

ORIGINAL ARTICLE

IETAFusion: An Illumination Enhancement and Target-aware Infrared and Visible Image Fusion Network for Security System of Smart City

Shuang Guo¹ | Kun Wu¹ | Seunggil Jeon² | Xiaomin Yang¹

¹College of Electronics and Information Engineering, Sichuan University, Chengdu, China

²Samsung Electronics 129, Samseong-ro Yeongtong-gu Suwon-si, Gyeonggi-do, South Korea

Correspondence

Xiaomin Yang, Sichuan University, Chengdu, 610064, China.

Email: arielyang@scu.edu.cn

Abstract

In the environmental security monitoring of smart cities, the infrared and visible image fusion method deployed on intelligent systems based on cloud and fog computing plays an vital role in providing enhanced images for target detection systems. However, the fusion quality can be significantly influenced by the illumination of the monitoring scenario in visible images. Therefore, conventional methods typically suffer a severe performance drop under the condition of insufficient illumination. To tackle this issue, we propose an illumination enhancement and target-aware fusion method (IETAFusion) based on artificial intelligence, which breaks the boundaries between the task of illumination enhancement and image fusion and provide a fusion result with better visual perception in nighttime scene. Specifically, we use a light-weight contrast enhancement module (CEM) restore the brightness of the visible image. Moreover, a Swin Transformer-based backbone network (STBNet) is utilized to facilitate information exchange between the source images and enhance the capabilities of target awareness. Finally, the fused images are reconstructed by the contrast-texture retention module (CTRM) and reconstructor. The extensive experiments indicates that the proposed approach achieves improved performance both in human perception and quantitative analysis compared with the state-of-the-art (SOTA) methods.

KEY WORDS

Image fusion, Cloud and Fog computing, Smart city, Artificial intelligence, Transformer

1 | INTRODUCTION

In recent years, owing to the remarkable advancements in AI, the domain of smart city has experienced substantial growth and development. Within this field, AI-enabled devices equipped with environmental security monitoring systems have been instrumental in safeguarding the security of the citizens. However, due to the optical limitations of imaging sensors, it is often unreachable for the devices to capture an image that contains the comprehensive information of entire scene (J. Ma, Ma, & Li, 2019) for security analysis. Therefore, a series of image fusion methods are emerging to obtain a high-quality single image that provide complementary information from multiple source images. In the sphere of image fusion, the visible and infrared image fusion (VIF) techniques have gained significant prominence and widespread adoption. Specifically, visible images are characterized by a wealth of textural details and encompass a diverse range of environmental contextual information. However, infrared images exploit the thermal imaging principle, thereby offering saliency information pertaining to entities such as pedestrians, animals, vehicles, *etc.* By combining visible and infrared imaging modalities, it becomes possible to overcome the limitations of improving the accuracy of object detection (Cao et al., 2019) and anomaly recognition in security systems.

Throughout recent decades, substantial advancements have been made to strengthen the accuracy and efficiency of security monitoring system by VIF. Existing fusion strategies can be roughly classified into two categories: conventional methods and

Abbreviations: AI, artificial intelligence; SOTA, state-of-the-art; VIF, Visible and infrared image fusion; CEM, contrast enhancement module; STBNet, Swin Transformer-based backbone network; CTRM, contrast-texture retention module;



FIGURE 1 Nighttime visible image enhancement result and corresponding infrared image fusion results. In comparison to the fusion output achieved by SwinFusion (J. Ma et al., 2022) on the enhanced version by KinD (Y. Zhang et al., 2019), our proposed method effectively preserves a greater extent of enhanced details and yields a more favorable visual perception while the SOTA method hides the texture details again. To facilitate a more comprehensive comparison, regions exhibiting abundant textures were magnified within the red boxes.

deep learning-based methods. Furthermore, conventional methods can be conceptually generalized into a three-step framework. Initially, an optimal transformation domain is selected to extract salient features from the source images. Subsequently, various fusion methodologies are employed to integrate the extracted features. Finally, the fused image is reconstructed by leveraging the corresponding inverse transformation to recover the fused features into a visually coherent representation. Depending on different way of measuring activity level and various fusion strategies, there are five research orientations in conventional approaches, *i.e.*, multi-scale transform-based (Chen, Li, Luo, Mei, & Ma, 2020; Ben Hamza, He, Krim, & Willsky, 2005), saliency-based (Bavirisetti & Dhuli, 2016), representation-based (Q. Zhang, Fu, Li, & Zou, 2013), subspace-based (Bai, Zhou, & Xue, 2011), and hybrid-based approaches (Y. Liu, Liu, & Wang, 2015; J. Ma, Zhou, Wang, & Zong, 2017). Despite the promising performance exhibited by traditional methods, they are inherently constrained by certain limitations and drawbacks that present significant challenges for facile mitigation. First and foremost, a major challenge in conventional methods is the development of an universally applicable feature extraction strategy due to the limited generalization capacity of such approaches. Secondly, representation-based techniques prone to (G. Liu et al., 2012) suffer from high computational costs, primarily attributed to the time-intensive process of dictionary learning. Thirdly, the intricate degradation patterns manifested by source images present a lot of hurdles for traditional methods to overcome.

Over the past year, due to the robust vitality displayed by AI in the field of image processing, AI enabled fusion approaches have demonstrated superior performance and opened up new avenues of exploration in security systems compared to conventional ones. Deep learning-based VIF methods can be categorized into three distinct classes: CNN-based approaches (Y. Liu, Chen, Cheng, Peng, & Wang, 2018), GAN-based approaches (J. Ma, Yu, Liang, Li, & Jiang, 2019), and AE-based approaches (H. Li & Wu, 2018). The CNN-based approach utilizes a specific network architecture to efficiently perform feature extraction, feature fusion, and image reconstruction with the integrated features obtained from the fusion process. Nevertheless, harnessing the potent image generation capabilities of GAN networks, GAN-based approaches obtain the fused images by exploiting the generative adversarial relations between the generator and discriminator. Specifically, within the GAN framework, the generator is tasked with generating the fused image, while the discriminator aims to discern whether the generated image is derived from the source image. Through unsupervised adversarial learning, the discriminator guides the network to generate results that closely approximate the source image. Additionally, AE-based approaches utilize an auto-encoder architecture for extracting features and incorporate a meticulously crafted fusion layer to integrate the extracted features. Consequently, the fused image is generated through the subsequent decoding process.

However, despite conventional VIF approaches based on deep learning have been extensively applied in the security systems, they often overlook extreme conditions such as the low-illumination degradation in visible images, which may cause substantial information loss. In particular, conventional methods can solely rely on infrared information to reconstruct contextual details due to the significant loss of details in invisible images within dark scenes. As a result, traditional approaches frequently fall short in representing the abundant scenario information, leading to an unnatural visual perception (Tang, Xiang, Zhang, Gong, & Ma, 2023). Hence, in highly dim environments, the primary focus lies in mitigating the degradation of illumination in visible images, which entails the task of low-light image enhancement (Wei, Wang, Yang, & Liu, 2018; Y. Zhang et al., 2019; Wu et al., 2022). One plausible approach involves pre-processing the low-light restoration network followed by fusing the enhanced visible and infrared images. However, experimental findings indicate several challenges when embedding the existing low-light restoration network into the fusion network. Firstly, an inherent incompatibility arises between the fusion task and the low-light

enhancement task, whereby the background details recovered during low-light enhancement may be obscured again in the fusion process, which is as shown in Fig. 1(d). Moreover, low-light enhancement networks already entail a certain computational cost, and combining the two networks inevitably leads to a reduction in efficiency.

To overcome the inherent limitations mentioned above, we propose an illumination enhancement and target aware network that seamlessly integrates the tasks of illumination enhancement and image fusion within a unified framework. This framework incorporates a pre-trained lightweight network as our contrast enhancement module (CEM) (L. Ma, Ma, Liu, Fan, & Luo, 2022), which predicts the illumination component of visible images. Subsequently, based on the Retinex theory, we are able to generate an enhanced visible image that preserves contrast and detail information. Then, the enhanced visible image, along with the infrared image, is fed into a fusion network based on the Swin Transformer (STBNet). To ensure the preservation of restored details, we introduce a carefully designed contrast texture preservation module before the process of reconstruction. Furthermore, to avoid the loss of enhanced details during the fusion process, we utilize annotated infrared images to generate salient target masks. These masks are integrated into the loss functions to improve the capability of target awareness, enabling the network to retain valid detail information that might be obscured in darker regions.

In conclusion, the primary contributions of our work are outlined as follows:

1. Our proposed IETAFusion framework not only effectively restores the details of visible images that are degraded by illumination, but also successfully preserves abundant of texture detail information and salient target information;
2. A light-weight pre-trained contrast enhancement module (CEM) is embedded in our fusion network to eliminate illumination degradation in visible images. The Swin Transformer-based backbone network (STBNet) generate the fused features with target information to minimize the incompatibility between the two enhancement processes;
3. The contrast-texture retention module (CTRM) implemented in the network reconstruction phase ensures the preservation of enhanced contrast and texture at the feature level. Simultaneously, the utilization of a loss function incorporating an infrared image mask enhances the network's ability to discern salient targets, resulting in the preservation of recovered details.

2 | RELATED WORKS

This section commences with a concise introduction of the SOTA learning-based methods for VIF. Then, we present an introduction to the Retinex-based low-light image enhancement techniques.

2.1 | CNN-based Fusion Strategies

The CNN-based method, characterized by its intricate network architecture and meticulously crafted loss function, demonstrates superior performance and robustness in the context of image fusion tasks. A prominent example of a CNN-based fusion method is SDNet (H. Zhang & Ma, 2021), which employs a specially designed squeeze-and-decomposition network to make the fusion results achieve a higher degree of scene information integration. To facilitate the network's ability to identify salient targets, STDFusionNet (J. Ma, Tang, Xu, Zhang, & Xiao, 2021) incorporates salient target masks within a dedicated loss function, which fosters the extraction of the features from different modal and ensures a comprehensive fusion process. Moreover, Long *et al.* (Long, Jia, Zhong, Jiang, & Jia, 2021) proposed the RXDNFuse which opened up a new direction for designing fusion rules by integrating the structural advantages of ResNet and DenseNet. To conduct long-range modeling of the fusion process, Ma *et al.* (J. Ma et al., 2022) introduced the self-attention mechanism into their fusion framework called SwinFusion by employing the Swin Transformer architecture. However, owing to the multi-head attention mechanism and image blocking strategy employed by the transformer, this architecture incurs higher computational and time costs. Although the above methods have shown promising results, none of them have taken the illumination conditions of visible images into account. Consequently, in PIAFusion, Tang *et al.* (Tang, Yuan, Zhang, Jiang, & Ma, 2022) devised an additional sub-network before the training to predict the distribution of the scene illumination. Nevertheless, the simplicity of its fusion framework limits its ability to effectively adapt to complicated lighting conditions and dynamic environmental variations. With the goal of reducing the illumination loss, DIVFusion (Tang et al., 2023) employed a scene-illumination disentangled network to restore brightness while preserving the contextual information of the original images. However, in spite of utilizing a color consistency loss to calibrate the color, the fusion process, which exclusively manipulates the intensity component while leaving the color component unchanged, still leads to color deviation in the resulting fused results.

