Optimizing Nighttime Infrared and Visible Image Fusion for Long-haul Tactile Internet

Wenhao Song, Mingliang Gao, *Member, IEEE*, Qilei Li, *Graduate Student Member, IEEE*, Xiangyu Guo, Zenghui Wang, and Gwanggil Jeon

***Abstract*—In the domain of infrared and visible image fusion, the majority of existing methods are designed for infrared and visible images with normal illumination conditions. However, these methods may not effectively address the challenges presented by long-haul transmission scenarios in the Tactile Internet. To meet the requirements of nighttime infrared and visible image fusion in long-haul network architectures for the tactile internet, an illumination component adjusting network (ICANet) is built. Firstly, an illumination adjustment denoising subnetwork (IADSubNet) is designed to enhance the illumination component of nighttime visible images and simultaneously eliminate noise. Secondly, a local-global perception fusion subnetwork (LGPFSubNet) is built to dynamically extract and fuse both global and local information of the source images. Furthermore, we leverage a mutual consistency loss to generate fused images that are both visually appealing and rich in information. This ensures the fidelity and consistency of the fused images during long-distance transmission. Comprehensive experimental results demonstrate that the proposed method outperforms state-of-the-art (SOTA) methods quantitatively and qualitatively, and prove that it has potential for high performance in the long-haul transmission scenarios of tactile Internet. Meanwhile, the fused images generated by the ICANet significantly enhance object detection tasks. It is a critical aspect for many tactile Internet applications dependent on real-time and accurate object recognition.**

***Index Terms*—Long-haul tactile Internet, deep learning, image fusion, transformer, Retinex theory.**

I. Introduction

THE RAPID advancement in infrared and visible imaging technology has opened up significant opportunities for applications across diverse fields. Infrared (IR) images excel in detecting and capturing thermal radiation information from targets, and they provide advantages in complex imaging conditions. In contrast, visible (VIS) images offer higher resolution, richer color information, and more detailed morphological features. The Tactile Internet is designed to enable ultra-low-latency and highly reliable communication between humans, machines, and the physical environment. It is focused on establishing real-time interactions with minimal data transmission delays, thereby extending the capabilities of the conventional Internet. In addition, infrared and visible image fusion enhances perception in the tactile Internet by integrating data from various sensors. Therefore, image fusion can contribute to improved immersion and response speed in the tactile Internet. However, the IR image is marred by issues such as texture blurring and an absence of color information. In addition, the VIS image often underperforms in low-light or nocturnal environments due to inadequate illumination. Under such circumstances, VIS images exhibit significant noise, diminished contrast, and limited visibility of objects.

In recent years, research has increasingly focused on advancing and refining infrared and visible image fusion (IVF) methods [1], [2], [3]. These methods aim to address the limitations of individual images and to enhance their utility in downstream computer vision tasks, such as object detection [4], [5]. By integrating the information from both infrared and visible images, fused images can provide a more comprehensive and accurate representation of target information. The IVF technology has found wide-ranging applications in fields such as military surveillance [6], object detection [7], and vehicle navigation [8].

In the domain of IVF, numerous techniques have been introduced, and they can be broadly classified into two primary categories, namely traditional methods and deep learning-based methods. Traditional methods [9], [10] typically treat IVF as a problem of feature representation. Initially, specific transforms are employed to extract crucial features from the source images. Subsequently, fusion strategies are employed to integrate these features. Finally, the fused image is reconstructed by applying corresponding inverse transformations. Nevertheless, with traditional methods continue to evolve, the increasing complexity of their transformation techniques makes real-time computation on modern computer systems increasingly challenging [11].

Recently, the field of computer vision has undergone significant advancements, largely driven by the swift development of deep learning techniques. These advancements have introduced more effective methods in IVF [12]. Deep learning-based methods can be broadly classified into three main

Manuscript received 3 August 2023; revised 7 October 2023 and 25 December 2023; accepted 5 February 2024. Date of publication 19 February 2024; date of current version 26 April 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62101310. *(Corresponding authors: Mingliang Gao; Gwanggil Jeon.)*

Wenhao Song, Mingliang Gao, Xiangyu Guo, and Zenghui Wang are with the School of Electrical and Electronic Engineering, Shandong University of Technology, Zibo 255000, China (e-mail: sdut_songwenhao@163.com; mlgao@sdut.edu.cn; xiangyvguo@163.com; sdut_zenghuiwang@163.com).

Qilei Li is with the School of Electronic Engineering and Computer Science, Queen Mary University of London, E1 4NS London, U.K. (e-mail: qilei.li@outlook.com).

Gwanggil Jeon is with the School of Electrical and Electronic Engineering, Shandong University of Technology, Zibo 255000, China, and also with the Department of Embedded Systems Engineering, Incheon National University, Incheon 22012, South Korea (e-mail: ggjeon@gmail.com).

Digital Object Identifier 10.1109/TCE.2024.3367667

categories according to their model architectures, namely convolutional neural network (CNN) based methods [13], [14], auto-encoder (AE) based methods [2], [15], and generative adversarial network (GAN) based methods [16], [17]. CNN-based methods utilize multiple parallel CNNs to extract features from the source images. Subsequently, the extracted features are fused to achieve end-to-end image fusion. CNN-based approaches effectively capture complementary information from both images and result in high-quality fused images. AE-based methods utilize auto-encoders for feature extraction and reconstruction. These methods then employ dedicated fusion strategies to effectively combine the extracted features for the final fused image. GAN-based methods consist of two critical components, *i.e.*, generator network and discriminator network. The generator network inputs both infrared and visible images and produces a fused image as its output. Meanwhile, the discriminator network plays a crucial role by providing feedback to the generator. The function of the discriminator is to constrain the distribution of the generated images, that resemble the characteristics of the source images as closely as possible.

Although deep learning has shown remarkable success in image fusion tasks, most existing approaches struggle in nighttime or low-light conditions. This necessitates the reliance on infrared information in many image fusion frameworks to counteract the deterioration of visual details in visible images caused by poor lighting. However, this dependence hinders their generalizability and negatively impacts overall fusion performance. On the other hand, convolutional neural networks are constrained by fixed-size local receptive fields and weight sharing during feature extraction. This limitation can lead to a loss of both texture details and global contextual information in the fused images. Furthermore, the information importance within the image can vary across different scenes. Nevertheless, existing methods often rely on fixed weights or simplistic weighting strategies to fuse the features, and they fail to adequately address the problem of balancing information from different modalities. This shortcoming can ultimately lead to information loss and suboptimal fusion outcomes.

