electronics

MDPI

Article

ESFuse: Weak Edge Structure Perception Network for Infrared and Visible Image Fusion

Wuyang Liu [1,2], Haishu Tan [3,4], Xiaoqi Cheng [1,2,*] and Xiaosong Li [2,3,4,*]

[1] School of Mechatronic Engineering and Automation, Foshan University, Foshan 528000, China; 2112203023@stu.fosu.edu.cn
[2] Guangdong Provincial Key Laboratory of Industrial Intelligent Inspection Technology, Foshan University, Foshan 528000, China
[3] School of Physics and Optoelectronic Engineering, Foshan University, Foshan 528225, China; tanhaishu@fosu.edu.cn
[4] Guangdong-HongKong-Macao Joint Laboratory for Intelligent Micro-Nano Optoelectronic Technology, Foshan University, Foshan 528225, China
* Correspondence: chexqi@163.com (X.C.); lixiaosong@fosu.edu.cn (X.L.)

Abstract: Infrared and visible image fusion (IVIF) fully integrates the complementary features of different modal images, and the fused image provides a more comprehensive and objective interpretation of the scene compared to each source image, thus attracting extensive attention in the field of computer vision in recent years. However, current fusion methods usually center their attention on the extraction of prominent features, falling short of adequately safeguarding subtle and diminutive structures. To address this problem, we propose an end-to-end unsupervised IVIF method (ESFuse), which effectively enhances fine edges and small structures. In particular, we introduce a two-branch head interpreter to extract features from source images of different modalities. Subsequently, these features are fed into the edge refinement module with the detail injection module (DIM) to obtain the edge detection results of the source image, improving the network's ability to capture and retain complex details as well as global information. Finally, we implemented a multiscale feature reconstruction module to obtain the final fusion results by combining the output of the DIM with the output of the head interpreter. Extensive IVIF fusion experiments on existing publicly available datasets show that the proposed ESFuse outperforms the state-of-the-art(SOTA) methods in both subjective vision and objective evaluation, and our fusion results perform well in semantic segmentation, target detection, pose estimation and depth estimation tasks. The source code has been availabled.

Keywords: infrared and visible image fusion; weak edge structure perception; multiscale feature

Citation: Liu, W.; Tan, H.; Cheng, X.; Li, X. ESFuse: Weak Edge Structure Perception Network for Infrared and Visible Image Fusion. *Electronics* **2024**, *13*, 4115. https://doi.org/10.3390/electronics13204115

Academic Editor: Silvia Liberata Ullo

Received: 2 September 2024
Revised: 11 October 2024
Accepted: 15 October 2024
Published: 18 October 2024

1. Introduction

Images captured by a single sensor or within a single shooting setup can only represent the imaging scene from a limited perspective, and conventional image processing techniques often struggle to acquire multiple pieces of information simultaneously [1]. Therefore, IVIF is essential for generating richer and clearer images. Recently, IVIF has been widely applied in scene understanding [2], saliency detection [3], and pedestrian re-identification [4].

Numerous image fusion methods have been developed over the past few decades. Traditional image fusion techniques include multiscale decomposition [5–7], sparse representation [8,9], and hybrid methods [10–12]. However, these traditional methods primarily rely on pixel-level operations, and their lack of semantic understanding and effective analysis of complex environments can result in fusion outputs that lack semantic consistency and plausibility. With the rise of deep learning in recent years, end-to-end image fusion solutions have demonstrated significant potential. Mainstream deep learning methods

can be broadly categorized into three types: auto-encoder (AE), generative adversarial networks (GANs), and convolutional neural networks (CNNs). Recent learning-based approaches leverage CNNs [13–17] to address the shortcomings of traditional methods. The feature extraction capability of CNNs excels in capturing valuable information about image features. However, CNNs struggle with the challenge of lacking ground truth data, necessitating sophisticated training strategies to enhance their fusion capabilities. AE-based methods first train auto-encoders as feature extractors on large natural image datasets. The feature extractor is then utilized to extract complementary information from multi-modal images, merging these features using specific fusion rules [18], such as splicing [19] and element summation [20,21]. However, the AE-based fusion framework is not fully learnable, as it employs handcrafted fusion rules to combine depth features. Consequently, Ma et al. [22] were the first to define image fusion as a game between a generator and a discriminator. Specifically, they constrained the probability distribution between the fused and source images to ensure the fused image possesses rich texture details. However, excessively strong constraints may introduce artificial textures into the fused image, limiting the realism and accuracy of the fusion results. Recently, algorithmic unfolding models have garnered significant attention. The core principle behind these models is that mathematical optimization algorithms guide the construction of the network framework, enhancing the interpretability of neural networks and effectively avoiding time-consuming empirical network design. The most commonly used models for IVIF tasks include convolutional sparse coding models [23] and learned low-rank representation models [24]. However, these methods lack flexibility and may struggle to handle varying magnification factors, blurring kernels, and other variations.

Significant modal differences exist between infrared and visible images due to variations in wavelengths, radiation sources, and acquisition sensors. These differences manifest in texture, luminance, and structure, impacting the quality of image fusion. When such modal differences lead to inconsistent feature fusion, the overall quality of the fusion result often deteriorates. Decomposition-based representation methods can align the feature spaces of images from different domains, reducing the impact of modal differences. However, these methods typically require complex decomposition and fusion rules, and when dealing with complex inter-domain transformations or significant modal differences, feature space alignment may fail to capture the intricate relationships both within and between domains. Additionally, many methods prioritize the extraction of salient features, neglecting subtle image texture information from intermediate layers, which is crucial for effective model learning [25]. Although dense connections [26] have been introduced in fusion networks, they often result in increased computational costs.

Therefore, the shortcomings of the current research can be summarized in the following five points. (1) Lack of semantic understanding: Traditional image fusion methods primarily rely on pixel-level operations, lacking effective analysis and semantic understanding of complex environments. (2) Over-reliance on ground truth: CNN-based methods struggle in the absence of ground truth and require complex training strategies to enhance fusion capabilities. (3) Limitations of manual fusion rules: While AE methods can extract features, their manual fusion rules limit learning capacity and adaptability to complex scenes. (4) Inadequate handling of modal differences: Current methods often fail to capture complex feature relationships effectively when addressing significant modal differences between infrared and visible images. (5) High computational cost: Although dense connections are introduced to extract more information, they result in a significant computational overhead.