2.2 | GAN-based Fusion Strategies

Ever since the introduction of the generative adversarial network (GAN) into the domain of VIF by Ma *et al.* in their work FusionGAN (J. Ma, Yu, et al., 2019), a variety of GAN-based fusion methods have emerged owing to their remarkable adaptability within this field. The GAN-based network architecture leverages the adversarial interplay between the generator and discriminator, effectively guiding the generator towards the production of fused images that show enhanced texture details. Inspired by FusionGAN, Han *et al.* (J. Ma, Xu, Jiang, Mei, & Zhang, 2020) recognized the potential inadequacy of a single adversarial game in establishing a stable system. Consequently, they proposed a novel architecture known as DDcGAN. This framework was devised to diminish the information loss, which may arise in the single-adversarial network, and to achieve a more balanced image fusion outcome. To preserve the target boundary and enhance the texture details, Ma *et al.* (J. Ma, Liang, et al., 2020) presented an end-to-end fusion framework called Detail-GAN, which utilize the specially designed loss function to improve the fusion quality and sharpen the edges of infrared targets. However, the aforementioned methods still encounter challenges in generating fused images that possess appropriate visual perception and naturalness. To this end, Fu *et al.* (Fu, Wu, & Durrani, 2021) proposed a framework named as perception-GAN to improve the visual quality of the fused images. Building upon previous studies, Li *et al.* (J. Li, Huo, Li, Wang, & Feng, 2020) presented AttentionFGAN, which applies a multi-scale attention mechanism into a generative adversarial network to emphasize salient regions in the scene. Nonetheless, the network entails a excessive parameters, and the training process is characterized by considerable time consumption.

2.3 | AE-based Fusion Strategies

In addition to the aforementioned frameworks, auto-encoder (AE) based architectures have also gained significant popularity. Within the AE framework, features from the source images are extracted using an auto-encoder, and these features are typically fused through a manually designed fusion rule. Subsequently, the fused features are sent to the Decoder for generating the fusion outcomes. The typical AE-based approach is the previously mentioned DenseFuse (H. Li & Wu, 2018), which involves the common structure of auto-encoder. Based on DenseFuse, Li *et al.* proposed the NestFuse (H. Li, Wu, & Durrani, 2020), which utilize the nest connect-based network to extract the multi-scale features. To address the challenge of developing an adaptable fusion strategy, Li *et al.* further proposed RFN-Nest (H. Li, Wu, & Kittler, 2021), which deploys a residual fusion network to replace the conventional fusion manners. Similarly, since the existing handcrafted fusion rules restrain the development of learning-based fusion methods, Xu *et al.* developed a classification saliency-based rule (CSF) (Xu, Zhang, & Ma, 2021) to obtain a saliency map and fuse the features in terms of the acquired saliency weights. As the information redundancy and the computational load cannot be diminished in the former works, Jian *et al.* (Jian et al., 2020) devised a symmetric framework (SEDRFuse) to better retain the thermal information and texture details.

Despite the notable advancements in fusion performance and extended application domains achieved by existing methods, there remains a limited number of approaches that demonstrate satisfactory fusion quality when dealing with visible images captured under extremely low illumination conditions. Moreover, while certain methods have payed attention to the issue of lightness degradation, they often leads to undesirable color distortion because the illumination enhancement process is solely conducted on the conventional Y-channel. Hence, there is a pressing demand to explore techniques that can efficiently restore the lost details in low-illumination visible images without altering the color distribution, while simultaneously achieving a harmonious fusion with salient targets presented in the infrared images.

2.4 | Retinex-based Low-light Image Enhancement Strategies

Recently, low-light image enhancement (LLIE) has also become a prominent area of interest in the field of image processing. After years of exploration, there has been a growing emphasis on utilizing Retinex theory as the foundation for LLIE. Retinex theory presumes that color images can be decomposed into reflection component and illumination component, which can be mathematically represented as:

$$I = R \cdot L \quad (1)$$

Where R and L indicate the reflection component and illumination component of a color image I . In the Retinex model, the reflection component is regarded as an intrinsic attribute of the image, thereby maintaining consistency between the reflection

components of low light images and corresponding normal light images. On the other hand, the illumination component represents the overall light intensity of the image, and its distribution indicates variations across different lighting conditions.

Learning based Retinex approaches have gained prominence in LLIE because of the superior stability and robustness of the system. One classic work is RetinexNet (Wei et al., 2018), which employs a decomposition module to estimate the two components. Subsequently, the illumination components is enhanced through an illumination adjustment module, leading to improved visual quality. Inspired by RetinexNet, KinD (Y. Zhang et al., 2019) allows users to obtain enhanced results at any lighting level by learning a flexible mapping function. Aiming to reduce noise artifacts and enhance the preservation of image details, URetinex (Wu et al., 2022) introduced an unfolding optimization module to iteratively refine the reflection components and illumination components. In addition to the above supervised manner, zhang *et.al* (F. Zhang et al., 2021) conducts self-supervised learning through histogram equalization prior to adapt to different lighting conditions and different imaging devices. To minimize computational complexity and time consumption, SCI (L. Ma et al., 2022) develops a lightweight illumination prediction network within a meticulously designed lighting self-calibration framework, aiming to expedite the low-light enhancement procedure. It is important to highlight that our CEM framework is derived from the prediction model of SCI, owing to its simplicity and outstanding performance.

3 | METHODOLOGY

In this section, we provide an elaborate explanation of our IETAFusion. In Section 3.1, we first discuss the problem formulation. In Section 3.2, we present detailed explanation of our proposed designs, including the contrast enhancement module (CEM) for restoring the illumination of the visible images, the backbone fusion network based on Swin Transformer (STBNet) for feature extraction and fusion, and the contrast-texture retention module (CTRM) for image reconstruction. Finally, we describe our loss functions for training the CEM and fusion network in Section 3.3.

3.1 | Problem formation

Given that illumination degradation predominantly occurs in visible images, how to eliminate the degradation within the network and reasonably integrate it into the overall architecture has become the crucial factors in our work. Inspired by SCI (L. Ma et al., 2022), we utilize a light weight contrast illumination enhancement module to effectively restore lightness while minimizing computational complexity. Specifically, in accordance with Retinex theory in Eq.1, there exists a linear correlation between illumination insufficient image y and its enhanced version z , which can be expressed as: $y = i \otimes z$, where i represent the illumination component and \otimes indicates the element-wise multiplication. Within our contrast enhancement module (CEM), we employ a straightforward network for the estimation of visible image illumination. Subsequently, we attain the desired bright image through element-wise division and send it to the fusion process. The structure of CEM is depicted in Fig.3, and the progressive illumination prediction process can be formulated as follows:

$$\begin{cases} u^t = \mathcal{H}_{\theta^c}(x^t), & x^0 = y \\ i^t = u^t \oplus x^t \\ x^{t+1} = i^t \end{cases} \quad (2)$$

where u^t and x^t indicates the residual term and the input at t -th ($t = 0, \dots, T - 1$) stage and \oplus represents the element-wise addition operation. \mathcal{H}_{θ^c} denotes the mapping network in CEM with learnable parameter θ^c , which would be shared in whole stage. This approach is motivated by a widely accepted consensus, *i.e.*, the similarity or existence of a residual difference between the illumination and low-light observation. subsequently, the desired bright image can be described as: $z^t = y \oslash i^t$. where \oslash represents the element-wise division. It is noteworthy that while the training process of CEM is conducted in phases, the similarity of outputs between the preceding and subsequent stages allows for the utilization of a single stage *i.e.*, the first stage, to accelerate the inference process in the testing phase. In Eq. (2), we transform the input at each stage into the illumination component derived from the output of the previous stage for stage-wise training. However, relying solely on illumination components can lead to color shifts in the final output. Therefore, we employ a self-calibration module (SCM) to calibrate the

input. The process of self-calibration can be presented as:

$$\begin{cases} \mathbf{r}^t = \mathcal{H}_{\theta^k}(\mathbf{z}^t) \\ \mathbf{s}^t = \mathbf{z}^t \ominus \mathbf{r}^t \\ \mathbf{x}^{t+1} = \mathbf{y} \oplus \mathbf{s}^t \end{cases} \quad (3)$$

where \mathbf{s}^t and \mathbf{r}^t represent the output and difference term of the self-calibration module, respectively. \mathcal{H}_{θ^k} denotes the self-calibrated mapping network with learnable parameters θ^k . The entire self-calibration process is integrated into the architecture of CEM and exclusively applied during the training phase.

After that, we propose a progressive backbone network based on Swin Transformer (STBNet) for feature fusion, as depicted in Fig. 2, which illustrates the over structure of our network. By leveraging the computational efficiency and long-range modeling capability offered by the Swin Transformer, it serves as an effective backbone network for feature extraction and fusion, yielding impressive fusion results. The whole feature extraction and fusion process can be defined as:

$$F_{vi,ir}^{l_{n+1}} = \begin{cases} o(\text{cov}_{vi,ir}^{l_{n+1}}(I_{vi,en,ir})), & n = 0 \\ o(\text{cov}_{vi,ir}^{l_{n+1}}(F_{attn}^{l_n} \oplus F_{vi,ir}^{l_n})), & n = 1, \dots, N \end{cases} \quad (4)$$

$$F_{attn}^{l_{n+1}} = \begin{cases} \mathcal{H}_{\theta^s}^{n+1}(F_{ir}^{l_{n+1}} \ominus F_{vi}^{l_{n+1}}), & n = 0 \\ \mathcal{H}_{\theta^s}^{n+1}(o(\text{cov}_{attn}^{l_n}(F_{attn}^{l_n})) \oplus F_{ir}^{l_{n+1}} \ominus F_{vi}^{l_{n+1}}), & n = 1, \dots, N-1 \end{cases} \quad (5)$$

where $F_{vi,ir}^{l_n}$ represents the features extracted from one of the two modal inputs and $F_{attn}^{l_n}$ indicates the attention map at the n -th stage respectively owing to the symmetrical structure of the backbone network. $\text{cov}_{vi,ir}^{l_n}$, $\text{cov}_{attn}^{l_n}$ denotes the convolutional layers with the LeakyReLU activation function set behind at the n -th stage. \mathcal{H}_{θ^s} indicates the Swin Transformer with parameter θ^s at the n -th stage. $I_{vi,en,ir}$ represents brightness restored visible image on Y-channel or the origin infrared image. N is the total number of the stages, which is set to 3 in our work.