To tackle these challenges, we propose a method that decomposes the image fusion task under extreme conditions into two sub-problems, namely image enhancement and image fusion. The proposed method effectively integrates both tasks while mitigating noise and artifacts, thereby ensuring compatibility between image enhancement and image fusion. Specifically, a Retinex-based illumination-adjustment and denoising subnetwork (IADSubNet) is designed. This subnetwork is used to decompose the mixed features at the feature level, simultaneously generating enhanced source images. Notably, IADSubNet possesses the ability to fine-tune the weights. This capability allows for control of the degree of enhancement for images under varying illumination conditions. Subsequently, a local-global perceptive fusion subnetwork (LGPFSubNet) is constructed. It comprises the local feature extraction module, the global feature extraction module, and the lightweight adaptive feature fusion module. This subnetwork serves feature extraction, fusion, and reconstruction. The local feature extraction module employs a CNN, while the global feature extraction module utilizes a Transformer architecture with self-attention mechanisms for comprehensive global feature extraction. Combining the advantages of local and global feature extraction enables the model to comprehend information from diverse dimensions and levels. The lightweight adaptive feature fusion module effectively fuses information from both images by dynamically combining their feature representations. Furthermore, the lightweight adaptive feature fusion module employs a lightweight design to ensure computational efficiency and real-time performance. We further introduced mutual consistency loss in both IADSubNet and LGPFSubNet. This loss function ensures that the generated images through the proposed method exhibit smoother variations and consistent structures. In summary, this study mainly contributes as follows,

- A joint network is proposed to enhance the visual perception of infrared and visible images in extreme conditions. This network effectively leverages the complementary information from the source image.
- An IADSubNet is built to enhance the degraded illumination component and the unique features of both the infrared and visible images. Additionally, an LGPFSubNet is developed to effectively utilize both local and global features and dynamically fuse the complementary information.
- A mutual consistency loss is introduced to mitigate color distortion and structural inconsistencies and improve the quality of the fused image.
- Extensive experiments demonstrate that the proposed method can adaptively fuse complementary information based on illumination conditions and generate fused images with brighter scenes and richer details.

The rest of this paper is arranged as follows. Section II reviews the related work on infrared and visible image fusion methods and retinex-based low-light image enhancement methods. Section III describes the proposed method in detail. Section IV presents the experimental setup and results. Section V concludes this paper and discusses future work.

II. Related Work

In this section, we initially introduce the existing infrared and visible image fusion (IVF) methods. We categorize these methods into four main groups, namely traditional IVF methods, CNN-based IVF methods, AE-based IVF methods, and GAN-based IVF methods. Subsequently, we review image enhancement techniques based on the Retinex theory [18].

A. IVF Methods

1) Traditional IVF Methods: Traditional IVF methods primarily focus on extracting the intrinsic features of both images for feature representation and applying specific fusion rules to combine them [19]. For example, Li et al. [20] introduced a multi-scale transformation and norm optimization technique to enhance the quality of fused images. Sparse representation-based fusion methods have also been widely explored. Wang et al. [21] introduced an image fusion

method based on sparse representation and geometric dictionary learning. This approach exploits the sparsity inherent in image representations to preserve information during fusion. Ma et al. [22] proposed a multi-scale IVF method that relies on visual saliency maps and weighted least squares optimization. This method retains specific scale information through a multi-scale decomposition method. Mou et al. [10] established an IVF method based on the principles of the non-negative matrix factorization and infrared target extraction. Li et al. [23] proposed a multi-focus image fusion method based on sparse feature matrix decomposition and morphological filtering. These traditional methods have made significant contributions to the field of infrared and visible image fusion. Nevertheless, they often face challenges in capturing differences between modalities and preserving fine details in complex scenes [12].

2) CNN-Based IVF Methods: CNN-based methods can automatically extract image features, fuse multi-modal features, and reconstruct images through an end-to-end network with carefully designed loss functions. Zhang et al. [24] proposed a network that separates information extraction into two distinct paths, namely the gradient path and the intensity path. They also formulated a unified loss function to guide the network to generate fused images directly. Tang et al. [25] integrated the image fusion and semantic segmentation tasks and incorporated a semantic loss to enhance the image fusion task performance. Sun et al. [26] proposed an image fusion method that enhances detection performance through task-driven image fusion. This method utilizes information from both infrared and visible images to improve the results of tasks such as object detection and tracking. PIAFusion [27] introduces an illumination-aware network to achieve more robust and effective image fusion results under extreme illumination conditions. Tang et al. [28] proposed a fusion method that injects semantic information and preserves scene fidelity. This network considers both semantic information and scene fidelity to ensure that the fusion results perform well in advanced tasks.

3) AE-Based IVF Methods: Numerous image fusion methods based on auto-encoders (AE) have been proposed due to their flexibility and interpretability. Most methods employ auto-encoders to extract features from the source images and reconstruct images. The feature fusion process mainly depends on manually designed fusion rules [29], [30]. Li and Wu [31] adopted an auto-encoder for feature extraction and fusion and incorporated dense connections within the encoder to extract deep features. They then proposed NestFuse [32] and RFN-Nest [15]. Specifically, NestFuse [32] introduces nested connections in the network to extract multi-scale features from source images. RFN-Nest [15] is designed with a loss function that preserves details and another that enhances features, compelling the network to acquire higher-quality integrated detail features. Xu et al. [29] proposed a learnable fusion rule that explores the interpretability of feature maps through saliency map-based feature fusion. Zhao et al. [33] proposed a model-based infrared and visible image fusion method (AUIF) that improves the efficiency and performance of image fusion while preserving accuracy based on the physical model.

4) GAN-Based IVF Methods: GANs can estimate probability distributions in an unsupervised fashion and enforce network constraints at a distributional level via adversarial loss. Therefore, they are suitable for unsupervised tasks such as image fusion [12]. FusionGAN [16] incorporates GAN into the image fusion process and eliminates the need for manually designing complex fusion rules. Nevertheless, using a single discriminator can result in imbalanced fusion outcomes. To address this issue, Ma et al. [17] introduced a Dual Discriminator Conditional Generative Adversarial Network (GAN), which discriminates the structural differences between the fused image and the two source images individually. Additionally, Xu et al. [34] incorporated a self-attention mechanism into the GAN that enables attention-driven operations and captures long-range dependencies to mitigate the problem of local distortions. Le et al. [35] proposed an unsupervised continual-learning generative adversarial network (UIFGAN). It achieves unified image fusion without the need for a large amount of annotated data under supervision.

B. Retinex-Based Low-Light Image Enhancement

The Retinex theory is employed to explain and simulate the perception of brightness and color in the human visual system. It considers that the perceived brightness of each pixel in an image can be expressed as the multiplication of two components, namely the reflectance component and the illumination component. The reflectance component conveys the surface properties and color information of the object, and the illumination component represents the impact of ambient illumination, which is denoted as,

$$I = R \odot L, \tag{1}$$

where I is the input low-light image. R and L denote the reflectance, and illumination components of the image, respectively. The $\odot$ represents element-wise multiplication.

Many methods have been proposed to enhance low-light images based on the Retinex theory, which decomposes an image into reflectance and illumination components. They then modify the estimated illumination to recover the image quality. Guo et al. [36] proposed a structural prior model to refine the initial illumination map for better image enhancement. Hao et al. [37] applied Gaussian total variation as a regularization term to build a decomposition model that reduces noise and artifacts in the enhanced results.

With the development of deep learning, researchers have attempted to utilize convolutional neural networks to estimate reflectance and illumination maps [38], [39]. Lore et al. [40] proposed the LLNet, which utilized stacked sparse denoising autoencoders to enhance and denoise low-light images. Jiang et al. [41] proposed a generative adversarial network-based approach to create an unpaired mapping between low-light and normal-light images. This method addresses the issue of data dependency in low-light enhancement methods. Guo et al. [42] introduced a reference-free enhancement approach that transformed images using pixel value mapping curves to produce the final output.