To address these challenges, we propose a novel weak edge structure perception network (ESFuse). To generate high-quality fused images, our method first employs a two-branch head interpreter to extract common features and reduce feature differences between the infrared and visible light modalities while preserving their unique information. The two-branch outputs from the head interpreter are then combined and processed by an edge refinement module, which captures feature location information from the source images of each modality and integrates these data into the fusion process to enhance

performance. Second, since the performance of a fusion network relies heavily on the spatial location of its features, we propose the DIM. This module utilizes the deep semantic information extracted by the head interpreter module, along with the edge detection results from the edge refinement, to update the gradient map using residuals and generate attention weights. These attention weights emphasize the complementary regions of the source image, thereby enhancing the network's performance. Finally, we implement a multiscale feature reconstruction module, which combines the outputs of the DIM and the head interpreter to produce the final fusion results. Our method outperforms existing SOTA approaches both quantitatively and qualitatively on publicly available datasets. Additionally, we demonstrate the effectiveness of our approach in various downstream tasks, including semantic segmentation, target detection, pose estimation, and depth estimation. We also conduct ablation studies to evaluate the effectiveness of each component of our approach. In summary, the contributions of this paper are as follows:

- We propose a novel two-branch unsupervised end-to-end IVIF model that effectively predicts weak edge structure and texture features in different modal images, and reduces the feature differences between modalities to preserve the unique information of each modal image.
- We propose a feature reconstruction module that optimizes the fusion process by comprehensively preserving the details and structural features by multiscale computation; thus, the fused image can more realistically reflect the information of the source images, significantly improving the fusion performance.
- We propose using the DIM to highlight the features in the complementary regions of the source images by multiplying the attention weights generated from the gradient map with the depth semantic information.The DIM can effectively enhance the importance and expressiveness of their features by means of fine-tuning them, and efficiently preserves and exploits the semantic structure.
- Extensive experiments cover IVIF image fusion, and downstream tasks such as semantic segmentation, target detection, pose estimation and depth estimation, the corresponding subjective and quantitative evaluation results consistently demonstrate the competing SOTA performance of the proposed ESFuse.

2. Related Work

In this section, we review various IVIF methods, categorizing them into traditional, AE-based, GAN-based and diffusion model-based methods.

2.1. Traditional-Based Methods

In the study of traditional methods for IVIF, various techniques have been proposed, including multiscale decomposition and saliency detection [27]. Multiscale decomposition methods [28–31] decompose and reconstruct the features of infrared and visible images at various levels to better fuse details and structures. These approaches align the processing of scale information with the human visual system. Saliency detection methods [32] enhance fusion performance for important targets by assigning higher weights to salient regions or objects. Sparse representation techniques [33] utilize dictionaries learned from large sets of images to encode and preserve essential information from the source images during the fusion process. Together, these traditional approaches lay the groundwork for IVIF, enabling the retention of image details and improvement of visual quality.

2.2. AE-Based Methods

The core idea of auto-encoder AE-based IVIF algorithms is to achieve information extraction and fusion from multiple input modalities by learning a compact representation (coding) and the reconstruction process of an image. Li et al. [34] first proposed the DenseFuse method, connecting the output of each layer in the encoder to every other layer to obtain more features from the source image. Following the introduction of DenseFuse, AE-based IVIF methods have seen significant development. Huang et al. [35] enhanced the

fusion and denoising capabilities of the model by employing decomposition techniques. Tang et al. [19] utilized convolutional neural networks (CNNs) as the basic units for the encoder and decoder, constructing a global feature extraction module embedded after feature encoding through the use of a transformer. This combination of CNNs and transformers is a common approach in AE-based models, effectively capturing local and global features of multimodal images, respectively.

2.3. GAN-Based Methods

The GAN-based IVIF algorithm primarily achieves multimodal information extraction through the adversarial training of generators and discriminators. Ma et al. [22] were the first to extend GANs to the IVIF task by balancing the fusion of infrared intensity information and visible gradient information during the adversarial process, thereby demonstrating the effectiveness of generative adversarial networks in infrared and visible image fusion. Additionally, Xu et al. [36] developed a conditional GAN with dual discriminators, each trained on infrared and visible images, respectively. This approach effectively balances the features of the two image types, thus enhancing fusion performance. Notable architectural innovations based on the GAN approach have emerged, as researchers have explored the use of multiple discriminators to improve fusion outcomes. For instance, Song et al. [37] introduced a novel GAN-based method for IVIF that employs a triple discriminator to generate detailed fused images.

2.4. Diffusion Model-Based Methods

Recently, diffusion models have been enhanced to produce images of higher quality than previous generative models, such as GANs [38,39]. This approach generates high-quality images by simulating the diffusion process, which transforms noise-corrupted images into clean ones, thereby alleviating common issues such as training instability and mode collapse associated with GANs. Zhao et al. [40] were the first to propose the use of Denoising Diffusion Probabilistic Models (DDPM) for fusion tasks, employing a hierarchical Bayesian approach to model the subproblem of maximum likelihood estimation. Xu et al. [41] introduced the FS-Diff diffusion model, which is based on a stochastic iterative denoising process for the task of tri-modal image fusion with super-resolution. Additionally, DPS [42] employs the Laplace approximation to compute the log-likelihood gradient for posterior sampling, effectively addressing many noisy nonlinear inverse problems.

3. Method

In this section, we introduce the head interpreter, edge refinement, detail injection module, and fusion module. The framework of the proposed method is illustrated in Figure 1.

3.1. Head Interpreter

The head interpreter is designed to learn and comprehend the embeddings of high-level image semantics. Therefore, a large receptive field is essential for effectively encoding structures that exhibit variations in scale. To address this requirement, we employed convolutional layers with varying kernel sizes to fully leverage contextual information. As shown in Figure 1, by utilizing a mixture of convolutional kernel sizes (3, 5, and 7) instead of a fixed size of 3, we obtained dense receptive fields of different dimensions. To facilitate information propagation in convolutional neural networks (CNNs), we combined residual connections with convolutional layers of varying kernel sizes, forming a residual block. Additionally, we utilized the spatial and channel attention block (SCAB) [43] to generate an attention map for visible image information, which we then multiplied element-wise with the infrared feature map. This process injects crucial information captured by the visible image into the infrared image features, thereby aiding the head interpreter in extracting these features.