Once the final infrared features $F_{vi}^{l_{out}}$, visible image features $F_{ir}^{l_{out}}$, and attention map features $F_{attn}^{l_{out}}$ are obtained, they are collectively forwarded to the reconstruction module for the purpose of reconstructing the images. Within the reconstruction module, the meticulously designed Contrast-texture Retention Module (CTRM), as illustrated in Fig. 4, is employed to retain the extracted texture details and restored contrast information with salient target awareness. The features of target awareness ϕ_a can be defined as:

$$\phi_a = o(\text{conv}_{1 \times 1}(F_{attn}^{l_{out}})) \quad (6)$$

where $\text{conv}_{1 \times 1}$ refers to the convolutional layer with a kernel size of 1×1 . Here, we exploit the feature attention mechanism (Cho, Ji, Hong, Jung, & Ko, 2021) to selectively emphasize or suppress the contrast and gradient information. The procedure is expressed as follows:

$$\phi_g = o\left(\nabla\left(\text{conv}_{3 \times 3}\left(\text{concat}\left(F_{vi}^{l_{out}}, F_{ir}^{l_{out}}\right) \otimes \phi_a\right)\right)\right) \quad (7)$$

$$\phi_c = o\left(\text{conv}_{3 \times 3}\left(o\left(\text{conv}_{3 \times 3}\left(\text{concat}\left(F_{vi}^{l_{out}}, F_{ir}^{l_{out}}\right) \otimes \phi_a\right)\right)\right)\right) \quad (8)$$

where ϕ_g and ϕ_c denote the features of gradient and contrast, respectively. concat represents the concatenation in channel dimensions and ∇ denotes the texture extraction operation achieved by Sobel operator. Furthermore, the final fused image I_f^Y is reconstructed as follows:

$$I_f^Y = \text{Re}\left(\text{ADD}(\phi_a, \phi_c, \phi_g)\right) \quad (9)$$

where $\text{ADD}(\cdot)$ represents element-wise addition operation and $\text{Re}(\cdot)$ denote the reconstructor comprising a sequential arrangement of convolutional layers. It is noteworthy that during the color space transformation process, the CbCr channels of enhanced visible image are concatenated with the I_f^Y , rather than relying on the CbCr channel of the origin visible image. This decision is based on the fact that the previous contrast enhancement module (CEM) not only modifies the Y channel but also alters the distribution of the color channels. Therefore, utilizing the CbCr component of the source image would introduce color distortion due to the mismatch between the modified Y channel and the original color components, which will be described in Section

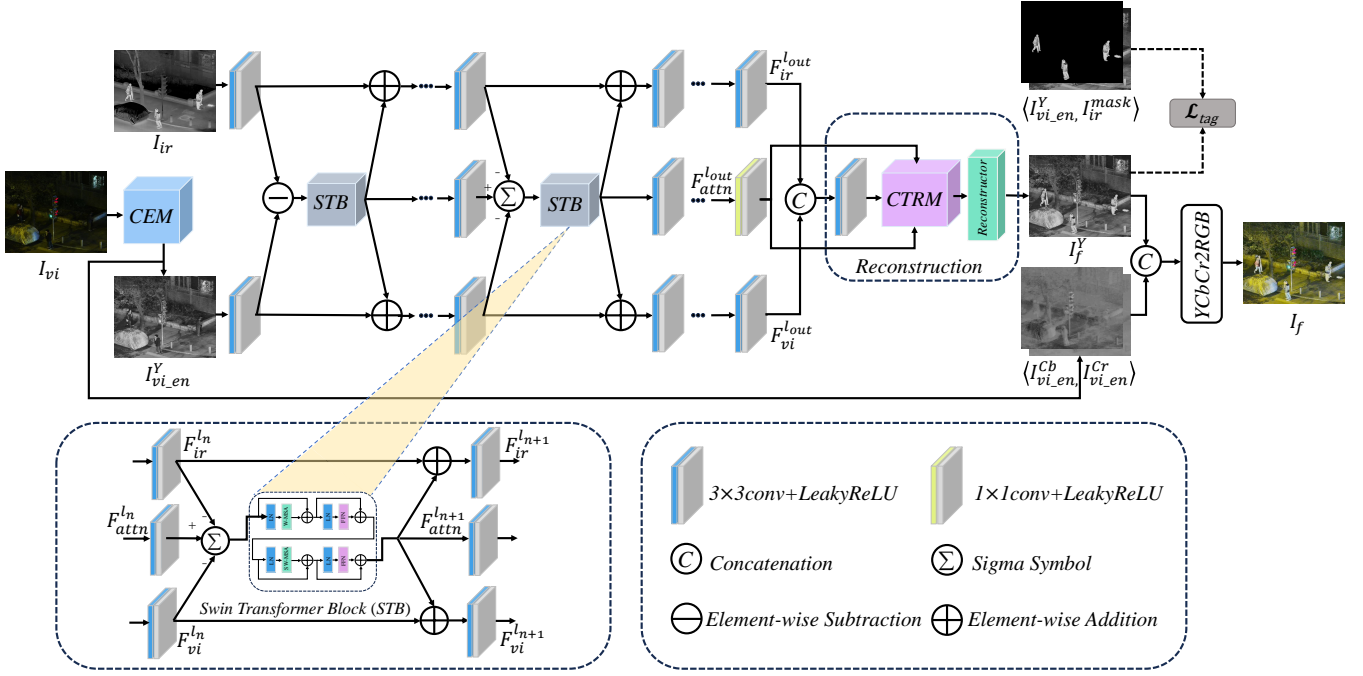


FIGURE 2 The overall framework of the proposed IETA Fusion.

4.6.5. Finally, the transformation of color space from YCbCr to RGB can be expressed as the following procedure:

$$I_f = \Gamma \left(\text{concat} \left(I_f^Y, I_{vi_en}^{Cb}, I_{vi_en}^{Cr} \right) \right) \quad (10)$$

where $\Gamma(\cdot)$ indicates the transformation function used for converting the color space from YCbCr to RGB. $I_{vi_en}^{Cb}$ and $I_{vi_en}^{Cr}$ represent the Cb and Cr channels of the enhanced visible images.

3.2 | Network Architecture

3.2.1 | Contrast enhancement module (CEM)

The architecture of the Contrast Enhancement Module (CEM) is illustrated in Fig. 3. CEM is composed of three convolutional layers with a kernel size of 3×3 and one batch-norm layer, with the activation function applied to the first two convolutional layers being the Leaky Rectified Linear Unit (LeakyReLU), while the final activation function is Sigmoid. Furthermore, each convolutional layer within CEM is configured with 3 input and output channels, which means that the architecture shows a remarkably low parameter quantity. Additionally, the self-calibration module (SCM) in training phase follows a similar structure to that of CEM, comprising two input-output convolutions and a set of residual blocks (RB_n). The ReLU activation function is used in all layers, except for the final layer which employs the Sigmoid activation function. To accelerate the training process, all convolutional layers, except for the input-output convolution, are configured with a channel count of 16. As CEM remains in the testing stage during the entirety of IETA Fusion, the enhanced visible image obtained from the initial stage of the pretrained CEM is transformed into the YCbCr color space and subsequently utilized as the input for the STBNet.

3.2.2 | Swin Transformer-based backbone network (STBNet)

As presented in Fig. 2, we employ the Swin Transformer-based framework as the backbone network (STBNet) for feature extraction and fusion. Specifically, a N-layer convolutional structure is utilized as the bypass of the backbone network to extract common and supplementary features from the source image. Each convolutional layer comprises of a 3×3 kernel, with the

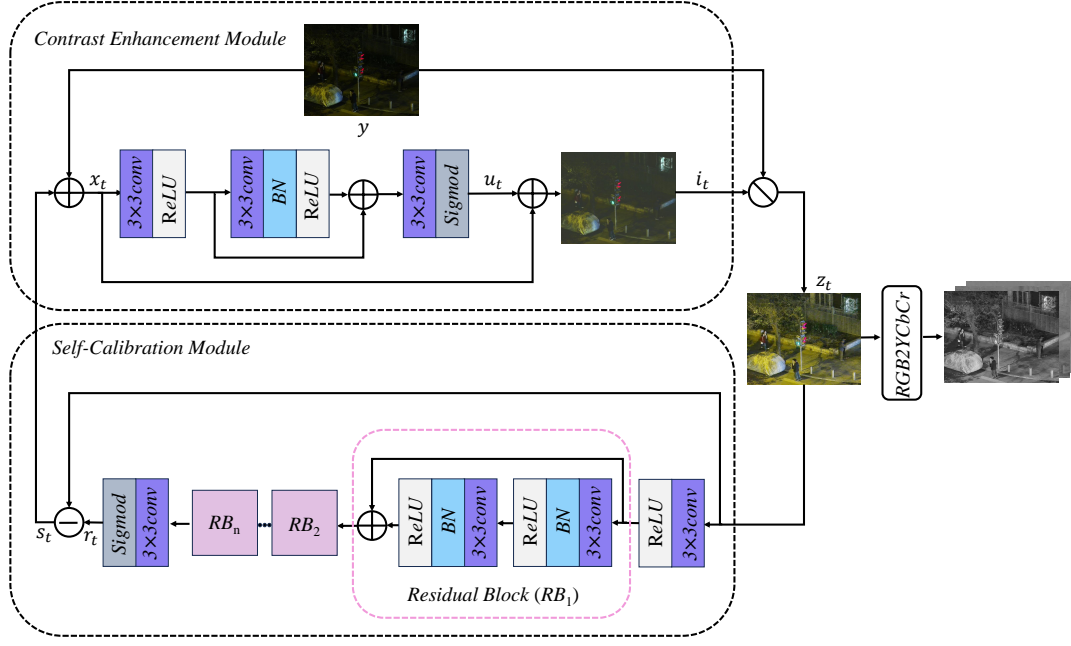


FIGURE 3 The architecture of employed CEM and SCI in training phase.

