods compared to traditional methods in practical applications, we conducted a computational efficiency experiment on the MSRS dataset. We compare the computational efficiency of the proposed method with six traditional image fusion methods, namely ADF [1], CNN [10], FPDE [2], GFCE [33], GTF [11], and IFEVIP [30].

The unit fusion time of the proposed method and six traditional methods on the MSRS dataset are shown in Table 2. It can be seen that the proposed method has a significant time efficiency advantage over traditional methods.

Table 2. Computational efficiency of F3Net with the compared fusion methods.

Methods	ADF [1]	CNN [10]	FPDE [2]	GFCE [33]	GTF [11]	IFEVIP [30]	Ours
Time (s)	0.61	25.52	1.11	0.93	2.88	0.25	0.11

4 Conclusion

This work introduces a Frequency-spatial Feature Fusion Network (F3Net) for infrared and visible image fusion. The proposed model can effectively extract and fuse frequency and spatial domain features from the source images. It consists of three modules, namely Frequency-spatial Feature Extraction Module (FSFEM), Feature Fusion Module (FFM), and Image Reconstruction Module (IRM). The FSFEM is designed to extract the frequency domain and spatial domain features from source images. FFM is used to fuse frequency domain and spatial domain features adaptively. Experimental results demonstrate that the proposed method can achieve superior performance over the existing SOTA methods, both subjectively and objectively.

Acknowledgement. This work has been funded by PCI2022-134990-2 (MARTINI) of the CHISTERA IV Cofund 2021 program, funded by MCIN/AEI/10.13039/501100011033 and by the "European Union NextGenerationEU/PRTR; by Grant PLEC2021-007681 (XAI-DisInfodemics) and PID2020-117263GB-100 (FightDIS) funded by MCIN/AEI/ 10.13039/501100011033 and, as appropriate, by "ERDF A way of making Europe," by the "European Union" or by the "European Union NextGenerationEU/PRTR".

References

1. Bavirisetti DP, Dhuli R (2015) Fusion of infrared and visible sensor images based on anisotropic diffusion and Karhunen-Loeve transform. IEEE Sens J 16(1):203–209
2. Bavirisetti DP, Xiao G, Liu G (2017) Multi-sensor image fusion based on fourth order partial differential equations. In: 2017 20th international conference on information fusion (Fusion), pp 1–9. IEEE
3. Gao M, Zhou Y, Zhai W, Zeng S, Li Q (2023) Saregan: a salient regional generative adversarial network for visible and infrared image fusion. Multimedia Tools Appl 1–13
4. He K, Zhou D, Zhang X, Nie R (2018) Infrared and visible image fusion combining interesting region detection and nonsubsampled contourlet transform. J Sens 2018
5. Hermessi H, Mourali O, Zagrouba E (2021) Multimodal medical image fusion review: theoretical background and recent advances. Signal Process 183:108036