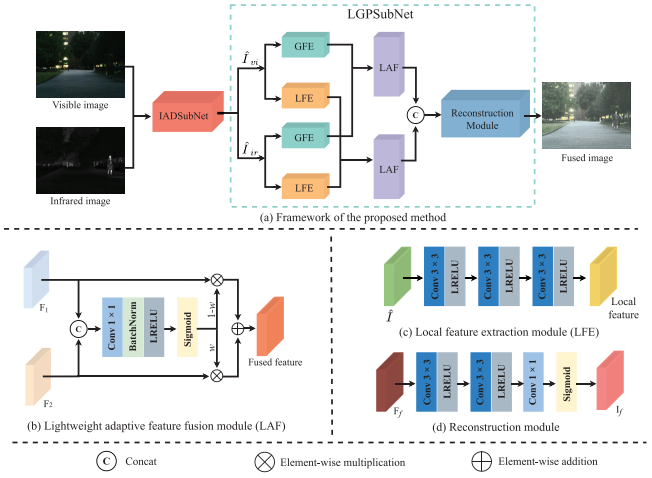


Fig. 1. Architecture of the proposed method. The sub-figure (b) provides a detailed description of the lightweight adaptive feature fusion module. Sub-figure (c) and (d) present the architecture of the local feature extraction and reconstruction modules, respectively.

III. Proposed Method

A. Overview

To address the incompatibility between image enhancement and fusion, we propose a novel framework consisting of two specialized subnetworks: the illumination-adjustment denoising subnetwork (IADSubNet) and the local-global perceptive fusion subnetwork (LGPFSubNet). The framework of the network is shown in Fig. 1. These subnetworks collaborate to maximize compatibility between the two tasks, leading to improved fusion outcomes. Specifically, the visible image is initially transformed into the YCbCr color space [43]. Subsequently, the Y channel of the visible image and the infrared image are concatenated along the channel dimension and fed into IADSubNet. This subnetwork decomposes and enhances the input, and it generates enhanced infrared and visible images as outputs, which is defined as,

$$\left\{\widehat{I}_{ir}, \widehat{I}_{vi}\right\} = \text{IADSubNet}\left(I_{vi}^{Y}, I_{ir}\right), \tag{2}$$

where I_{ir} and I_{vi}^{Y} denote the infrared and the Y channel of the visible image respectively. The enhanced infrared and visible images are represented as $\widehat{I}_{ir}$ and $\widehat{I}_{vi}$, respectively.

The enhanced infrared and visible images are then fed into LGPFSubNet, which performs feature extraction, fusion, and image reconstruction to produce the fused image I_f^Y. The process can be formulated as,

$$I_f^Y = \text{LGPFSubNet}\left(\widehat{I}_{ir}, \widehat{I}_{vi}\right), \tag{3}$$

Finally, to obtain the fused color image I_f, we concatenate I_f^Y, Cb, and Cr along the channel dimension, and then convert them from the $YCbCr$ domain to the RGB domain. This process can be expressed as,

$$I_f = \mathcal{H}\left(\text{concat}\left(I_f^Y, I_{vi}^{Cb}, I_{vi}^{Cr}\right)\right), \tag{4}$$

where $\mathcal{H}(\cdot)$ represents the conversion of the image from the $YCbCr$ colour space to the RGB colour space. I_{vi}^{Cb} and I_{vi}^{Cr} indicates the Cb and Cr channels of the visible images.

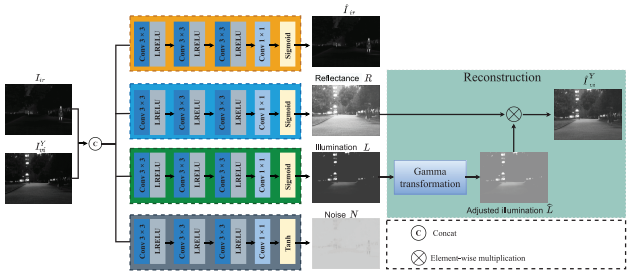


Fig. 2. Architecture of the illumination-adjustment denoising subnetwork.

B. Illumination-Adjustment Denoising Subnetwork

The architecture of the illumination-adjustment denoising subnetwork is illustrated in Fig. 2. This subnetwork is designed based on the robust Retinex model [44]. Specifically, the mixed features of infrared and visible are first decomposed into four components, namely enhanced infrared features $\widehat{I}_{ir}$, reflectance R, illumination L, and noise N as follows:

$$\text{concat}\left(I_{vi}^{Y}, I_{ir}\right) = R \odot L + N + \widehat{I}_{ir}. \tag{5}$$

The IADSubNet is a fully convolutional neural network that employs identical convolutional layers to create four branches. It comprises three 3×3 convolutional layers and a 1×1 convolutional layer. All layers except the last one use LReLU activation. The last layer of the noise branch differs from the others. To effectively handle additive noise, the noise branch employs a tanh layer as its final layer and keeps noise values within the range $[-1, 1]$. The other branches utilize sigmoid layers.

In the reconstruction stage, the illumination component is adjusted using a Gamma transformation, which can be represented as,

$$\widehat{L} = L^{\gamma}, \tag{6}$$

where γ is the adjustable parameter in the gamma transformation. The final restoration result combines the adjusted illumination and noise-free reflectance,

$$\widehat{I}_{vi}^{Y} = R \odot \widehat{L}. \tag{7}$$

C. Local-Global Perceptive Fusion Subnetwork

The architecture of the local-global perception subnetwork is depicted in Fig. 1. It consists of four modules *i.e.*, the local feature extraction module, the global feature extraction module, the lightweight adaptive feature fusion module, and the reconstruction module. The local feature extraction module consists of 3×3 convolutional layers with LReLU activation functions.

The specific architecture of the global feature extraction module is illustrated in Fig. 3. This module combines multi-layer perception and self-attention mechanisms to extract global image features. The multi-layer perception effectively stacks multiple 3×3 convolutional layers with GELU non-linear activation functions to extract high-level feature representations from the input image. These convolutional layers effectively capture both local and global features of the

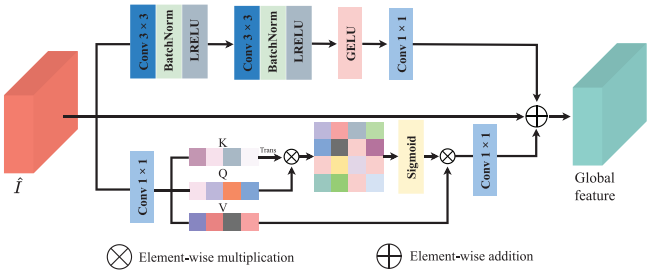


Fig. 3. Architecture of the global feature extraction module.

input image. It allows the model to comprehend information from various dimensions and levels within the image.

The self-attention mechanism enables the model to learn informative feature representations by attending to the relationships between different spatial locations within the input image. Thereby, it can enhance the ability of the model to capture long-range dependencies. To reduce computational complexity while emphasizing crucial features, a 1×1 convolutional layer is employed to reduce the input dimension. The self-attention mechanism process involves splitting the input into query (Q), key (K), and value (V) components, followed by computing attention weights that measure the relevance of each key to each query. These weights are then used to aggregate the corresponding values and generate an attention representation. Finally, the projection layer uses the 1×1 convolutional maps of the features to generate the output with the original dimension. The input image passes through self-attention and multilayer perceptron layers, and the original input is added to get the final output. This residual connection helps preserve the input information and mitigate the vanishing gradients issue.

The architecture of the lightweight adaptive feature fusion module is shown in Fig. 1. This module has a lightweight architecture with a convolutional layer, batch normalization, and LReLU activation function. The lightweight adaptive feature fusion module initially concatenates information from two input feature maps. Subsequently, it uses a 1×1 convolutional layer to transform the input features into a lower-dimensional space. This transformation is followed by batch normalization to standardize the channel-wise activations, and the application of the LReLU activation function to introduce non-linearity. Furthermore, a Sigmoid function is utilized to scale the features within the range [0, 1], and generates channel weights that represent the relevance of each input feature map. This enables the model to selectively emphasize or suppress information from different input sources during the fusion process. Finally, it concatenates the fused local and global features along the channel dimension and feeds them into the reconstruction module. The reconstruction module consists of 3×3 and 1×1 convolutional layers and a sigmoid function, which produce the fused image.