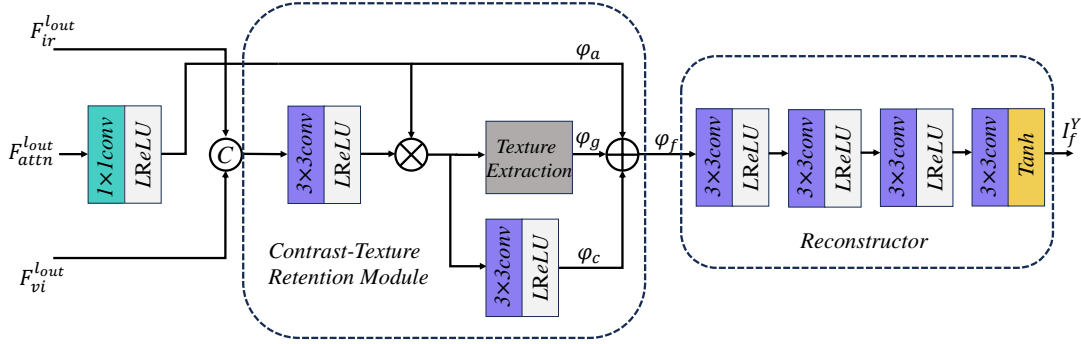


FIGURE 4 The framework of image reconstruction and proposed CTRM.

LeakyReLU activation function employed after each convolutional layer. Furthermore, to enhance the network's capability to capture salient target information, we adopt the Swin Transformer Block (J. Ma et al., 2022) as the main stream of the backbone network. This choice is motivated by the block's inherent capacity for long-range modeling, enabling it to effectively retain the recovered details and important target-specific information. It is worth emphasizing that, for the purpose of enhancing efficiency and reducing computational complexity, our Swin Transformer Block is deliberately constrained to incorporate a concise arrangement of solely two successive Swin Transformer layer layers. Moreover, in order to ensure consistent dimensions between the bypass and mainstream features, a convolutional layer with a kernel size of 3×3 is employed after each Swin Transformer Block layer, with the exception of the final layer which utilizes a 1×1 convolutional kernel. The subsequent activation function employed is LeakyReLU.

3.2.3 | Contrast-texture retention module (CTRM)

The Contrast-texture Retention Module (CTRM) is designed to preserve the extracted details and restored contrast information during the reconstruction phase of the network, while reduce the information loss caused by simplistic structures. As depicted in Fig. 4, CTRM comprises a residual structure composed of two convolutional layers and a Sobel operator serving as texture extraction. Each convolutional layer utilizes a 3×3 convolution kernel with the LeakyReLU. Additionally, to retain salient target

information, we introduce the extracted attention map into CTRM through the feature attention mechanism. Ultimately, the fused image is generated by a reconstructor, which is composed of four convolutional layers. All convolutional layers use the LeakyReLU activation function, except for the final layer, which utilizes Tanh.

3.3 | Loss function

3.3.1 | Illumination enhancement loss

The effectiveness of contrast enhancement in our proposed approach is heavily dependent on the successful restoration of contrast through the CEM. Therefore, it is imperative to pre-train the CEM to ensure the desired enhancement effect is achieved. To address the challenge posed by the lighting inconsistencies present in the paired high and low lighting image datasets, we employ the unsupervised illumination enhancement loss in SCI (L. Ma et al., 2022). This loss enhances the robustness and generalization capability of the network, allowing it to effectively adapt to various lighting conditions and produce visually appealing and contrast-enhanced results. The whole illumination loss is calculated as:

$$\mathcal{L}_{illu} = \alpha_1 \mathcal{L}_f + \alpha_2 \mathcal{L}_s \quad (11)$$

where \mathcal{L}_f and \mathcal{L}_s represent the fidelity and smoothing loss. α_1 and α_2 denote the hyper-parameters responsible for balancing each loss. The fidelity loss is employed to constrain the offset between the illumination component of the output and input at each training stage, which can be written as:

$$\mathcal{L}_f = \|i^0 - y\|_2 + \sum_{t=1}^{T-1} \|i^t - (y + s^{t-1})\|_2 \quad (12)$$

where T represents the number of the training stage. The smoothness loss incorporates a spatially-variant l_1 -norm to maintain the smoothness of the illumination component, which is formulated as follows:

$$\mathcal{L}_s = \sum_{m=1}^N \sum_{n \in \mathcal{N}(i)} w_{m,n} \|i_m^t - i_n^t\|_1 \quad (13)$$

where N indicates the overall number of pixels and m refers to the index of each pixel. $\mathcal{N}(i)$ denotes the 5x5 neighborhood. The weight $w_{m,n}$ of each neighborhood is calculated as follows :

$$w_{m,n} = \exp \left(-\frac{\sum_c ((y_{m,c} + s_{m,c}^{t-1}) - ((y_{n,c} + s_{n,c}^{t-1})))^2}{2\sigma^2} \right) \quad (14)$$

where c represents the channel in YCbCr color space, and σ denotes the standard deviations for the Gaussian kernels.

3.3.2 | Fusion loss

Since our objective is to retain the background details of the visible image after restoration by CEM, while simultaneously incorporating salient targets information from the infrared domain during the fusion process. Inspired by STDFusion (J. Ma et al., 2021), we create a mask-based loss function \mathcal{L}_{tag} by marking the salient targets in the infrared images. This loss function replaces the conventional constraint on the infrared image with the mask derived from the infrared image, which aims to combine significant infrared information while preserving as much of the recovered background detail as possible. Consequently, our target perception loss can be defined as follows:

$$\mathcal{L}_{tag} = \frac{1}{HW} \left\| I_f^Y - \max \left(I_{vi_en}^Y, I_{ir}^{mask} \right) \right\|_1 \quad (15)$$

where $\max(\cdot)$ represents the element-wise maximum calculation. H and W refer to the height and width of the input images. I_{ir}^{mask} represents the grayscale mask of the infrared image, where the less meaningful background regions are assigned a value of 0.

While the target perception loss effectively guides the network to fuse relevant information from the source images, the background information of infrared images is not to be supplemented. To address this limitation, we introduce an supplementary loss, which is expressed as follows:

$$\mathcal{L}_{sup} = \gamma_1 \left\| I_f^Y - \max \left(I_{vi_en}^Y, I_{ir} \right) \right\|_2 + \gamma_2 \left\| I_f^{back} - I_{ori}^{back} \right\|_2 \quad (16)$$

where γ_1 and γ_2 are the weights to control the two loss terms. I_f^{back} and I_{ori}^{back} refer to the background information from the fused image and source image, which can be defined as:

$$I_f^{back} = |I_f^Y - I_{ir}^{mask}| \quad (17)$$

$$I_{ori}^{back} = \left| \max \left(I_{vi_en}^Y, I_{ir} \right) - I_{ir}^{mask} \right| \quad (18)$$

where $|\cdot|$ represents the mathematical operation of absolute value. Furthermore, we use the texture loss to force the network to transfer fine-grained details from both the source and enhanced images to the fused image. The texture loss is written as follows:

$$\mathcal{L}_{tex} = \frac{1}{HW} \left\| |\nabla I_f^Y| - \max \left(|\nabla I_{vi_en}^Y|, |\nabla I_{ir}| \right) \right\|_1 \quad (19)$$

where ∇ represents the Laplace operator. Finally, the fusion loss is a weighted combination of target loss, supplementary loss and detail loss, which is presented as follows:

$$\mathcal{L}_{fuse} = \beta_1 \mathcal{L}_{tag} + \beta_2 \mathcal{L}_{sup} + \beta_3 \mathcal{L}_{tex} \quad (20)$$

where $\beta_1, \beta_2, \beta_3$ represents the hyper-parameters that regulate the balance between each loss term.

4 | EXPERIMENTS

In the following section, we firstly provide a comprehensive overview of the experimental configurations and training specifics employed for our network. Then, we carry out several comparison experiments, including fusion comparisons, two-stage enhancement and fusion comparisons and generalization experiments, to highlight the advancement of our network. Moreover, We undertake ablation studies to authenticate the effectiveness of the distinctive designs in our work. Finally, we conduct computational complexity to evaluate the efficiency of our framework.

4.1 | Experimental configurations

In our primary experimental evaluations, we have selected 25 pairs of aligned images from LLVIP dataset (Jia, Zhu, Li, Tang, & Zhou, 2021) to comprehensively assess the fusion performance of our IETA Fusion. Furthermore, we have also utilized the MSRS dataset (Tang et al., 2022) which contains 30 pairs of images to investigate the generalization capability of our approach in our generalization experiments.

It is worth noting that we performed comparative evaluations of our proposed approach against seven latest fusion algorithms, which include one conventional method, *i.e.*, GTF (J. Ma, Chen, Li, & Huang, 2016), six learning based method, *i.e.*, FusionGAN (J. Ma, Yu, et al., 2019), SDNet (H. Zhang & Ma, 2021), RFN-Nest (H. Li et al., 2021), PIAFusion (Tang et al., 2022), SwinFusion (J. Ma et al., 2022), DIVFusion (Tang et al., 2023). In addition to conventional comparative experiments, we have conducted a two-stage enhancement and fusion comparative experiments inspired by DIVFusion. Specifically, recognizing that the majority of methods solely utilize the initial low-light image as input without incorporating low-light enhancement process, we have employed three latest low-light enhancement algorithms, namely KinD (Y. Zhang et al., 2019), URetinex (Wu et al., 2022), and SCI (L. Ma et al., 2022), to pre-process the visible image before its utilization as input for these fusion algorithms. Notably, it should be emphasized that the inherent enhancement process already exists in DIVFusion, rendering any additional pre-enhancement processing unnecessary for its input.

For the quantitative evaluations, we have selected four metrics to assess the network's performance: average gradient (AG), entropy (EN), visual information fidelity (VIF), and spatial frequency (SF). These metrics were chosen as they provide comprehensive measures of various aspects of the image fusion quality. The AG metric calculates the level of detailed information

present in the image. The EN metric evaluates the information content of the image from an information theory standpoint. The VIF metric assesses the faithfulness of the information representation according to human visual perception. SF metric captures the frequency of variations in the image, providing an assessment of texture complexity. A fusion algorithm that yields higher values of AG, EN, VIF, and SF indicates superior performance in terms of fusion quality.