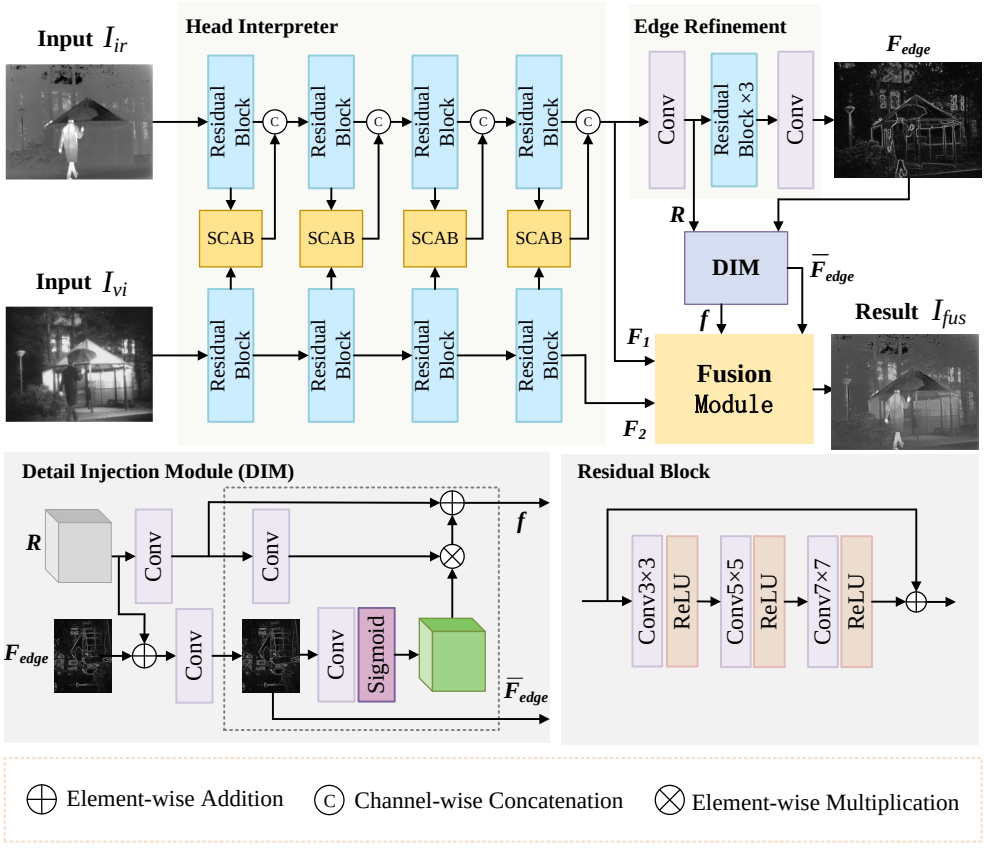

Figure 1. Workflow of the proposed framework, head interpreter is our proposal to implicitly learn contextual information from the source image using residual block. To further highlight the features of the source image, we propose edge refinement and the DIM, which utilizes the attention mechanism to regulate the head interpreter. The last part is the fusion model, which utilizes the learned head interpreter and the DIM to improve the fusion performance.

The input paired infrared and visible images are denoted as $I_{ir} \in R^{H\times W}$ and $I_{vi} \in R^{H\times W\times 3}$, respectively, and the head interpreter is denoted as $H(\cdot)$. The head interpreter aims at extracts the respective depth features $\{F_1,F_2\}$ from the infrared and visible inputs $\{I_{ir},I_{vi}\}$, which are formulated as follows:

$$\{F_1, F_2\} = H(I_{ir}, I_{vi}) \tag{1}$$

3.2. Spatial and Channel Attention Block

As shown in Figure 1, adaptive feature enhancement and fusion is performed using SCAB after each multi-layer feature extraction in the head interpreter.

As shown in Figure 2, the SCAB consists of two cascaded attention units. The node data I_1 and I_2 from the head interpreter are processed by the 1D channel attention graph $\mathbf{M}_c \in \mathbb{R}^{C_i\times 1\times 1}$ and a 2D spatial attention graph $\mathbf{M}_s \in \mathbb{R}^{1\times H_i\times W_i}$. The channel attention first performs global maximum pooling and global average pooling of spatial dimensions on an input feature map I of size $H_i \times W_i \times C_i$ to obtain two $1 \times 1 \times C_i$ feature maps; then, the results of global maximum pooling and global average pooling are fed into a shared multilayer perceptual machine (MLP) for learning, respectively, to obtain two $1 \times 1 \times C_i$ feature maps. Finally, the outputs of the MLP are summed up and then go through the mapping process of the Sigmoid activation function to obtain the channel attention weight matrix. The channel attention process can be summarized as follows:

$$\mathbf{M}_c(\mathbf{I}) = \sigma\Big[MLP(AvgPool(\mathbf{I})) + (MLP(MaxPool(\mathbf{I}))\Big] \tag{2}$$

$$\mathbf{I}' = \mathbf{M}_c(\mathbf{I}) \otimes \mathbf{I} \tag{3}$$

where $\otimes$ denotes the element-wise multiplication, σ denotes sigmoid function, C_i, H_i, W_i denote the number of channels, height, and width of input I_1 and I_2. $AvgPool(\cdot)$ and $MaxPool(\cdot)$ denote average pooling and maximum pooling with a step size of 2, respectively. Before multiplication, the attention maps $M_c(I)$ are stretched to the size of $\mathbf{M}_c(I) \in \mathbb{R}^{C_i \times H_i \times W_i}$.