D. Loss Function

1) Illumination-Adjustment Denoising Loss: To update the weights of the IADSubNet, it is necessary to employ a loss function that ensures the network generates more accurate components. Consequently, we have designed a loss function $\mathcal{L}_{\text{IAD}}$ consisting of five parts. It can be formulated as,

$$\begin{aligned}\mathcal{L}_{\text{IAD}} &= \lambda_1 \mathcal{L}^{\text{ir}}_{\text{recon}} + \lambda_2 \mathcal{L}^{vi}_{\text{recon}} + \lambda_3 \mathcal{L}^{\text{illu}}_{\text{smooth}} \\ &\quad + \lambda_4 \mathcal{L}_{mc} + \lambda_5 \mathcal{L}_{\text{noise}}\,, \end{aligned} \tag{8}$$

where $\mathcal{L}^{\text{ir}}_{\text{recon}}$ and $\mathcal{L}^{vi}_{\text{recon}}$ are the reconstruction losses for infrared and visible images, respectively. $\mathcal{L}^{\text{illu}}_{\text{smooth}}$ and $\mathcal{L}_{mc}$ represent the illumination smoothness loss and the mutual consistency loss, respectively. $\mathcal{L}_{\text{noise}}$ denote the noise estimation loss $\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5$ stands for the corresponding balance weight factors.

We introduce the infrared and visible image reconstruction losses to preserve high-fidelity information from the original images in the IADSubNet. The reconstruction losses are defined as follows,

$$\mathcal{L}^{ir}_{\text{recon}} = \left\| I_{ir} - \widehat{I}_{ir} \right\|_1, \tag{9}$$

$$\mathcal{L}^{vi}_{\text{recon}} = \left\| I^{Y}_{vi} - (R \odot L + N) \right\|_1, \tag{10}$$

where $\|\cdot\|_1$ denote the l_1-norm.

We further incorporate illumination smoothness and mutual consistency losses to generate consistent and seamless illumination components. These losses are inspired by DIVFusion [45] and formulated as follows,

$$\mathcal{L}^{\text{illu}}_{\text{smooth}} = \left\| \frac{\nabla L}{\max\left(\left|\nabla I^{Y}_{vi}, \varepsilon\right|\right)} \right\|_1, \tag{11}$$

$$\mathcal{L}_{mc} = \left\| \nabla L \odot \exp(-c \odot \nabla L) \right\|_1, \tag{12}$$

where ∇ denotes the gradient operation, and ε is a positive constant. We apply the maximum operation to constrain the minimum value of the denominator to be ε (0.01 in this work). The parameter c (set as 10 in this work) plays a crucial role in shaping the mutual consistency loss function. By setting c, we use the mutual consistency loss to enhance the medium-gradient parts of the images.

When enhancing low-light scenes, the noise concealed within the shadowed areas may also be amplified. To suppress the amplified noise more accurately, we adopt a noise estimation loss, which is formulated as,

$$\mathcal{L}_{noise} = \left\| L \odot N \right\|_2, \tag{13}$$

where $\|\cdot\|_2$ denotes the l_2-norm. The noise estimation loss uses the illumination component as a weight to constrain the noise component. It facilitates the model in separating the noise component more effectively.

2) Local-Global Perceptive Fusion Loss: To improve the fusion performance, we employ three loss functions namely, texture loss, pixel loss, and mutual consistency loss. The local-global perceptive fusion loss is represented as,

$$\mathcal{L}_{\text{LGP}} = \alpha_1 \mathcal{L}_{\text{edge}} + \alpha_2 \mathcal{L}_{\text{pix}} + \alpha_3 \mathcal{L}_{\text{fmc}}, \tag{14}$$

where $\mathcal{L}_{\text{edge}}$ represents the edge loss, which helps the fused image retain more edge detail information. $\mathcal{L}_{\text{pix}}$ is the pixel loss that aims to preserve prominent target information from the infrared image. $\mathcal{L}_{\text{fmc}}$ represents the mutual consistency loss that enhances the gradient consistency and visual quality of the fused image.

The edge loss is utilized to preserve the distinctive texture details of the enhanced infrared and visible images. It can be formulated as follows,

$$\mathcal{L}_{\text{edge}} = \left\| \nabla I_f^Y - \max\left(\nabla \widehat{I}_{ir}, \nabla \widehat{I}_{vi}\right) \right\|_1. \tag{15}$$

Similarly, pixel loss is a pixel-level constraint that measures the difference between the fused image and the source image. It can be represented as follows,

$$\mathcal{L}_{\text{pix}} = \|I_f^Y - \widehat{I}_{ir}\|_1. \tag{16}$$

The fused images often suffer from inconsistent gradient distributions, resulting in discontinuous edges or blurry details. To address this issue, we use a mutual consistency loss function on the fused image that minimizes the overall consistency of gradients within the fused image, thereby preserving critical details while mitigating high-gradient regions. This function enhances the visual quality of the fused image and is defined as,

$$\mathcal{L}_{fmc} = \|\nabla I_f^Y \odot \exp\left(-c \odot \nabla I_f^Y\right)\|_1, \tag{17}$$

where c is a parameter set to 10.

IV. Experimental Results and Analysis

A. Experimental Configurations

The MSRS [27] dataset is used to evaluate the performance of the ICANet. The MSRS dataset includes high-resolution infrared and visible images. It covers various object types in different scenes, and these images have been pre-aligned. We also use the LLVIP [46] and RoadScene [47] datasets to illustrate the generalization ability of our method. These datasets consist of infrared and visible images covering various scenes and lighting conditions. The LLVIP dataset contains images captured in extremely dark environments, and these images are rigorously aligned in both time and space. The RoadScene dataset contains a rich collection of scenes, such as roads, vehicles, and pedestrians. This dataset has been preprocessed to remove background thermal noise from the original infrared images and to align the infrared and visible images precisely. The illumination-adjustment denoising and Local-global perceptive fusion subnetworks are trained on the MSRS dataset. To evaluate the effectiveness of ICANet, a total of 185 image pairs from the MSRS datasets are utilized. In addition, 50 image pairs from the LLVIP dataset and 40 image pairs from the RoadScene dataset were selected to evaluate the generalization of the ICANet.

In this work, We compare our method with nine SOTA methods, including CSF [29], CUFD [48], IFCNN [14], PIAFusion [27], FusionGAN [16], U2Fusion [49], UMF-CMGR [50], SDNet [51], and RFN-Nest [15]. We implemented all comparison methods using their open-source code and set the parameters as reported in the original papers. To evaluate the fusion performance quantitatively, five metrics are employed as objective measures. These metrics encompass entropy (EN), spatial frequency (SF), average gradient (AG), standard deviation (SD), and visual information fidelity (VIF). EN quantifies the complexity and amount of information in the image. SF describes the variations in different scales and frequencies in the image. It indicates the sharpness, clarity, and fine details of the image. AG measures the gradient information of the fused images by the intensity changes across adjacent pixels. SD quantifies the distribution and contrast of the fused images. It reflects the statistical quality of the image. VIF measures the fidelity of the fused image in terms of human visual perception. A fusion method with higher values in EN, SF, AG, SD, and VIF has superior fusion performance.