4.2 | Training details

The training of our two-stage IETAFusion model involves separate training of the Contrast Enhancement Module (CEM) and the fusion network. The training procedure for the CEM follows the same settings as in SCI. Specifically, the DarkFace dataset (Yang et al., 2020) was employed to train the CEM. It is noteworthy that the images were randomly cropped into patches with a size of 400×400 and normalized to $[0, 1]$ before training. The number of training stages T was set to 3, and the number of residual blocks (RB_n) in the self-Calibration Module (SCM) was set to 2. We have utilized the Adam optimizer to update the parameters. The batch size and learning rate was set to 4 and 10^{-4} . Moreover, $\alpha_{(\cdot)}$ of Eq. (11) were configured as follows: $\alpha_1 = 1.5$, $\alpha_2 = 1$. The total number of training epoch was set to 1000.

The fusion network was trained using 1700 pairs of visible and infrared images, as well as aligned mask images in the night scene which were randomly selected from LLVIP dataset. These images were randomly cropped into patches of size 128×128 and normalized to $[0, 1]$ before being fed into the network for training. The batch size was set to 8 and all parameters in the fusion network were updated by the Adam optimizer. The initial learning rate was set to 0.001, and it was halved every 100 epochs to facilitate the training process. The total number of training epochs were also set to 1000. Additionally, $\gamma_{(\cdot)}$ of Eq. (16) were set as follows: $\gamma_1 = 0.5$, $\gamma_2 = 0.5$. $\beta_{(\cdot)}$ of Eq. (20) were experimentally specified as follows: $\beta_1 = 100$, $\beta_2 = 50$, $\beta_3 = 10$. The whole network was trained and tested with Pytorch framework on GeForce 1080Ti GPU and 4.2 GHz Intel Core i7-7700k CPU.

4.3 | Comparative experiments

4.3.1 | Qualitative analysis

To assess the fusion quality in terms of human visual perception, two pairs of infrared and visible images (labeled as c_1, c_2) were selected from the LLVIP test set. The visual fusion outcomes are depicted in Fig. 5. Owing to substantial illumination loss in the source visible image, conventional fusion outcomes tend to exhibit reduced detail and poorly visualized contrast in the initial scene. Considering the overall perspective of the two images, FusionGAN, SDNet, and RFN-nest tend to produce darker fusion results with a loss of fine details. Although GTF yields brighter outcomes, the overall image shows a stronger inclination towards the infrared component, thereby overshadowing the details in the visible light image. SwinFusion, while maintaining a normal contrast, displays relatively more blurriness in the fusion results, as evidenced by the red box in c_2 . PIAfusion is capable of illumination awareness, but falls short in delivering satisfactory fusion results in extremely dark scenes. DIVFusion minimize the degradation of illumination, resulting in higher scene contrast, but it also introduces overexposure and some color distortion. It is noteworthy that, apart from DIVFusion and our proposed method, none of the conventional approaches are able to unveil intricate details in dark regions. This phenomenon has been illustrated through the magnified views of the flag region in the green box of c_1 and the fence region in c_2 . By contrast, our method demonstrates superior visual perception by effectively recovering a substantial amount of detail and incorporating salient target information.

4.3.2 | Quantitative analysis

We further conducted a quantitative comparison to verify the advancement of our proposed approach. Table 1 displays the quantitative comparison results. The values in the table correspond to the average value of each metric. The table shows that our method achieves the highest ranking in three metrics and ranks fourth in EN metric. The superior AG metric reflects that our results exhibit a greater amount of texture details. Additionally, our AG metric demonstrates a substantial improvement compared to the seven SOTA methods, providing further evidence of our method's exceptional capability to recover extensive details from visible images. Then, The best VIF metric suggests that our results show the highest level of information fidelity. The optimal SF metric indicates that our results have retained more high spatial frequency detail information, which is also demonstrates that our approach can produce fusion outcomes with higher texture complexity and detail richness. Although our

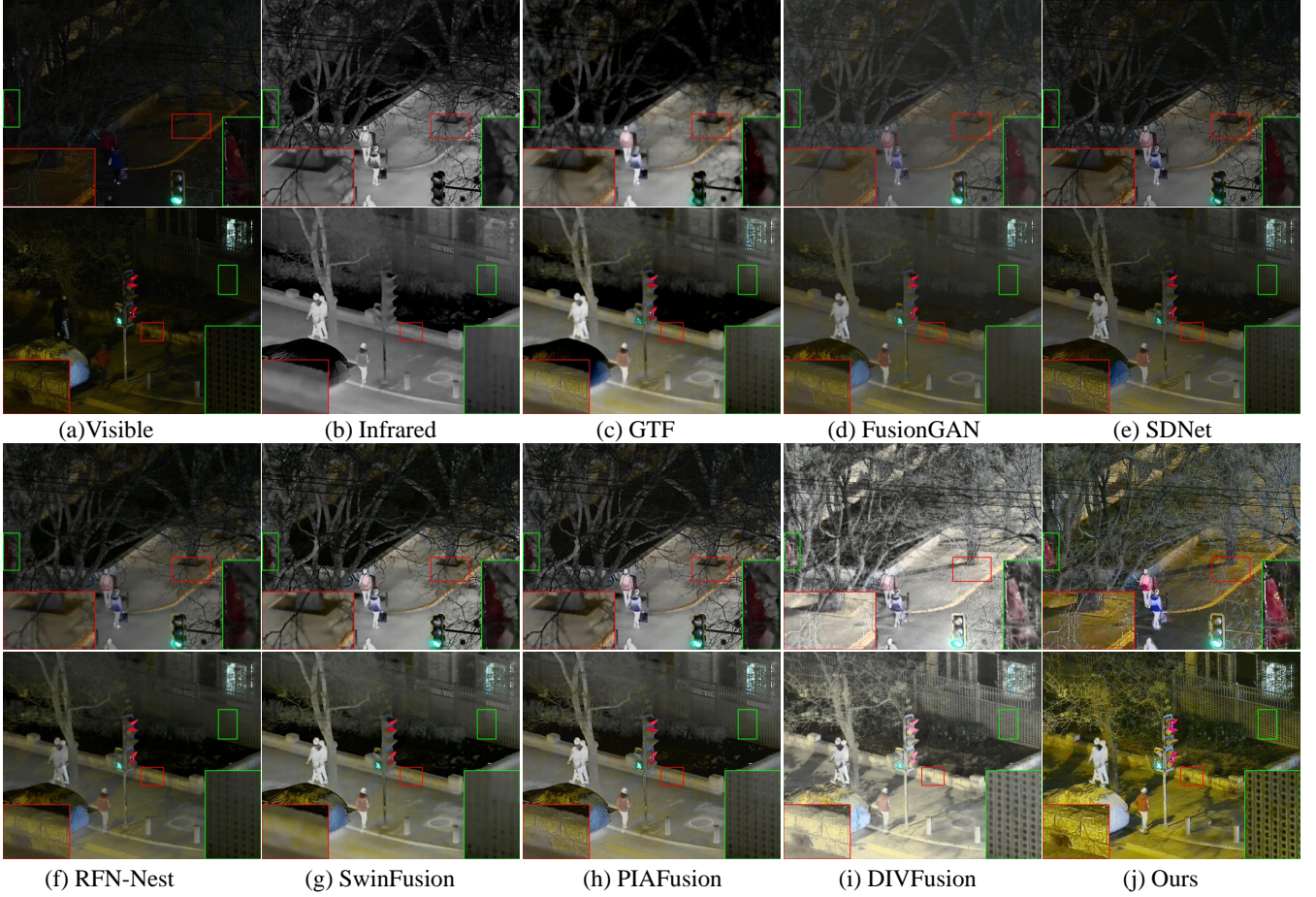


FIGURE 5 Visual quality results on two image pairs (denoted as c_1 and c_2 from top to bottom) from the LLVIP dataset. To facilitate a clearer comparison, we have zoomed in the regions with rich textures, which are shown in red box and green box.

TABLE 1 Quantitative results of the 4 metrics on 50 image pairs selected from LLVIP dataset. red value indicates the best result and blue value denotes the second best result.

	AG	EN	VIF	SF
GTF	3.9580	6.9671	0.7412	0.0547
FusionGAN	2.3155	6.3354	0.5062	0.0299
SDNet	4.8898	6.7064	0.6315	0.0606
RFN-Nest	2.7922	6.8931	0.7447	0.0300
SwinFusion	4.5237	7.3379	0.9576	0.0569
PIAFusion	5.7927	7.2860	0.9067	0.0703
DIVFusion	5.5227	7.6067	1.0583	0.0603
Ours	9.5508	7.2479	1.1875	0.1134

method ranks lower than other methods in terms of the EN metric, this can be justified. Specifically, as our fusion model is designed to prioritize target awareness, the emphasis is placed on preserving visible image details in the background, which may result in a slight loss of infrared information in non-prominent target regions. However, this trade-off is acceptable as it allows for the enhancement of visible details in exchange. In summary, our approach shows clear advantage over the SOTA methods.