Similar to the channel attention process, spatial attention is paid to the fact that the input feature map of size $H \times W \times C$ is first subjected to global maximum pooling and global average pooling in the channel dimension to obtain two feature maps of size $H \times W \times 1$. Then, the results of global maximum pooling and global average pooling are spliced according to the channel to obtain a feature map of size $H \times W \times 2$, and finally, a 7×7 convolution operation is performed on the spliced results to obtain a feature map of size $H \times W \times 1$, and then the Sigmoid activation function is used to obtain the spatial attention weight matrix. The spatial attention process can be summarized as follows:

$$M_s(\mathbf{I}') = \sigma(f^{7\times7}([AvgPool(\mathbf{I}'); MaxPool(\mathbf{I}')])) \tag{4}$$

$$\mathbf{I}'' = \mathbf{M}_s(\mathbf{I}') \otimes \mathbf{I}' \tag{5}$$

where $f^{7\times7}$ represents a convolutional operation with the filter size of 7×7. The attention maps $M_s(I)$ are also stretched to the size of $\mathbf{M}_s(I) \in \mathbb{R}^{C_i \times H_i \times W_i}$ before multiplication.

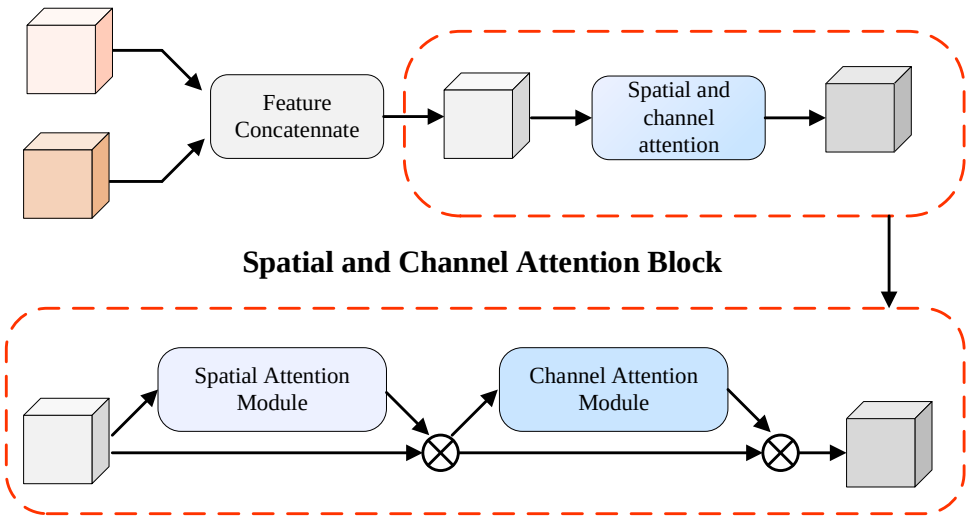

Figure 2. Spatial and channel attention block. The SCAB is used to reduce the semantic gap at the multi-layer feature fusion stage in the head interpreter.

3.3. Edge Refinement

CNNs are typically more adept at identifying the texture of an image than its shape [25]. This characteristic simplifies the learning process; however, it compromises the robustness when applied to real-world scenarios. To improve the retention of subtle and small-scale structures within an image, we implemented an edge refinement module. This module was designed to bridge the gap between the fusion outcome and the actual structural layers of the original image. Rather than using existing edge prediction methods to generate pseudo ground truth, this aims to give the model an overall improved ability to capture and preserve complex detail, thus enhancing its generalisation and utility. Edge refinement (denoted by $E(\cdot)$) aims to extract the edge details of the source image and is formulated as

$$F_{edge} = E(F_1) \tag{6}$$

As shown in Figure 1, this module explores more new features by encouraging the edge refinement module to learn residual from the input image. The refinement module comprises two convolutional layers and three cascaded residual blocks.

3.4. Detail Injection Module

In the image fusion task, the performance of a fused image depends heavily on the spatial location of its features. Therefore, an image fusion model should be able to handle individual regions with different priorities. Using a head interpreter to represent the spatial information of the regions to be fused, we further emphasize the importance of these regions in the attention mechanism. Specifically, we propose a detail injection module (DIM). It uses the deep semantic information extracted by the head interpreter module and updates the gradient map with residuals. The updated gradient map is then used to generate attention weights, which increase the importance of the complementary regions of the source image and suppress other irrelevant regions. Finally, the features within the complementary regions are emphasized by multiplying the elements of deep semantic information and attention weights. The DIM can be described as follows:

$$\{\overline{F}_{\text{edge}}, f\} = D(F_{\text{edge}}, R) \tag{7}$$

where F_{edge} and R are the final gradient map and the deep semantic information extracted from the residual blocks in the edge refinement module, respectively. $D(\cdot)$ denotes our detail injection module. $\overline{F}_{\text{edge}}$ and f are the modulation outputs and feature tensor, respectively, ready to be fed into the feature reconstruction module.

3.5. Fusion Module

To retain as much structural information as possible in the final fused image, we used multiscale reconstruction. As shown in Figure 3, multiscale reconstruction uses residual blocks combined in a parallel fashion to capture structural information at different scales in the source image and enhance the fusion effect. Denoting the fusion module as $MU(\cdot)$, which is formulated as follows:

$$I_{\text{fus}} = MU\{(F_1 + \overline{F}_{\text{edge}}), (F_2 + \overline{F}_{\text{edge}}), f\} \tag{8}$$

where f is the a prior output of the DIM, F_1, F_2, $\overline{F}_{\text{edge}}$ are the outputs of the head interpreter and update edge refinement, respectively.