B. Training Details

The IADSubNet and LGPFSubNet are trained on the MSRS dataset. To enrich the training data, we cropped the images into multiple pairs of image patches. All images are preprocessed by normalization and scaling to the range [0, 1] before being fed into the subnetworks.

We adopted a two-stage training approach, where each subnetwork is trained separately. In the first stage, the IADSubNet is trained with a batch size of 128, and the learning rate is initialized as 0.0001. The training epoch is set as 100. Subsequently, the output images from the IADSubNet are utilized as the input for LGPFSubNet. In the second stage, a batch size of 64 is set, and the learning rate is adjusted to 0.001. The LGPFSubNet training epoch is set to 30. We employed the Adam [52] optimizer to update the parameters in both subnetworks. The hyperparameters in the loss functions Eq. (8) λ_1, λ_2, λ_3, λ_4, and λ_5 are set to 1000, 2000, 7, 9 and 1, respectively. The α_1, α_2, and α_3 in the Eq. (14) are seated as 200, 0.1, and 1.1, respectively. The γ in Eq. (6) is set to 0.4. Both subnetworks are implemented using the PyTorch framework and trained on the NVIDIA GeForce RTX 3090 Ti GPU.

C. Fusion Performance Analysis

To comprehensively evaluate the performance of ICANet, we conducted a comparative analysis with nine SOTA methods on the MSRS dataset.

1) Qualitative Evaluation: The qualitative evaluation measures the level of satisfaction that humans have with an image. A superior low-light fusion algorithm should extract and enhance valuable information from the source images, and produce a scene with high contrast and well-illuminated details for images degraded by low-light conditions. We selected a pair of typical infrared and visible images to demonstrate the fusion performance of different algorithms on the MSRS dataset visually in Fig. 4. It shows that the results of PIAFusion and UMF-CMGR preserve some details from the visible image, but they have an overall dark appearance and are not visually pleasing. Additionally, other comparative methods suffer from significant loss of details from the visible image. In contrast, the proposed fusion method enhances the texture information from the visible image and leads to a brighter scene and a high-contrast fused image. The ICANet also balances the intensity information from the infrared image and the texture information from the visible image effectively.

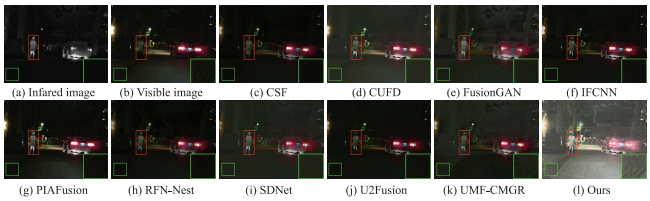


Fig. 4. Visualized results of different methods on the MSRS dataset.

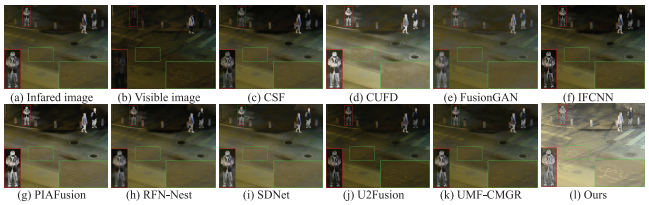


Fig. 5. Visualized results of different methods on the LLVIP dataset.

TABLE I
QUANTITATIVE COMPARISONS ON 182 PAIRS OF IMAGES FROM THE MSRS DATASET. THE DISPLAYED VALUES REPRESENT THE MEAN ± STANDARD DEVIATION. (THE BEST, SECOND-BEST, AND THIRD-BEST RESULTS ARE MARKED IN RED, BLUE, AND GREEN, RESPECTIVELY

	EN ↑	SF ↑	AG ↑	SD ↑	VIF ↑
CSF [29]	5.104±0.613	5.180±1.122	1.480±0.397	20.411±6.406	0.617±0.091
CUFD [48]	5.209±0.567	6.668±1.219	1.808±0.413	24.227±7.807	0.571±0.085
IFCNN [14]	5.281±0.506	9.075±1.883	2.631±0.691	22.830±7.025	0.683±0.050
PIAFusion [27]	5.960±0.600	9.240±1.680	2.831±0.662	34.998±9.840	1.063±0.074
FusionGAN [16]	5.270±0.424	3.715±0.754	1.198±0.269	15.713±4.371	0.493±0.114
U2Fusion [49]	4.518±0.675	6.442±1.464	1.609±0.502	19.599±6.445	0.493±0.055
UMF-CMGR [50]	5.060±0.328	5.405±1.200	1.534±0.357	15.297±5.175	0.336±0.071
SDNet [51]	4.833±0.441	6.443±1.258	1.899±0.446	14.021±4.363	0.433±0.073
RFN-Nest [15]	5.573±0.563	4.678±1.133	1.386±0.364	23.169±6.751	0.667±0.128
Ours	6.853±0.416	11.478±1.553	3.908±0.639	42.153±8.159	1.101±0.830

TABLE II
QUANTITATIVE COMPARISONS ON 50 PAIRS OF IMAGES FROM THE LLVIP DATASET. THE DISPLAYED VALUES REPRESENT THE MEAN ± STANDARD DEVIATION. (THE BEST, SECOND-BEST, AND THIRD-BEST RESULTS ARE MARKED IN RED, BLUE, AND GREEN, RESPECTIVELY)

	EN ↑	SF ↑	AG ↑	SD ↑	VIF ↑
CSF [29]	6.698±0.383	8.21±4.129	2.575±1.534	32.502±9.808	0.695±0.113
CUFD [48]	6.681±0.313	7.234±3.39	2.03±1.246	33.846±4.771	0.627±0.141
IFCNN [14]	6.734±0.466	14.408±8.599	4.184±2.906	34.84±12.336	0.762±0.084
PIAFusion [27]	7.178±0.327	14.685±8.537	4.312±2.906	44.701±10.499	0.952±0.109
FusionGAN [16]	6.308±0.312	6.921±3.059	1.947±1.042	24.822±4.483	0.476±0.079
U2Fusion [49]	6.360±0.581	11.047±6.177	3.289±2.243	31.219±11.383	0.674±0.091
UMF-CMGR [50]	6.462±0.501	9.916±6.569	2.504±1.895	29.38±10.066	0.521±0.070
SDNet [51]	6.680±0.338	12.098±7.258	3.439±2.361	31.802±6.250	0.635±0.078
RFN-Nest [15]	6.820±0.331	6.321±2.917	2.158±1.155	34.646±8.666	0.669±0.131
Ours	7.303±0.289	19.149±8.883	6.068±3.232	43.982±8.504	1.076±0.302

2) Quantitative Evaluation: To further validate the effectiveness of ICANet, we selected 182 pairs of images from the MSRS dataset for quantitative comparisons. As illustrated in Table I, the results demonstrate that the ICANet outperforms other methods in all five metrics. The highest EN metric indicates that the results of the ICANet with higher information content and better visual effects. The highest SF and AG metrics indicate that ICANet preserves more texture information. The highest SD metric suggests that the fusion results of the proposed method have a higher contrast. The synergistic effect of image enhancement and fusion in ICANet leads to higher VIF metrics than other SOTA methods. In summary, the ICANet can effectively extract valuable information in low-light scenes. It leverages IADSubNet to enhance the illumination component and reduce noise, which is then integrated into the fused image. Consequently, the proposed method exhibits clear advantages over other SOTA methods.