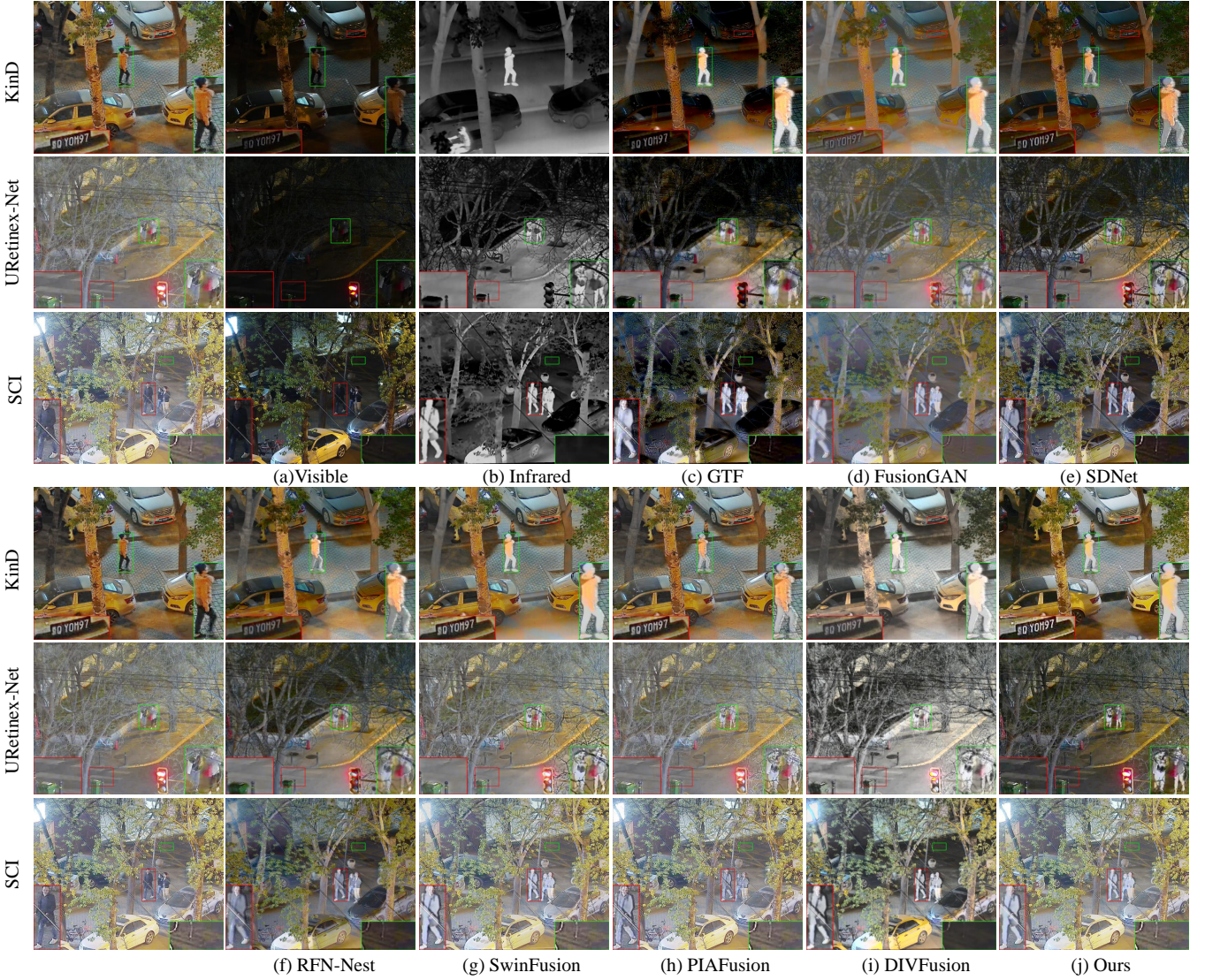


FIGURE 6 Comparison of the visual quality of two-stage enhancement and fusion experiments, where each row corresponds to a different scene labeled as s_1 , s_2 , s_3 from top to bottom. The first column illustrates the results obtained by applying different enhancement algorithms to the visible images.

4.4 | Two-stage enhancement and fusion performance analysis

For further validating the advantages of our method, We performed a comparative experiment on two-stage enhancement and fusion by applying pre-enhancement to the input visible images using selected SOTA methods that do not include a low-light enhancement process., *i.e.*, GTF, FusionGAN, SDNet, RFN-Nest, SwinFusion, PIAfusion. Specifically, three SOTA low-light enhancement algorithms namely KinD (Y. Zhang et al., 2019), URetinex-Net (Wu et al., 2022) and SCI (L. Ma et al., 2022) were chosen to perform illumination restoration of visible images.

4.4.1 | Enhanced by KinD

Given the inability of conventional fusion methods to address the issue of illumination loss in low-light visible images, we employ a classic illumination enhancement algorithm, *i.e.*, KinD (Y. Zhang et al., 2019) to restore the illumination in the visible images prior to image fusion. Specifically, KinD employs a decomposition network, a reflection recovery network, and a light

adjustment network to optimize the reflection component and constrain the illumination component, achieving an effective light enhancement. Fig. 6 s_1 illustrates the visual outputs of the low-light image enhancement and fusion process.

By examining Fig. 6 s_1 , it is evident that the application of KinD leads to a certain enhancement of ambient illumination in each fusion result, compared to the original visible image. However, the fusion results achieved by GTF, SDNet, and RFN-Nest remain comparatively dark and fail to capture the detail information recovered by KinD. Furthermore, FusionGAN yields fusion outcomes that exhibit increased blurriness when compared to the other methods, displaying incompatibility with the employed enhancement methods. Comparing SwinFusion and PIAFusion with our method, we observe a similar contrast in their fusion outcomes. However, as highlighted by the red box in s_1 , our method excels in capturing and presenting recovered texture details, while DIVFusion shows a suboptimal performance in this dark region. Furthermore, in the zoomed green box of s_1 , it becomes apparent that PIAFusion does not retain as much salient target infrared information as our method. Overall, our approach demonstrates subjective advantages over the two-stage fusion approach with KinD as a pre-processing step.

The quantitative results are shown in Table 2. Compared with the data in Table 1, we can demonstrate that the application of KinD as a pre-processing technique leads to enhancements in AG, EN, VIF, and SF metrics. However, it should be noted that in the case of SwinFusion and PIAFusion, there is a slight reduction in the EN metric. This can be attributed to the replacement of infrared background information, which typically shows higher pixel values, with visible background information that tends to yield lower EN values during the calculation process. Our method shows superior performance in terms of AE, VIF, and SF metrics, indicating its ability to preserve finer texture details and significant target information compared to the two-stage combined fusion method. While our approach ranks lower than DIVFusion in terms of the EN metric, it still possesses an advantage over the SOTA combined method.

4.4.2 | Enhanced by URetinex

In addition to KinD, we have also selected URetinex as the low-light enhancement pre-processing algorithm. URetinex differs from traditional Retinex-based methods by employing an unfolding optimization module to iteratively optimize the reflection and illumination components. The visual quality outcomes are presented in Fig. 6 s_2 .

As observed in the magnified red box in s_2 , URetinex successfully restored abundant contrast information in this intricate scene. However, GTF, FusionGAN, SDNet, and SwinFusion replaced this effective information with low-value infrared information during the fusion process. In comparison to this fusion approach, although our method does not yield a brighter scene, it preserves the highest amount of contrast information, resulting in improved visual perception.

The quantitative outcomes are presented in the second row of Table 2. Similar to the KinD combination, our method achieves the highest scores on AG, VIF, and SF, while ranking below DIVFusion on EN. As previously discussed, our method effectively incorporates the background details from the enhanced visible images while disregarding irrelevant infrared information. Consequently, our approach shows an average performance in EN metric. However, both subjective and objective assessments consistently indicate that our method outperforms this combination.

4.4.3 | Enhanced by SCI

To explore the compatibility of our fusion strategy with the enhancement module CEM, we choose one model from SCI, which has an identical structure to CEM, as the low-light pre-processing algorithm. The visual outcomes are depicted in Fig. 6 s_3 .

As evident in the highlighted region within the green box in s_3 , the enhanced markings on the ground, which were challenging to discern in the original visible image, are effectively recovered by SCI. However, GTF, FusionGAN, SDNet, and RFN-Nest fail to preserve this level of detail. It is noteworthy that although the combination of PIAFusion and SCI manages to retain the recovered details, it falls short of our method in terms of capturing salient target information, which is demonstrated in the red box of s_3 . DIVFusion, on the other hand, shows undesirable enhancement performance in some specific darker regions. Hence, our approach demonstrates superior adaptability when combined with the SCI-based enhancement algorithm.

The quantitative results of the experiments are presented in the last row of Table 2. Our approach achieves the best scores in VIF and SF metrics, while slightly trailing behind PIAFusion in AG metric. This discrepancy can be attributed to the fact that PIAFusion extracts excessive information from visible image through its illumination-aware network but causes a loss of target information which can be observed in Fig. 6 s_3 . Our ranking in the EN metric aligns with the findings discussed in Section 4.3.2, attributing the same reason for the observed results.

TABLE 2 Quantitative results of the two-stage enhancement and fusion experiment.

Enhancement Method: KinD							
	GTF	FusionGAN	SDNet	RFN-Nest	SwinFusion	PIAFusion	DIVFusion
AG	5.0187	3.0934	6.1588	3.5652	6.4287	7.4479	5.5227
EN	7.0776	6.5870	7.1689	7.1549	7.0816	7.0762	7.6067
VIF	0.7444	0.5266	0.7622	0.7924	0.9500	1.0327	1.0583
SF	0.0634	0.0370	0.0703	0.0358	0.0746	0.0856	0.0603
Ours							
							9.5508
							7.2479
							1.1875
							0.1134
Enhancement Method: URetinex-Net							
	GTF	FusionGAN	SDNet	RFN-Nest	SwinFusion	PIAFusion	DIVFusion
AG	5.2115	3.2629	6.5755	3.8947	6.5709	7.7929	5.5227
EN	7.0672	6.6321	7.0858	7.1698	7.1696	7.1782	7.6067
VIF	0.7477	0.5255	0.7235	0.8301	0.9639	1.0067	1.0583
SF	0.0369	0.0370	0.0726	0.0371	0.0724	0.0841	0.0603
Ours							
							9.5508
							7.2479
							1.1875
							0.1134
Enhancement Method: SCI							
	GTF	FusionGAN	SDNet	RFN-Nest	SwinFusion	PIAFusion	DIVFusion
AG	6.7785	3.9077	8.2195	4.5803	8.5190	10.0590	5.5227
EN	7.0466	6.5751	7.1704	7.1475	7.3324	7.3593	7.6067
VIF	0.7284	0.5045	0.7066	0.8353	0.9821	1.0415	1.0583
SF	0.0832	0.0459	0.0921	0.0452	0.0959	0.1116	0.0603
Ours							
							9.5508
							7.2479
							1.1875
							0.1134