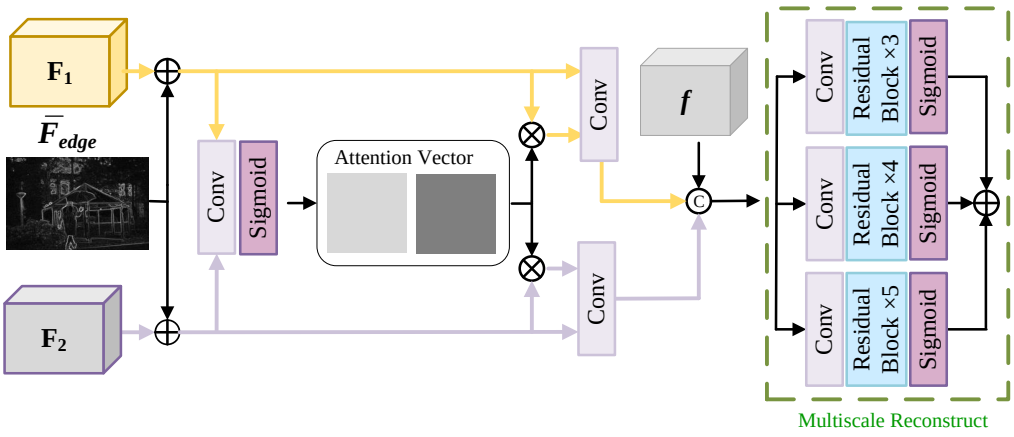

Figure 3. Workflow of the fusion module, which uses the learned head interpreter and the DIM as inputs and adds multiscale reconstruct at the end to emphasize the recovery of the source image structure.

3.6. Loss Function

Fusion loss: Constraining the network using the structural similarity index measure (SSIM) [44] is an effective approach for preserving luminance information. The SSIM

considers luminance, contrast, and structural metrics, which agree with the perception of human vision. L_{SSIM} is expressed by the following loss function:

$$L_{\text{SSIM}} = (1 - \text{SSIM}(I_{\text{fus}}, I_{\text{ir}})) + (1 - \text{SSIM}(I_{\text{fus}}, I_{\text{vi}})) \tag{9}$$

In addition, we expected the fused image to maintain an optimal intensity distribution consistent with that of the source image. We designed the following intensity loss L_{int} to guide our fusion model in capturing the appropriate intensity information:

$$L_{\text{int}} = \|\max(I_{\text{ir}}, I_{\text{vi}}), I_{\text{fus}}\|_1 \tag{10}$$

Here, $\|(\cdot)\|_1$ denotes the L_1 norm and $max(\cdot)$ denotes the element-by-element maximum selection.

Edge loss: The main idea of our network was to introduce additional effective constraints on the texture details. For edge detection, we used the L_1 norm to constrain the detailed prediction results directly, which can be written as

$$L_e = \left\| \nabla I_{\text{ir}}, F_{\text{edge}} \right\|_1 + \left\| \nabla I_{\text{vi}}, F_{\text{edge}} \right\|_1 \tag{11}$$

Here, $\nabla(\cdot)$ denotes the gradient map in the horizontal and vertical directions, and in this study, we utilize the Sobel operator to compute the gradient map.

Total loss: The ultimate goal is to achieve the combined goal of edge detection and image fusion. The overall loss function is expressed as

$$L_{\text{total}} = \lambda_1 L_{\text{SSIM}} + \lambda_2 L_{\text{int}} + \lambda_3 L_e \tag{12}$$

Here, λ_1, λ_2, and λ_3 are hyperparameters that control the trade-offs of each sub-loss term. In this study, we empirically set them to 5, 10, and 3, respectively.

4. Experiments

4.1. Datasets

We used four popular benchmarks to validate our fusion model: MSRS [45], Roadscene [46], TNO [47] and M3FD [14]. We trained our network on the MSRS training set (1083 pairs). The test datasets included the MSRS (361 pairs), RoadScene (221 pairs), TNO (107 pairs) and M3FD (300 pairs) test sets. In total, we use 1083 image pairs for training, and 989 for testing. Among these datasets, RoadScene includes both daytime and nighttime road scenarios, MSRS and M3FD offer a variety of scenarios across different conditions, while TNO focuses specifically on infrared and visible light data in military applications.

4.2. Implementation Details

The source images of the training dataset were cropped into 64 × 64 patches to increase data volume. The parameters were updated using the Adam optimizer at a learning rate of 1×10^{-4}. Training was performed for 30 epochs, with a batch size of 32. The proposed approach was implemented in PyTorch1.12.1 with two NVIDIA 3090 GPUs for training. The source code is available at https://github.com/lwy12345678/ESFuse.

4.3. Evaluation Metrics

To evaluate the image fusion capability of our model, we employed four objective evaluation metrics: normalized mutual information (Q_{MI}), phase congruency (Q_P), average gradient (AG), Chen–Blum Metric (Q_{CB}), and Piella's Metric (Q_S). Higher values for these metrics generally indicate better quality in the fused images. These metrics assess various aspects of the merged images, including the amount of information conveyed, fidelity to the original source images, and overall visual quality.

Q_{MI} is a modification of mutual information (MI). Specifically, MI is a quantitative measure of the mutual dependence of two variables. The definition of mutual information for two discrete random variables U and V is

$$MI(U;V) = \sum_{v \in V} \sum_{u \in U} p(u,v) \log_2 \frac{p(u,v)}{p(u)p(v)}, \tag{13}$$

where $p(u,v)$ is the joint probability distribution function of U and V , and $p(u)$ and $p(v)$ are the marginal probability distribution functions of U and V , respectively. Actually, MI quantifies the distance between the joint distribution of U and V, i.e., $p(u,v)$, and the joint distribution when U and V are independent, i.e., $p(u)p(v)$. Mutual information can be equivalently expressed with joint entropy $\{H(U,V)\}$ and marginal entropy $\{H(U,V)\}$ the two variable U and V as

$$MI(U,V) = H(U) + H(V) - H(U,V), \tag{14}$$

where

$$\begin{aligned} H(U) &= -\sum_{u} p(u) \log_2 p(u), \\ H(V) &= -\sum_{v} p(v) \log_2 p(v), \\ H(U,V) &= -\sum_{u,v} p(u,v) \log_2 p(u,v). \end{aligned}$$

Qu et al. [48] used the summation of the MI between the fused image $F(i,j)$ and two input images, $A(i,j)$ and $B(i,j)$, to represent the difference in quality. The expression of the MI-based fusion performance measure M_F^{AB} is

$$\begin{aligned} M_F^{AB} &= MI(A,F) + MI(B,F) \\ &= \sum_{i,j} \left(h_{AF}(i,j) \log_2 \frac{h_{AF}(i,j)}{h_A(i) h_F(j)} \right. \\ &\quad \left. + h_{BF}(i,j) \log_2 \frac{h_{BF}(i,j)}{h_B(i) h_F(j)} \right), \end{aligned} \tag{15}$$

where $h_{AF}(i,j)$ indicates the normalized joint gray level histogram of images $A(i,j)$ and $F(i,j)$; $h_K(i,j)(K = A, B,$ and $F)$ is the normalized marginal histogram of images A, B, or F, respectively. However, (15) mixes two joint entropies measured at different scales. This can lead to unstable measurements, so Hosny et al. [49] modified (15) as follows:

$$Q_{MI} = 2 \left[\frac{MI(A,F)}{H(A) + H(F)} + \frac{MI(B,F)}{H(B) + H(F)} \right]. \tag{16}$$

We used Hossny's definition in our experiments.