D. Generalization Evaluation

In the domain of deep learning, the generalization ability of a model is a crucial metric. Generalization ability refers to the performance of a model on unseen data. It indicates the adaptability of the model to new samples. To evaluate the generalization performance of the ICANet, we conducted evaluations on the LLVIP and RoadSence datasets. It is essential to mention that the ICANet is trained on the MSRS dataset and subsequently evaluated directly on the LLVIP and RoadSence datasets.

1) Generalization Evaluation on the LLVIP Dataset: Fig. 5 illustrates a pair of typical infrared and visible images captured under nighttime conditions. The fusion results of the proposed method reveal a bright scene with salient objects, uncovering the hidden details in dark areas. In the red box, the ICANet leverages salient information from the infrared image and improves the visibility of the targets by enhancing the contrast. In the green box, only the ICANet preserves crucial scene details, while other methods fail to provide information for this region. This observation indicates that ICANet improves the fusion performance by extracting information in the dark regions through IADSubNet.

Table II presents the results of quantitative experiments on the LLVIP dataset. We compare the performance of ICANet with nine SOTA methods. The results demonstrate that ICANet achieves the highest scores in EN, SF, AG, and VIF metrics. The higher scores in EN, SF, and AG indicate that ICANet effectively preserves texture information in the source images. The superior VIF metric implies that ICANet successfully transfers more information into the fused image. For the SD metric, the ICANet ranks second, only behind the PIAFusion. This high ranking indicates that the proposed fusion method maintains visual similarity with the source images, and highlights its potential for accurate and reliable fusion results.

2) Generalization Evaluation on the RoadScene Dataset: We further employed the RoadScene dataset to validate the generalization ability of ICANet with qualitative and quantitative experiments. The RoadScene dataset consists of grayscale images, and the infrared and visible images are mostly acquired under normal lighting conditions. Therefore, the full potential of the ICANet cannot be fully utilized. However, the ICANet still demonstrates comparable performance to SOTA methods. As shown in Fig. 6, this example illustrates a representative pair of images from the RoadScene dataset. The fusion results of CSF, CUFD, FusionGAN, RFN-Nest, and UMF-CMGR exhibit noticeable blurriness and significant loss of scene information. IFCNN, PIAFusion, SDNet, and U2Fusion preserve more texture information, but have low scene contrast and poor information balancing from the infrared and visible images. In contrast, the ICANet not only achieves superior scene contrast but also retains rich texture details.

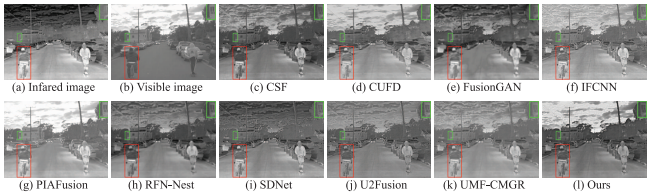


Fig. 6. Visualized results of different methods on the RoadScene dataset.

TABLE III
QUANTITATIVE COMPARISONS ON 40 PAIRS OF IMAGES FROM THE ROADSCENE DATASET. THE DISPLAYED VALUES REPRESENT THE MEAN ± STANDARD DEVIATION. (THE BEST, SECOND-BEST, AND THIRD-BEST RESULTS ARE MARKED IN RED, BLUE, AND GREEN, RESPECTIVELY)

	EN ↑	SF ↑	AG ↑	SD ↑	VIF ↑
CSF [29]	7.351±0.273	12.493±3.145	5.098±1.456	45.574±9.127	0.572±0.103
CUFD [48]	7.299±0.220	13.949±2.759	5.442±1.219	48.318±7.507	0.587±0.105
IFCNN [14]	7.202±0.287	15.991±4.167	6.373±1.735	40.518±8.371	0.568±0.109
PIAFusion [27]	6.949±0.227	12.262±2.967	4.410±1.263	42.007±5.609	0.663±0.094
FusionGAN [16]	7.017±0.252	8.126±1.455	3.205±0.648	37.486±6.243	0.364±0.058
U2Fusion [49]	7.140±0.303	14.705±3.648	6.004±1.611	38.666±8.520	0.537±0.104
UMF-CMGR [50]	6.995±0.349	10.461±2.923	4.089±1.229	35.583±8.655	0.584±0.108
SDNet [51]	7.294±0.247	15.112±3.624	6.119±1.631	43.505±8.096	0.577±0.103
RFN-Nest [15]	7.297±0.255	7.639±1.703	3.310±0.888	44.122±8.749	0.500±0.091
Ours	7.327±0.255	19.279±5.265	6.408±2.210	46.381±6.690	0.533±0.091

Table III summarizes the numerical results on the RoadScene dataset. The ICANet outperforms other methods in terms of AG and SF. This demonstrates that the fusion results of the ICANet exhibit high contrast. The ICANet also achieves the second-best performance on EN and SD, which indicates that the fused images of the ICANet contain richer detailed information. However, the ICANet shows suboptimal performance on the VIF metric. The challenge arises from the fact that the proposed method targets the adjustment of the image's illumination level to achieve a brighter scene, which might not be well-suited for grayscale images in the RoadScene dataset. Consequently, the method encounters difficulty in fine-tuning the image to closely align with human visual perception.

E. Application to Object Detection

This section validates the effectiveness of the ICANet for downstream object detection tasks using the fusion images generated by the proposed method. For the object detection task, the Yolov5 [53] is used to evaluate the performance of the source images and our fused images.

As shown in Fig. 7, the visible images have difficulties in capturing adequate information in low-light environments, which hinders the detection of pedestrians. In contrast, infrared images can capture thermal information. It helps emphasize prominent targets like pedestrians. However, infrared images lack detailed information about objects such as vehicles. This results in reduced detection accuracy for such objects. The proposed model effectively integrates meaningful information from the source images, thereby improving the detection accuracy of pedestrians and vehicles. Table IV presents the quantitative metrics for the object detection task. Precision represents the probability of correctly predicting positive samples among all samples predicted as positive. Higher precision means a higher rate of correctly detected positive samples. Recall is the probability of correctly predicting

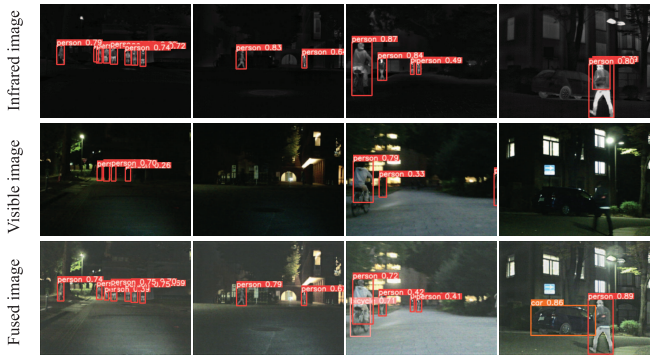


Fig. 7. Visual results of object detection on the MSRS dataset.