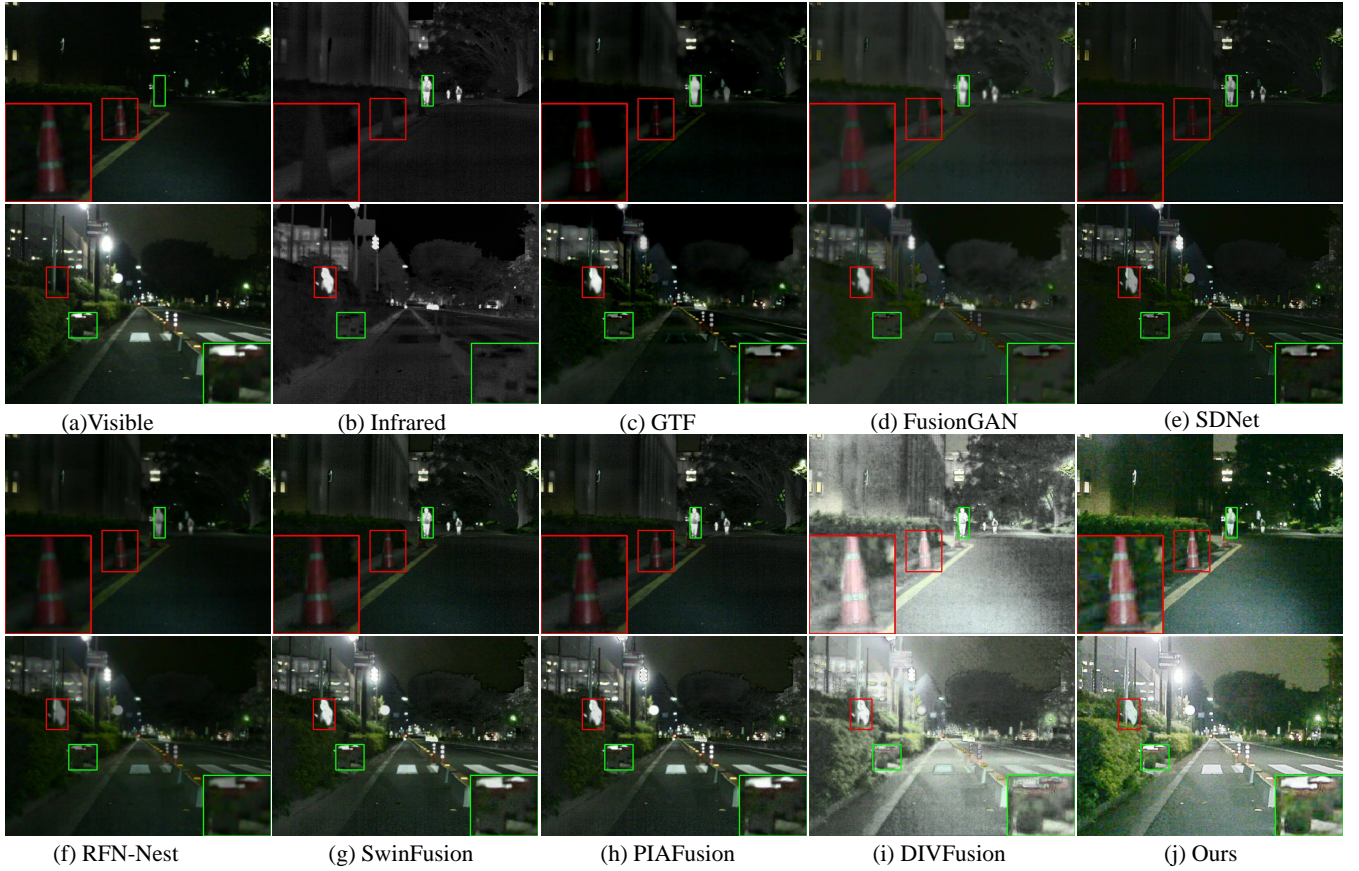


FIGURE 7 Visual quality comparison of our method with seven SOTA fusion methods on two image pairs (denoted as g_1 and g_2 from top to bottom) from the MSRS dataset .

TABLE 3 Quantitative results of generalization experiments.

	AG	EN	VIF	SF
GTF	1.4178	3.9251	0.4836	0.0241
FusionGAN	1.3141	5.4013	0.4962	0.0161
SDNet	1.9916	4.6086	0.4179	0.0294
RFN-Nest	2.7922	6.8931	0.7447	0.0300
SwinFusion	2.9566	6.2466	0.9815	0.0361
PIAFusion	3.1062	6.2630	0.9838	0.0380
DIVFusion	4.6666	7.5189	0.8261	0.0474
Ours	4.9385	7.0800	1.0048	0.0573

4.5 | Generalization experiment

For the purpose of verifying the generalization capability of our strategy to other datasets, we conducted generalization experiments on MSRS dataset. The visualization outcomes of the generalization experiment are presented in Fig. 7.

The visual results illustrate that our architecture effectively enhances the ambient lighting and highlight the salient target. In particular, the detailed information recovery achieved by our method surpasses that of conventional algorithms, as illustrated in the enlarged red box in g_1 . Notably, our method significantly enhances the contrast of the scene, capturing intricate details that are not reproduced by other conventional methods. Additionally, while DIVFusion also extracts scene information from dark areas, our method shows a more natural green color compared to the slight color distortion observed in DIVFusion. Similar observations can be made in the enlarged green box in g_2 .

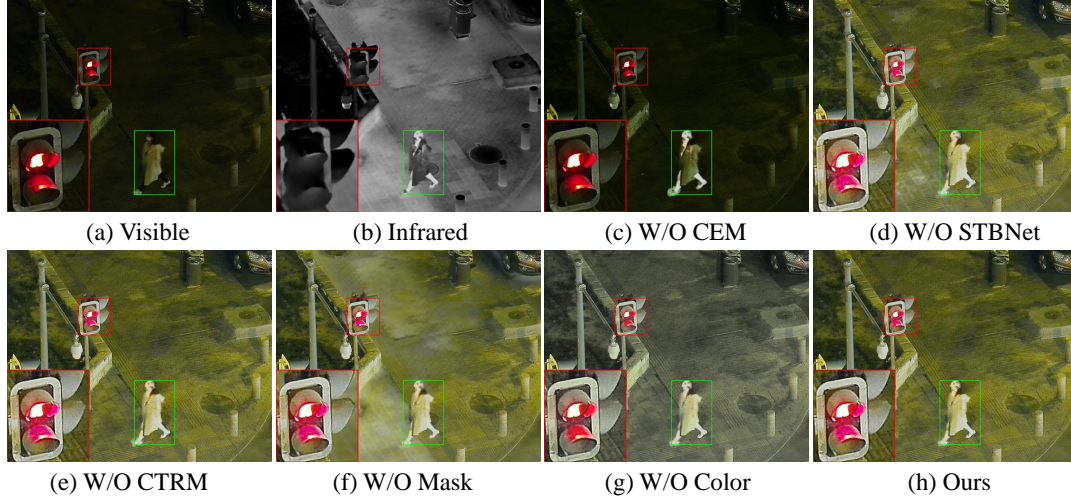


FIGURE 8 Visual results of ablation studies on the specific modules and loss functions.

TABLE 4 Quantitative results of ablation studies.

	AG	EN	VIF	SF
W/O CEM	4.9168	6.2711	0.9964	0.0664
W/O STBNet	9.4967	7.2649	1.1724	0.1125
W/O CTRM	9.5471	7.2473	1.1866	0.1133
W/O Mask	8.6387	7.3065	1.0455	0.1035
W/O Color	9.5101	7.2177	1.1872	0.1131
Ours	9.5508	7.2479	1.1875	0.1134

Table 3 presents the quantitative results of the four metrics. Our method attains the highest rankings in AG, VIF, and SF, while placing second below DIVFusion in EN. The superior performance in AG and SF highlights our ability to preserve texture detail and environmental information effectively. Furthermore, the best VIF metric confirms the exceptional visual perception offered by our method. It is important to note that DIVFusion employs histogram equalization as label in its illumination enhancement procedure, which contributes to higher EN score. In conclusion, our method demonstrates potential in generalization ability.

4.6 | Ablation study

To demonstrate the validity of the specific designs in our method, we conducted additional ablation studies to assess the influence of components such as CEM, STBNet, CTRM, I_{ir}^{mask} in \mathcal{L}_{tag} and \mathcal{L}_{sup} , and the enhanced color component on the fusion results.

4.6.1 | Ablation study for CEM

Given that our model involves both low-light enhancement and image fusion tasks, a crucial issue is how much the contrast enhancement module, namely CEM, affects the fusion results. To explore this matter, we conducted an experiment by excluding the CEM module. As illustrated in Fig. 8(c), the fusion outcomes of the proposed method without CEM, similar to conventional methods, exhibit a lack of texture detail recovery in the visible image, leading to fusion results with dark scene. In contrast, our proposed approach effectively illuminates the scene and restores significant details, thereby enhancing the visual quality of the fused image.

TABLE 5 Computational complexity and efficiency comparison with seven SOTA fusion methods.

	Parameters(M)	FLOPs(G)	FPS
GTF	-	-	0.0530
FusionGAN	0.9256	0.1136	0.5012
SDNet	0.0671	0.0088	0.7307
RFN-Nest	7.5242	0.1111	0.4715
SwinFusion	0.9275	0.0637	0.0496
PIAFusion	1.1761	0.0770	1.0220
DIVFusion	4.4028	0.7727	0.2170
Ours	0.6866	0.0450	0.5480

4.6.2 | Ablation study for STBNet

In our fusion network architecture, the Swin Transformer-based backbone network (STBNet) plays a crucial role in enhancing the network's target awareness capabilities. When we substituted the Swin Transformer in STBNet with a conventional channel attention module, the fusion result, *i.e.*, Fig. 8(d), showed unsatisfactory infrared noise in some regions. This phenomenon can be attributed to the limited ability of the normal channel attention module to effectively discern between the background and target regions, leading to the inclusion of redundant infrared information.

4.6.3 | Ablation study for CTRM

During the process of reconstructing the fused image, we employ the Contrast-Texture Retention Module (CTRM) to effectively guide the network in enhancing both the detailed texture and contrast information. When we exclude the entire CTRM and replace it with a single convolutional layer, the resulting fusion output is depicted in Fig. 8 (e). In contrast to our fusion results, the overall contrast of the image experiences a slight reduction with the omission of the CTRM, and there is also a minor loss of detail information.

subsubsectionAbation study for Mask As detailed in Section 3.3.2, our target perception is primarily achieved by data-driven training, relying on the I_{ir}^{mask} in \mathcal{L}_{tag} and \mathcal{L}_{aux} . To explore the impact of the mask, we replaced it with the original infrared image and remained the whole network unchanged. Subsequently, the fused image obtained by the model without the mask is presented in Fig. 8 (f). In comparison to our method employing a mask, the fusion output with the mask indicates a precise separation between the visible and infrared backgrounds, leading to an increased presence of texture detail information. Additionally, as indicated by the zoom red box Fig. 8 (f), the absence of the mask prevents the extraction of meaningful information from the tactile paving.

4.6.4 | Ablation study for Color Component

As stated in Eq.(10), during the conversion of the fused image to the RGB space, we concatenate the CbCr channels from the illumination enhanced version of visible image, instead of utilizing the CbCr channels from the original images. This is due to the incompatibility between the enhanced Y-channel and the original color channels. In this ablation experiment, we opted to apply the unchanged color channels into the color conversion process, and visual comparisons are depicted in Fig. 8(g). Notably, it is evident from the figure that when concatenate the color channel of the original visible image, the overall color of the resulting image tends towards a grayish tone, leading to the color distortion and an unfavorable visual perception.