Q_P is an image feature-based metric. Zhao et al. [50] and Liu et al. [51] used the phase congruency, which provides an absolute measure of image feature, to define an evaluation metric. The metric is defined as a product of three correlation coefficients:

$$Q_P = (P_p)^{\alpha} (P_M)^{\beta} (P_m)^{\gamma}, \tag{17}$$

where p, M, m refers to phase congruency (p), maximum, and minimum moments, respectively, and there are

$$\begin{aligned} P_p &= \max\left(C_{AF}^p, C_{BF}^p, C_{SF}^p\right), \\ P_M &= \max\left(C_{AF}^M, C_{BF}^M, C_{SF}^M\right), \\ P_m &= \max\left(C_{AF}^m, C_{BF}^m, C_{SF}^m\right). \end{aligned}$$

Herein, C_{xy}^k, $\{k|p,M,m\}$ stands for the correlation coefficients between two sets x and y:

$$C_{xy}^k = \frac{\sigma_{xy}^k + C}{\sigma_x^k \sigma_y^k + C}, \tag{18}$$

$$\sigma_{xy} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y}). \tag{19}$$

The suffixes A, B, F, and S correspond to the two inputs, fused image, and maximum-select map. The exponential parameters α, β, and γ can be adjusted based on the importance of the three components.

Q_{CB} is a human perception-inspired fusion metric. There are five steps involved:

Contrast sensitivity filtering: Filtering is implemented in the frequency domain. Image $I_A(i,j)$ is transformed into the frequency domain and we obtain $I_A(m,n)$. The filtered image is obtained: $\tilde{I}_A(m,n) = I_A(m,n)S(r)$, where $S(r)$ is the CSF filter in polar form with $r = \sqrt{m^2+n^2}$. In [52], there are three choices suggested for CSF, which include Mannos–Sakrison, Barton, and DoG filter.

Local contrast computation: Peli's contrast is defined as

$$C(i,j) = \frac{\phi_k(i,j) * I(i,j)}{\phi_{k+1}(i,j) * I(i,j)} - 1. \tag{20}$$

A common choice for ϕ_k would be

$$G_k(x,y) = \frac{1}{(\sqrt{2\pi}\sigma_k)} e^{\frac{x^2+y^2}{2\sigma_k^2}}, \tag{21}$$

with a standard deviation $\sigma_k = 2$.

Contrast preservation calculation: The masked contrast map for input image $I_A(i,j)$ is calculated as

$$C'_A = \frac{t(C_A)^p}{h(C_A)^q + Z}. \tag{22}$$

Here, t, h, p, q, and Z are real scalar parameters that determine the shape of the nonlinearity of the masking function.

Saliency map generation: The saliency map for $I_A(i,j)$ defined as

$$\lambda_A(i,j) = \frac{{C'_A}^2(i,j)}{{C'_A}^2(i,j) + {C'_B}^2(i,j)}. \tag{23}$$

The information preservation value is computed as

$$Q_{AF}(i,j) = \begin{cases} \frac{C'_A(i,j)}{C'_F(i,j)}, & \text{if } C'_A(i,j) < C'_F(i,j), \\ \frac{C'_F(i,j)}{C'_A(9i,j)}, & \text{otherwise.} \end{cases} \tag{24}$$

Global quality map:

$$Q_{GQM}(i,j) = \lambda_A(i,j)Q_{AF}(i,j) + \lambda_B(i,j)Q_{BF}(i,j). \tag{25}$$

The metric value is obtained by average the global quality map, i.e., $Q_{CB} = \overline{Q_{GQM}(i,j)}$.

Q_S is an image-structure similarity-based metric. Piella and Heijmans [53] defined three fusion quality indices. Assume the local $Q(A,B|w)$ value is calculated in a sliding window w. There are

$$Q_S = \frac{1}{|W|}\sum_{w\in W}[\lambda(w)Q_0(A,F|w) + (1-\lambda(w))Q_0(B,F|w)], \tag{26}$$

$$Q_W = \sum_{w\in W} c(w)[\lambda(w)Q_0(A,F|w) + (1-\lambda(w))Q_0(B,F|w)], \tag{27}$$

$$Q_E = Q_W(A,B,F)\cdot Q_W(A',B',F')^{\alpha}, \tag{28}$$

where the weight $\lambda(w)$ is defined as

$$\lambda(w) = \frac{s(A|w)}{s(A|w) + s(B|w)}. \tag{29}$$

Herein, $s(A|w)$ is a local measure of image salience. In Piella's implementation, $s(A|w)$ and $s(B|w)$ are the variance of images A and B within the window w, respectively. The coefficient $c(w)$ in (27) is

$$c(w) = \frac{\max[s(A|w), s(B|w)]}{\sum_{w' \in W}[s(A|w'), s(B|w')]}. \tag{30}$$

In (28), $Q_W(A', B', F')$ is the Q_w calculated with the edge images, i.e., A', B', and F' 0, and is a manually adjustable parameter to weight the edge-dependent information.

AG is a non-reference metric, which calculates the average gradient value across all pixels, representing the overall spatial variation (sharpness) in the images.

$$AG = \frac{1}{MN} \sum_{i=1}^{M} \sum_{j=1}^{N} \sqrt{\frac{\nabla F_x^2(i,j) + \nabla F_y^2(i,j)}{2}}. \tag{31}$$

where $\nabla F_x(i,j) = F(i,j) - F(i+1,j)$ and $\nabla F_y(i,j) = F(i,j) - F(i+1,j)$. A higher AG metric indicates an enhanced presence of gradient information within the fused image, implying that the algorithm effectively integrates gradient details, potentially leading to superior performance in image fusion tasks.