6. Karim S, Tong G, Li J, Qadir A, Farooq U, Yu Y (2023) Current advances and future perspectives of image fusion: a comprehensive review. Inf Fusion 90:185–217
7. Li H, Wu XJ (2019) Densefuse: a fusion approach to infrared and visible images. IEEE Trans Image Process 28(5):2614–2623
8. Li H, Wu XJ, Durrani T (2020) NestFuse: an infrared and visible image fusion architecture based on nest connection and spatial/channel attention models. IEEE Trans Instrum Meas 69(12):9645–9656. https://doi.org/10.1109/TIM.2020.3005230
9. Li S, Kang X, Fang L, Hu J, Yin H (2017) Pixel-level image fusion: a survey of the state of the art. Inf Fusion 33:100–112
10. Liu Y, Chen X, Cheng J, Peng H, Wang Z (2018) Infrared and visible image fusion with convolutional neural networks. Int J Wavelets Multiresolut Inf Process 16(03):1850018
11. Ma J, Chen C, Li C, Huang J (2016) Infrared and visible image fusion via gradient transfer and total variation minimization. Inf Fusion 31:100–109
12. Ma J, Yu W, Liang P, Li C, Jiang J (2019) Fusiongan: a generative adversarial network for infrared and visible image fusion. Inf Fusion 48:11–26
13. Ma J, Zhou Y (2020) Infrared and visible image fusion via gradientlet filter. Comput Vis Image Underst 197:103016
14. Qin Z, Zhang P, Wu F, Li X (2021) Fcanet: frequency channel attention networks. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 783–792
15. Singh P, Diwakar M, Cheng X, Shankar A (2021) A new wavelet-based multi-focus image fusion technique using method noise and anisotropic diffusion for real-time surveillance application. J Real-Time Image Process 18(4):1051–1068
16. Singh S, Singh H, Mittal N, Hussien AG, Sroubek F (2022) A feature level image fusion for night-vision context enhancement using arithmetic optimization algorithm based image segmentation. Expert Syst Appl 209:118272
17. Song W, Zhai W, Gao M, Li Q, Chehri A, Jeon G (2023) Multiscale aggregation and illumination-aware attention network for infrared and visible image fusion. Concurr Comput Pract Exp e7712
18. Tang L, Deng Y, Ma Y, Huang J, Ma J (2022) Superfusion: a versatile image registration and fusion network with semantic awareness. IEEE/CAA J Automatica Sinica 9(12):2121–2137
19. Tang L, Xiang X, Zhang H, Gong M, Ma J (2023) Divfusion: darkness-free infrared and visible image fusion. Inf Fusion 91:477–493
20. Tang L, Yuan J, Zhang H, Jiang X, Ma J (2022) Piafusion: a progressive infrared and visible image fusion network based on illumination aware. Inf Fusion
21. Wang Z, Ma Y, Zhang Y (2023) Review of pixel-level remote sensing image fusion based on deep learning. Inf Fusion 90:36–58
22. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP (2004) Image quality assessment: from error visibility to structural similarity. IEEE Trans Image Process 13(4):600–612
23. Woo S, Park J, Lee JY, Kweon IS (2018) Cbam: convolutional block attention module. In: Proceedings of the European conference on computer vision (ECCV), pp 3–19
24. Xu H, Gong M, Tian X, Huang J, Ma J (2022) Cufd: an encoder–decoder network for visible and infrared image fusion based on common and unique feature decomposition. Comput Vis Image Underst 103407 (2022)
25. Xu H, Ma J, Jiang J, Guo X, Ling H (2020) U2fusion: a unified unsupervised image fusion network. IEEE Trans Pattern Anal Mach Intell

26. Xu H, Zhang H, Ma J (2021) Classification saliency-based rule for visible and infrared image fusion. IEEE Trans Comput Imaging 7:824–836
27. Yan X, Qin H, Li J, Zhou H, Zong JG (2015) Infrared and visible image fusion with spectral graph wavelet transform. JOSA A 32(9):1643–1652
28. Zhang H, Ma X, Tian Y (2020) An image fusion method based on curvelet transform and guided filter enhancement. Math Probl Eng 2020
29. Zhang Y, Liu Y, Sun P, Yan H, Zhao X, Zhang L (2020) Ifcnn: a general image fusion framework based on convolutional neural network. Inf Fusion 54:99–118
30. Zhang Y, Zhang L, Bai X, Zhang L (2017) Infrared and visual image fusion through infrared feature extraction and visual information preservation. Infrared Phys Technol 83:227–237
31. Zhao Z, Bai H, Zhang J, Zhang Y, Xu S, Lin Z, Timofte R, Van Gool L (2023) Cddfuse: correlation-driven dual-branch feature decomposition for multi-modality image fusion. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 5906–5916
32. Zhao Z, Xu S, Zhang C, Liu J, Li P, Zhang J (2020) Didfuse: deep image decomposition for infrared and visible image fusion. arXiv preprint arXiv:2003.09210
33. Zhou Z, Dong M, Xie X, Gao Z (2016) Fusion of infrared and visible images for night-vision context enhancement. Appl Opt 55(23):6480–6490