4.6.5 | Quantitative analysis

Table 4 displays the quantitative assessments from the ablation studies. we can discover that our method shows the superior performance on AG, VIF, SF, which demonstrates the effectiveness of the specific designs in our IETAFAfusion. On the EN metric, our method is not the best, and the value of EN increases when the model removes STBNet and Mask. However, this can be reasonably explained. As discussed previously, the removal of STBNet and Mask causes a reduction in the network's ability

to perceive salient targets and backgrounds, leading to the introduction of redundant infrared information. Consequently, this contributes to an increase in the EN metric and a decrease in other evaluation metrics.

4.7 | Computational complexity analysis

To assess the computational complexity of our model, we conducted a comparative analysis of model parameters with SOTA methods. We chose floating-point operations per second (FLOP_s), and frames per second (FPS) as the quantitative evaluation of the comparison. The results of this comparison are presented in Table 5. The findings shows that our model parameters rank second in terms of quantity and computational complexity, primarily attributed to the lightweight architecture of our CEM module, which significantly reduces computational costs. Regarding computational efficiency, as measured by FPS, our method ranks third. This result is reasonable considering that our network incorporates transformers, which inherently requires more computational resources compared to conventional methods and CNN-based strategy.

5 | CONCLUSION

In this paper, we propose a comprehensive network for fusing infrared and visible images, which integrates the process of contrast enhancement and image fusion with target awareness. Specifically, the CEM is used to predict the illumination component and minimize the illumination degradation in the visible image. Subsequently, we employ the STBNet to extract and fuse details and contextual information from the source images. Furthermore, in the image reconstruction stage, CTRM is utilized to preserve contrast information and enhance the details of the fused image. To ensure compatibility between the two enhancement tasks, the masks are integrated into the loss function to bolster the network's ability to discern targets and scenes effectively. Our approach is evaluated through both quantitative and qualitative comparisons with SOTA fusion methods, showcasing its capability to generate visually brighter scenes while preserving more texture details and improving overall visual perception. Furthermore, the two-stage enhancement-fusion experiments affirm the harmonious interaction between our enhancement and fusion processes. Generalization experiments are carried out to prove the robustness of our method on the MSRS dataset. Finally, the ablation studies substantiate the efficacy of the specific designs employed in IETAFusion, contributing positively to the overall fusion outcomes.

ACKNOWLEDGMENTS

This work was partially supported by the funding from the Research Result of the Key Laboratory of Ultra HD Video Technology Application in Fusion Publishing, National Press and Publication Administration UHD-ZD-202306.

DATA AVAILABILITY STATEMENT

The data are available on request.

CONFLICT OF INTEREST

The authors declare no potential conflict of interests.

References

- Bai, X., Zhou, F., & Xue, B. (2011). Fusion of infrared and visual images through region extraction by using multi scale center-surround top-hat transform. *Optics express*, 19(9), 8444–8457.
- Bavirisetti, D. P., & Dhuli, R. (2016). Two-scale image fusion of visible and infrared images using saliency detection. *Infrared Physics Technology*, 76, 52–64. Retrieved from <https://www.sciencedirect.com/science/article/pii/S1350449515300955> doi: <https://doi.org/10.1016/j.infrared.2016.01.009>
- Ben Hamza, A., He, Y., Krim, H., & Willsky, A. (2005). A multiscale approach to pixel-level image fusion. *Integrated Computer-Aided Engineering*, 12(2), 135–146.
- Cao, Y., Guan, D., Huang, W., Yang, J., Cao, Y., & Qiao, Y. (2019). Pedestrian detection with unsupervised multispectral feature learning using deep neural networks. *information fusion*, 46, 206–217.

- Chen, J., Li, X., Luo, L., Mei, X., & Ma, J. (2020). Infrared and visible image fusion based on target-enhanced multiscale transform decomposition. *Information Sciences*, 508, 64–78.
- Cho, S.-J., Ji, S.-W., Hong, J.-P., Jung, S.-W., & Ko, S.-J. (2021). Rethinking coarse-to-fine approach in single image deblurring. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 4641–4650).
- Fu, Y., Wu, X.-J., & Durrani, T. (2021). Image fusion based on generative adversarial network consistent with perception. *Information Fusion*, 72, 110–125.
- Jia, X., Zhu, C., Li, M., Tang, W., & Zhou, W. (2021). Llvp: A visible-infrared paired dataset for low-light vision. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 3496–3504).
- Jian, L., Yang, X., Liu, Z., Jeon, G., Gao, M., & Chisholm, D. (2020). Sedrfuse: A symmetric encoder–decoder with residual block network for infrared and visible image fusion. *IEEE Transactions on Instrumentation and Measurement*, 70, 1–15.
- Li, H., & Wu, X.-J. (2018). Densefuse: A fusion approach to infrared and visible images. *IEEE Transactions on Image Processing*, 28(5), 2614–2623.
- Li, H., Wu, X.-J., & Durrani, T. (2020). Nestfuse: An infrared and visible image fusion architecture based on nest connection and spatial/channel attention models. *IEEE Transactions on Instrumentation and Measurement*, 69(12), 9645–9656.
- Li, H., Wu, X.-J., & Kittler, J. (2021). Rfn-nest: An end-to-end residual fusion network for infrared and visible images. *Information Fusion*, 73, 72–86.
- Li, J., Huo, H., Li, C., Wang, R., & Feng, Q. (2020). Attentionfgan: Infrared and visible image fusion using attention-based generative adversarial networks. *IEEE Transactions on Multimedia*, 23, 1383–1396.
- Liu, G., Lin, Z., Yan, S., Sun, J., Yu, Y., & Ma, Y. (2012). Robust recovery of subspace structures by low-rank representation. *IEEE transactions on pattern analysis and machine intelligence*, 35(1), 171–184.
- Liu, Y., Chen, X., Cheng, J., Peng, H., & Wang, Z. (2018). Infrared and visible image fusion with convolutional neural networks. *International Journal of Wavelets, Multiresolution and Information Processing*, 16(03), 1850018.
- Liu, Y., Liu, S., & Wang, Z. (2015). A general framework for image fusion based on multi-scale transform and sparse representation. *Information Fusion*, 24, 147–164. Retrieved from <https://www.sciencedirect.com/science/article/pii/S1566253514001043> doi: <https://doi.org/10.1016/j.inffus.2014.09.004>
- Long, Y., Jia, H., Zhong, Y., Jiang, Y., & Jia, Y. (2021). Rxdnfuse: A aggregated residual dense network for infrared and visible image fusion. *Information Fusion*, 69, 128–141.
- Ma, J., Chen, C., Li, C., & Huang, J. (2016). Infrared and visible image fusion via gradient transfer and total variation minimization. *Information Fusion*, 31, 100–109.
- Ma, J., Liang, P., Yu, W., Chen, C., Guo, X., Wu, J., & Jiang, J. (2020). Infrared and visible image fusion via detail preserving adversarial learning. *Information Fusion*, 54, 85–98.
- Ma, J., Ma, Y., & Li, C. (2019). Infrared and visible image fusion methods and applications: A survey. *Information Fusion*, 45, 153–178. doi: <https://doi.org/10.1016/j.inffus.2018.02.004>
- Ma, J., Tang, L., Fan, F., Huang, J., Mei, X., & Ma, Y. (2022). Swinfusion: Cross-domain long-range learning for general image fusion via swin transformer. *IEEE/CAA Journal of Automatica Sinica*, 9(7), 1200–1217.
- Ma, J., Tang, L., Xu, M., Zhang, H., & Xiao, G. (2021). Stdffusionnet: An infrared and visible image fusion network based on salient target detection. *IEEE Transactions on Instrumentation and Measurement*, 70, 1–13.
- Ma, J., Xu, H., Jiang, J., Mei, X., & Zhang, X.-P. (2020). Ddrgan: A dual-discriminator conditional generative adversarial network for multi-resolution image fusion. *IEEE Transactions on Image Processing*, 29, 4980–4995.
- Ma, J., Yu, W., Liang, P., Li, C., & Jiang, J. (2019). Fusiongan: A generative adversarial network for infrared and visible image fusion. *Information fusion*, 48, 11–26.
- Ma, J., Zhou, Z., Wang, B., & Zong, H. (2017). Infrared and visible image fusion based on visual saliency map and weighted least square optimization. *Infrared Physics & Technology*, 82, 8–17.
- Ma, L., Ma, T., Liu, R., Fan, X., & Luo, Z. (2022). Toward fast, flexible, and robust low-light image enhancement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5637–5646).
- Tang, L., Xiang, X., Zhang, H., Gong, M., & Ma, J. (2023). Divfusion: Darkness-free infrared and visible image fusion. *Information Fusion*, 91, 477–493.
- Tang, L., Yuan, J., Zhang, H., Jiang, X., & Ma, J. (2022). Piafusion: A progressive infrared and visible image fusion network based on illumination aware. *Information Fusion*, 83, 79–92.
- Wei, C., Wang, W., Yang, W., & Liu, J. (2018). Deep retinex decomposition for low-light enhancement. *arXiv preprint arXiv:1808.04560*.

- Wu, W., Weng, J., Zhang, P., Wang, X., Yang, W., & Jiang, J. (2022). Uretinex-net: Retinex-based deep unfolding network for low-light image enhancement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5901–5910).
- Xu, H., Zhang, H., & Ma, J. (2021). Classification saliency-based rule for visible and infrared image fusion. *IEEE Transactions on Computational Imaging*, 7, 824–836.
- Yang, W., Yuan, Y., Ren, W., Liu, J., Scheirer, W. J., Wang, Z., ... et al. (2020). Advancing image understanding in poor visibility environments: A collective benchmark study. *IEEE Transactions on Image Processing*, 29, 5737–5752. doi: 10.1109/TIP.2020.2981922
- Zhang, F., Shao, Y., Sun, Y., Zhu, K., Gao, C., & Sang, N. (2021). Unsupervised low-light image enhancement via histogram equalization prior. *arXiv preprint arXiv:2112.01766*.
- Zhang, H., & Ma, J. (2021). Sdnet: A versatile squeeze-and-decomposition network for real-time image fusion. *International Journal of Computer Vision*, 129, 2761–2785.
- Zhang, Q., Fu, Y., Li, H., & Zou, J. (2013). Dictionary learning method for joint sparse representation-based image fusion. *Optical Engineering*, 52(5), 057006–057006.
- Zhang, Y., Zhang, J., & Guo, X. (2019). Kindling the darkness: A practical low-light image enhancer. In *Proceedings of the 27th ACM international conference on multimedia* (pp. 1632–1640).