TABLE IV
OBJECT DETECTION QUALITATIVE EVALUATION OF INFRARED, VISIBLE, AND FUSED IMAGES ON THE MSRS DATASET. THE BEST AND SECOND-BEST RESULTS ARE HIGHLIGHTED IN RED AND BLUE, RESPECTIVELY

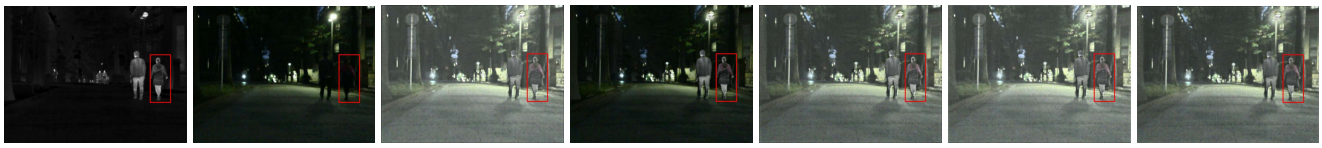
	Precision ↑	Recall ↑	mAP@0.50 ↑	mAP@[0.5:0.95] ↑
Infrared	0.88	0.686	0.812	0.569
Visible	0.913	0.765	0.804	0.533
Ours	0.929	0.747	0.875	0.604

positive samples among all positive samples. Higher recall indicates fewer missed detections. Another crucial detection performance indicator is the mean average precision (*mAP*), where a value closer to 1 signifies superior object detection. Specifically, *mAP*@0.50 represents the *mAP* value at a confidence threshold of 0.5, while *mAP*@[0.5:0.95] is the average of all *mAP* values at different *IoU* thresholds (ranging from 0.5 to 0.95 with an interval of 0.05). Table IV shows that the fused images produced by the proposed method achieve the best precision, *mAP*@0.50, and *mAP*@[0.5:0.95] values. The proposed method also achieves the second-highest recall.

In summary, ICANet effectively fuses the salient object information in infrared images and the texture structure information in visible light images through the collaborative action of IADSubNet and LGPFSubNet. It also improves the overall brightness of the fused image.The differentiation of the targets in the fused image is significantly improved, thus achieving superior performance in the target detection task.

F. Ablation Study

To evaluate the effectiveness of each component in the ICANet, we conduct an ablation evaluation on the four key components, *i.e.,* mutual consistency loss, illumination-adjustment denoising subnetwork, global feature extraction module, and lightweight adaptive feature fusion module. We present the quantitative results summarized in Table V and list the qualitative results in Fig. 8. Qualitative and quantitative experiments indicate that the removal of $\mathcal{L}_{\text{fmc}}$ from the proposed method results in increased noise and reduced smoothness in the fused image. Moreover, the EN, SF, AG, and SD indicators exhibit a decline. When the IADSubNet is removed, the visual results turn out to be obvious degradation. Removing IADSubNet results in considerable visual degradation and darkened imagery. For the validity of the



(a) Infared image (b) Visible image (c) W/O $\mathcal{L}_{\text{tex}}$ (d) W/O IADSubNet (e) W/O GFE (f) W/O LAF (g) Ours

Fig. 8. Visual results of ablation study.

TABLE V
QUANTITATIVE EVALUATION RESULTS OF ABLATION STUDY. THE BEST, SECOND-BEST, AND THIRD-BEST RESULTS ARE MARKED IN RED, BLUE, AND GREEN, RESPECTIVELY. "W/O" DENOTES "WITHOUT"

	EN ↑	SF ↑	AG ↑	SD ↑	VIF ↑
W/O $\mathcal{L}_{\text{fmc}}$	6.830±0.424	11.134±1.521	3.792±0.657	41.543±7.788	1.127±0.19
W/O IADSubNet	6.000±0.516	8.864±1.703	2.681±0.628	32.635±8.68	0.996±0.088
W/O GFE	6.835±0.429	11.385±1.608	3.923±0.649	41.642±8.161	1.084±0.195
W/O LAF	6.857±0.403	11.186±1.44	3.889±0.629	42.329±8.099	1.135±0.201
Ours	6.853±0.416	11.478±1.553	3.908±0.639	42.153±8.159	1.101±0.83

experiment, we replaced the GFE module with the local feature extraction module. When the global feature extraction module is replaced from the method, the fusion results lose a lot of detailed information, and in the quantitative experiments, the EN, SF, SD, and VIF indicators decrease significantly. The LAF module achieves a dynamic fusion of local and global information. The absence of the LAF module resulted in ghosting and color distortion issues, highlighting its role in balancing local and global information for dynamic fusion.

V. CONCLUSION

In this study, we proposed the illumination component adjusting network (ICANet) for infrared and visible image fusion. This network achieves the integration of image enhancement and image fusion tasks. Specifically, we initially constructed the IADSubNet to isolate and enhance the illumination, reflection, and noise components, thereby augmenting the scene information in images. Subsequently, we designed the LGPFSubNet for image fusion, which integrates local and global information through specialized extraction modules. A lightweight feature fusion module is then employed to dynamically fuse the features of the infrared and visible images. To enhance the fused image quality, we implemented mutual consistency loss. Experimental results demonstrate that the ICANet outperforms the SOTA methods in terms of performance qualitative and quantitative evaluations. Furthermore, the application of ICANet in object detection tasks underscores its effectiveness.

As part of future work, we intend to develop an adaptive illumination adjustment module tailored for fusing images with varying degradation levels. Additionally, we will explore methods to integrate semantic information into the fusion process. This investigation entails combining infrared images and visible images with semantic segmentation masks to augment the efficacy of downstream tasks, including object detection and tracking.

REFERENCES

[1] W. Song, W. Zhai, M. Gao, Q. Li, A. Chehri, and G. Jeon, "Multiscale aggregation and illumination-aware attention network for infrared and visible image fusion," *Concurr. Comput. Pract. Exp.*, Apr. 2023, Art. no. e7712.

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

[3] M. Gao, Y. Zhou, W. Zhai, S. Zeng, and Q. Li, "SaReGAN: A salient regional generative adversarial network for visible and infrared image fusion," *Multimedia Tools Appl.*, pp. 1–13, Jan. 2023.

[4] J. Liu et al., "Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 5802–5811.

[5] W. He, W. Feng, Y. Peng, Q. Chen, G. Gu, and Z. Miao, "Multi-level image fusion and enhancement for target detection," *Optik*, vol. 126, nos. 11–12, pp. 1203–1208, 2015.

[6] S. Das and Y. Zhang, "Color night vision for navigation and surveillance," *Transp. Res. Rec.*, vol. 1708, no. 1, pp. 40–46, 2000.

[7] X. Xu, G. Liu, D. P. Bavirisetti, X. Zhang, B. Sun, and G. Xiao, "Fast detection fusion network (FDFnet): An end to end object detection framework based on heterogeneous image fusion for power facility inspection," *IEEE Trans. Power Del.*, vol. 37, no. 6, pp. 4496–4505, Dec. 2022.

[8] G. Bhatnagar and Z. Liu, "A novel image fusion framework for night-vision navigation and surveillance," *Signal, Image Video Process.*, vol. 9, pp. 165–175, Jan. 2015.

[9] Z. Zhou, B. Wang, S. Li, and M. Dong, "Perceptual fusion of infrared and visible images through a hybrid multi-scale decomposition with gaussian and bilateral filters," *Inf. Fusion*, vol. 30, pp. 15–26, Jul. 2016.

[10] J. Mou, W. Gao, and Z. Song, "Image fusion based on non-negative matrix factorization and infrared feature extraction," in *Proc. 6th Int. Congr. Image Signal Process. (CISP)*, 2013, pp. 1046–1050.

[11] S. Li, X. Kang, L. Fang, J. Hu, and H. Yin, "Pixel-level image fusion: A survey of the state of the art," *Inf. Fusion*, vol. 33, pp. 100–112, Jan. 2017.

[12] H. Zhang, H. Xu, X. Tian, J. Jiang, and J. Ma, "Image fusion meets deep learning: A survey and perspective," *Inf. Fusion*, vol. 76, pp. 323–336, Dec. 2021.