These five metrics can comprehensively measure the performance of the fusion method. The higher the value of all metrics, the better.

4.4. Qualitative Comparison

The results of the qualitative comparison are presented in Figure 4. For each of the four datasets, we selected a set of images to compare the supervised fusion performance. To effectively distinguish the fusion effects of different methods across various scenarios, we chose diverse scenes, including nighttime, highway, battlefield, and street, from the respective datasets.

From the nighttime scene of the MSRS dataset, it is evident that most methods exhibit poor robustness under low-light conditions (see the red boxes in Figure 4). For instance, in the fused images produced by TarADL, U2Fusion, ReCoNet, and LRRNet, the human figure is difficult to recognize. In comparison, UMF-CMGR and MURF perform worse than DEFusion and the proposed method in terms of detail retention and character contrast. Only ESFuse and DEFusion achieved better fusion results in the night scenes.

In the highway scene from the RoadScene dataset, a comparison between the infrared and visible light images reveals that the character information on the road sign primarily originates from the visible light image. However, aside from LRRNet, most of the compared methods fail to retain the character information from the source visible image (see the yellow boxes in Figure 4). This indicates that most existing methods struggle to balance information between source images, often causing the fused image to be biased toward one source image, thereby losing complementary features across different modalities. In contrast, our ESFuse method enhances the complementary region features of the source images through the DIM, resulting in better fusion outcomes in the highway scene.

In the battlefield scene from the TNO dataset, efficient target discrimination is a crucial challenge. As shown in the orange box in Figure 4, the fusion results from U2Fusion, MURF, LRRNet, ReCoNet, and UMF contain less information, exhibit lower brightness, and retain more visible light data, leading to incomplete fusion of the visible light images. The targets in the fusion results of TarDAL and DeFusion are not prominent. In contrast, our ESFuse method fuses more information through a unified feature space, resulting in more prominent targets. Furthermore, due to the incorporation of multiscale reconstruction in the fusion module, our method produces results with clearer structures and improved contrast.

Figure 4. Qualitative comparisons of the SOTA methods with the proposed ESFuse on the Roadscene, MSRS, TNO, and M3FD datasets.

The green box in Figure 4 highlights the street scene from the M3FD dataset. The fused image produced by TarDAL contains more infrared information but lacks visible light details. In the fusion results from U2Fusion, MURF, and LRRNet, the overall brightness is relatively low, making the objects in the fused image less prominent. Although ReCoNet, DeFusion, and UMF-CMGR yield brighter fusion results, they exhibit over-smoothing, leading to reduced clarity and poor handling of object edges (e.g., tree edges in the green box). In contrast, our ESFuse method demonstrates superior fusion performance in street scenes, effectively integrating source information from both infrared and visible images. Our approach also provides enhanced background edge texture and improved character contrast.

We utilized remotely sensed data from natural environments, built landscapes, and urban scenes to evaluate the performance of our method. Figure 5 presents the fused images from these scenes. Our fusion technique effectively integrates valuable information from

the source images, achieving satisfactory results in terms of illumination, detail, and structural integrity. The fused images in the first, second, and third columns demonstrate our method's capability to successfully merge infrared and visible data with enhanced detail and structural clarity, as highlighted by the red boxes. Additionally, our approach enhances useful information while preserving critical features, as evidenced by the farmland scene in the first column. Despite the visible images in the second and third columns being somewhat dark and containing subtle details, our fusion results retain these details without being compromised by the unusual lighting conditions. Overall, our method is robust in preserving key information across various scenes and lighting conditions.

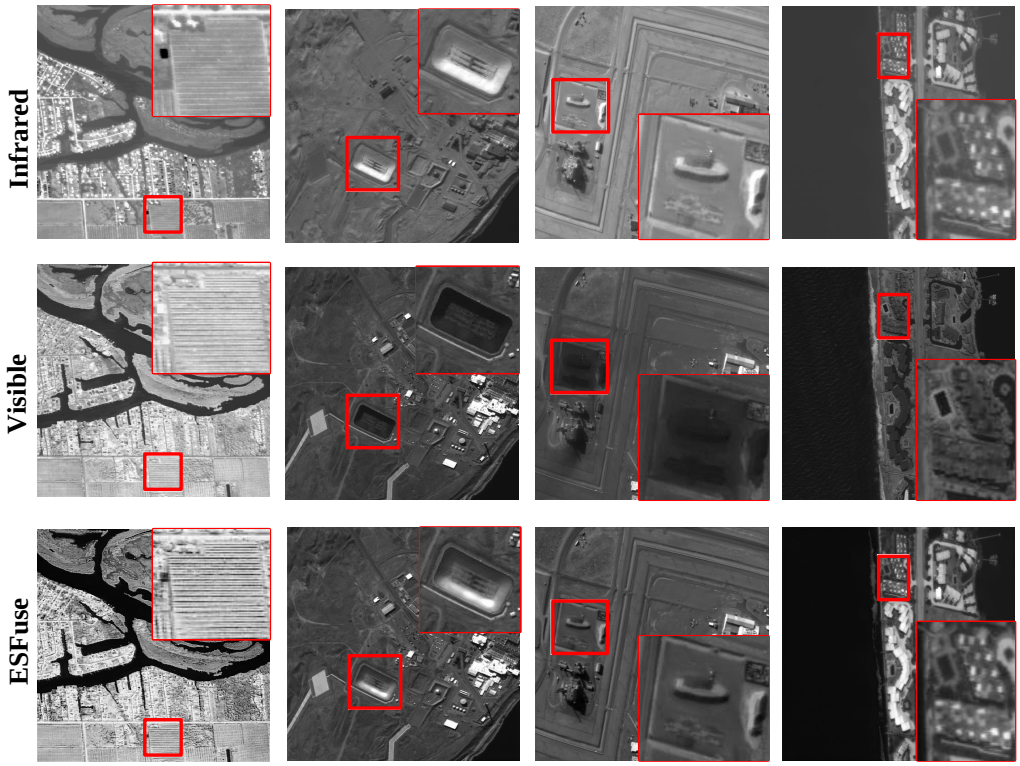

Figure 5. Fusion results in remote sensing imagery. The red boxes are enlarged to highlight the fusion performance on image details.