[13] J. Ma, L. Tang, M. Xu, H. Zhang, and G. Xiao, "STDFusionNet: An infrared and visible image fusion network based on salient target detection," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–13, Apr. 2021.

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

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

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

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

[18] E. H. Land, "The retinex theory of color vision," *Sci. Am.*, vol. 237, no. 6, pp. 108–129, 1977.

[19] J. Ma, Y. Ma, and C. Li, "Infrared and visible image fusion methods and applications: A survey," *Inf. Fusion*, vol. 45, pp. 153–178, Jan. 2019.

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

[21] K. Wang, G. Qi, Z. Zhu, and Y. Chai, "A novel geometric dictionary construction approach for sparse representation based image fusion," *Entropy*, vol. 19, no. 7, p. 306, 2017.

[22] J. Ma, Z. Zhou, B. Wang, and H. Zong, "Infrared and visible image fusion based on visual saliency map and weighted least square optimization," *Infrared Phys. Technol.*, vol. 82, pp. 8–17, May 2017.

[23] H. Li, L. Li, and J. Zhang, "Multi-focus image fusion based on sparse feature matrix decomposition and morphological filtering," *Opt. Commun.*, vol. 342, pp. 1–11, May 2015.
[24] H. Zhang, H. Xu, Y. Xiao, X. Guo, and J. Ma, "Rethinking the image fusion: A fast unified image fusion network based on proportional maintenance of gradient and intensity," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 12797–12804.
[25] L. Tang, J. Yuan, and J. Ma, "Image fusion in the loop of high-level vision tasks: A semantic-aware real-time infrared and visible image fusion network," *Inf. Fusion*, vol. 82, pp. 28–42, Jun. 2022.
[26] Y. Sun, B. Cao, P. Zhu, and Q. Hu, "DetFusion: A detection-driven infrared and visible image fusion network," in *Proc. 30th ACM Int. Conf. Multimedia*, 2022, pp. 4003–4011.
[27] L. Tang, J. Yuan, H. Zhang, X. Jiang, and J. Ma, "PIAFusion: A progressive infrared and visible image fusion network based on illumination aware," *Inf. Fusion*, vols. 83–84, pp. 79–92, Jul. 2022.
[28] L. Tang, H. Zhang, H. Xu, and J. Ma, "Rethinking the necessity of image fusion in high-level vision tasks: A practical infrared and visible image fusion network based on progressive semantic injection and scene fidelity," *Inf. Fusion*, vol. 99, Nov. 2023, Art. no. 101870.
[29] H. Xu, H. Zhang, and J. Ma, "Classification saliency-based rule for visible and infrared image fusion," *IEEE Trans. Comput. Imag.*, vol. 7, pp. 824–836, Jul. 2021.
[30] J. Ma et al., "Infrared and visible image fusion via detail preserving adversarial learning," *Inf. Fusion*, vol. 54, pp. 85–98, Feb. 2020.
[31] H. Li and X.-J. Wu, "DenseFuse: A fusion approach to infrared and visible images," *IEEE Trans. Image Process.*, vol. 28, pp. 2614–2623, 2019.
[32] H. Li, X.-J. Wu, and T. Durrani, "NestFuse: An infrared and visible image fusion architecture based on nest connection and spatial/channel attention models," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 12, pp. 9645–9656, Dec. 2020.
[33] Z. Zhao, S. Xu, J. Zhang, C. Liang, C. Zhang, and J. Liu, "Efficient and model-based infrared and visible image fusion via algorithm unrolling," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 3, pp. 1186–1196, Mar. 2022.
[34] H. Xu, J. Ma, and X.-P. Zhang, "MEF-GAN: Multi-exposure image fusion via generative adversarial networks," *IEEE Trans. Image Process.*, vol. 29, pp. 7203–7216, 2020.
[35] Z. Le et al., "UIFGAN: An unsupervised continual-learning generative adversarial network for unified image fusion," *Inf. Fusion*, vol. 88, pp. 305–318, Dec. 2022.
[36] X. Guo, Y. Li, and H. Ling, "LIME: Low-light image enhancement via illumination map estimation," *IEEE Trans. Image Process.*, vol. 26, pp. 982–993, 2017.
[37] S. Hao, X. Han, Y. Guo, X. Xu, and M. Wang, "Low-light image enhancement with semi-decoupled decomposition," *IEEE Trans. Multimedia*, vol. 22, no. 12, pp. 3025–3038, Dec. 2020.
[38] X. Li, M. Gao, J. Shang, J. Pan, and Q. Li, "A complexity reduction based retinex model for low luminance retinal fundus image enhancement," *Netw. Model. Anal. Health Inform. Bioinf.*, vol. 11, no. 1, p. 30, 2022.
[39] X. Li, Q. Li, M. Anisetti, G. Jeon, and M. Gao, "A structure and texture revealing retinex model for low-light image enhancement," *Multimedia Tools Appl.*, vol. 83, pp. 2323–2347, Jan. 2024.
[40] K. G. Lore, A. Akintayo, and S. Sarkar, "LLNet: A deep autoencoder approach to natural low-light image enhancement," *Pattern Recognit.*, vol. 61, pp. 650–662, Jan. 2017.
[41] Y. Jiang et al., "EnlightenGAN: Deep light enhancement without paired supervision," *IEEE Trans. Image Process.*, vol. 30, pp. 2340–2349, 2021.
[42] C. Guo et al., "Zero-reference deep curve estimation for low-light image enhancement," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 1780–1789.
[43] K. R. Prabhakar, V. S. Srikar, and R. V. Babu, "DeepFuse: A deep unsupervised approach for exposure fusion with extreme exposure image pairs," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 4714–4722.
[44] M. Li, J. Liu, W. Yang, X. Sun, and Z. Guo, "Structure-revealing low-light image enhancement via robust retinex model," *IEEE Trans. Image Process.*, vol. 27, pp. 2828–2841, 2018.
[45] L. Tang, X. Xiang, H. Zhang, M. Gong, and J. Ma, "DIVFusion: Darkness-free infrared and visible image fusion," *Inf. Fusion*, vol. 91, pp. 477–493, Mar. 2023.
[46] X. Jia, C. Zhu, M. Li, W. Tang, and W. Zhou, "LLVIP: A visible-infrared paired dataset for low-light vision," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 3496–3504.
[47] H. Xu, J. Ma, Z. Le, J. Jiang, and X. Guo, "FusionDN: A unified densely connected network for image fusion," in *Proc. 34th AAAI Conf. Artif. Intell.*, 2020, pp. 12484–12491.
[48] H. Xu, M. Gong, X. Tian, J. Huang, and J. Ma, "CUFD: An encoder–decoder network for visible and infrared image fusion based on common and unique feature decomposition," *Comput. Vis. Image Understand.*, vol. 218, Apr. 2022, Art. no. 103407.
[49] H. Xu, J. Ma, J. Jiang, X. Guo, and H. Ling, "U2Fusion: A unified unsupervised image fusion network," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 1, pp. 502–518, Jan. 2022.
[50] W. Di, L. Jinyuan, F. Xin, and R. Liu, "Unsupervised misaligned infrared and visible image fusion via cross-modality image generation and registration," in *Proc. Int. Joint Conf. Artif. Intell. (IJCAI)*, 2022, pp. 1–8.
[51] H. Zhang and J. Ma, "SDNet: A versatile squeeze-and-decomposition network for real-time image fusion," *Int. J. Comput. Vis.*, vol. 129, no. 10, pp. 2761–2785, 2021.
[52] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2017, *arXiv:1412.6980*.
[53] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 779–788.