4.5. Quantitative Comparison

Table 1 provides quantitative comparisons between our method and the SOTA methods on the MSRS, RoadScene, TNO, and M3FD datasets, respectively, and summarizes the average metric values for our ESFuse and these methods. Our ESFuse stands out in terms of overall performance.

On the MSRS and M3FD datasets, our ESFuse scores very highly on all five metrics. In the RoadScene dataset, our method performs well in Q_{MI}, Q_P, and Q_S, and ranks second in AG, and in the TNO dataset, we obtain the best results in Q_{MI}, Q_P, Q_S, and AG. It shows that the information of the source image is effectively integrated in our ESFuse while preserving the rich details in the fused image. It further confirms its excellent overall performance. In addition, Defusion achieves the second best results for Q_{MI} and Q_{CB} on the MSRS dataset, but it relies on complex decomposition algorithms and faces challenges in preserving the rich information of the source images. In contrast, U2Fusion achieves the second best scores for Q_P and Q_{CB} on the TNO dataset and Q_S and Q_{CB} on the M3FD dataset by combining the properties of the different similarities between the source images, but it relies on a specific multitask-oriented loss function and does not retain the source image contrast and details well.

Table 1. Quantitative comparison of the proposed algorithm and different comparison methods on the Roadscene, MSRS,TNO and M3FD datasets. **Bold** is the best, red is the second.

Methods	Venue	MSRS				
		Q_{MI}	Q_S	Q_P	Q_{CB}	AG
MURF [13]	*TPAMI 2023*	0.2575	0.6626	0.2242	0.3665	3.0720
LRRNET [24]	*TPAMI 2023*	0.4607	0.7110	0.3372	0.3928	2.6509
U2Fusion [54]	*TPAMI 2020*	0.3834	0.7599	0.3141	0.4686	2.3203
ReCoNet [55]	*ECCV 2022*	0.3927	0.3705	0.3760	0.3780	3.0006
Defusion [56]	*ECCV 2022*	0.4702	0.7483	0.3757	0.5142	2.6539
TarDAL [14]	*CVPR 2022*	0.3841	0.4855	0.1639	0.4088	1.7156
UMF-CMGR [57]	*IJCAI 2022*	0.3278	0.6560	0.2104	0.3510	2.1364
ESFuse	-	**0.5690**	**0.7714**	**0.4771**	**0.5247**	**3.3181**
		Road Sence				
MURF [13]	*TPAMI 2023*	0.3489	0.8017	0.3370	0.4852	**6.6474**
LRRNET [24]	*TPAMI 2023*	0.3936	0.5942	0.2460	0.5096	4.6399
U2Fusion [54]	*TPAMI 2020*	0.3604	**0.8135**	0.3705	**0.5178**	4.6377
ReCoNet [55]	*ECCV 2022*	0.4341	0.7414	0.3049	0.4785	3.8011
Defusion [56]	*ECCV 2022*	0.4085	0.7583	0.2870	0.4982	3.5582
TarDAL [14]	*CVPR 2022*	0.4536	0.7462	0.3038	0.4060	4.1746
UMF-CMGR [57]	*IJCAI 2022*	0.4183	0.8133	0.4022	0.5029	4.0954
ESFuse	-	**0.5486**	0.7581	**0.4536**	0.4986	4.7343
		TNO				
MURF [13]	*TPAMI 2023*	0.2297	0.7809	0.2077	0.4906	4.7921
LRRNET [24]	*TPAMI 2023*	0.3479	0.6771	0.2348	0.5514	4.3380
U2Fusion [54]	*TPAMI 2020*	0.2791	**0.8333**	0.2657	0.5592	3.4423
ReCoNet [55]	*ECCV 2022*	0.3304	0.7771	0.2291	0.5403	3.0014
Defusion [56]	*ECCV 2022*	0.3027	0.7671	0.1465	0.5165	2.2102
TarDAL [14]	*CVPR 2022*	0.4110	0.7684	0.2394	0.4402	3.3909
UMF-CMGR [57]	*IJCAI 2022*	0.2836	0.8178	0.2503	0.4990	2.7568
ESFuse	-	**0.4845**	0.7852	**0.3167**	**0.5575**	**4.8459**
		M3FD				
MURF [13]	*TPAMI 2023*	0.3468	0.7973	0.1816	0.4931	3.4529
LRRNET [24]	*TPAMI 2023*	0.3917	0.8447	0.3218	0.4479	2.4001
U2Fusion [54]	*TPAMI 2020*	0.4202	0.8790	0.4132	0.5481	2.7574
ReCoNet [55]	*ECCV 2022*	0.4498	0.8421	0.3304	0.4510	2.7424
Defusion [56]	*ECCV 2022*	0.4348	0.8512	0.3585	0.4677	1.8383
TarDAL [14]	*CVPR 2022*	0.4777	0.8257	0.2584	0.4030	1.8877
UMF-CMGR [57]	*IJCAI 2022*	0.4655	0.8624	0.4200	0.4273	2.0321
ESFuse	-	**0.5582**	**0.8832**	**0.4653**	**0.5487**	**3.5437**

4.6. Task-Driven Evaluation

To assess the generalization of the proposed approach to advanced tasks, we used (1) DeepLabV3+ [58], a semantic segmentation model, (2) YOLOv5 [59], a target detection model, (3) OpenPose [60], a pose estimation model, and (4) MiDaS [61], a depth estimation model to measure the contribution of various fusion algorithms to advanced visual tasks.

Figure 6 presents the visualization results of semantic segmentation and target detection on the MSRS dataset. Since both DeepLabV3+ and YOLOv5 are trained on visible image datasets, a fusion method that mitigates the interference from thermal radiation in infrared images yields better segmentation and detection performance. Our method preserves structural information while enhancing texture details and effectively fuses the complementary information from both infrared and visible images. For instance, as shown in Figure 6, the car in the yellow box and the pedestrian in the red box demonstrate how our approach significantly improves the scene understanding for segmentation and target detection models.

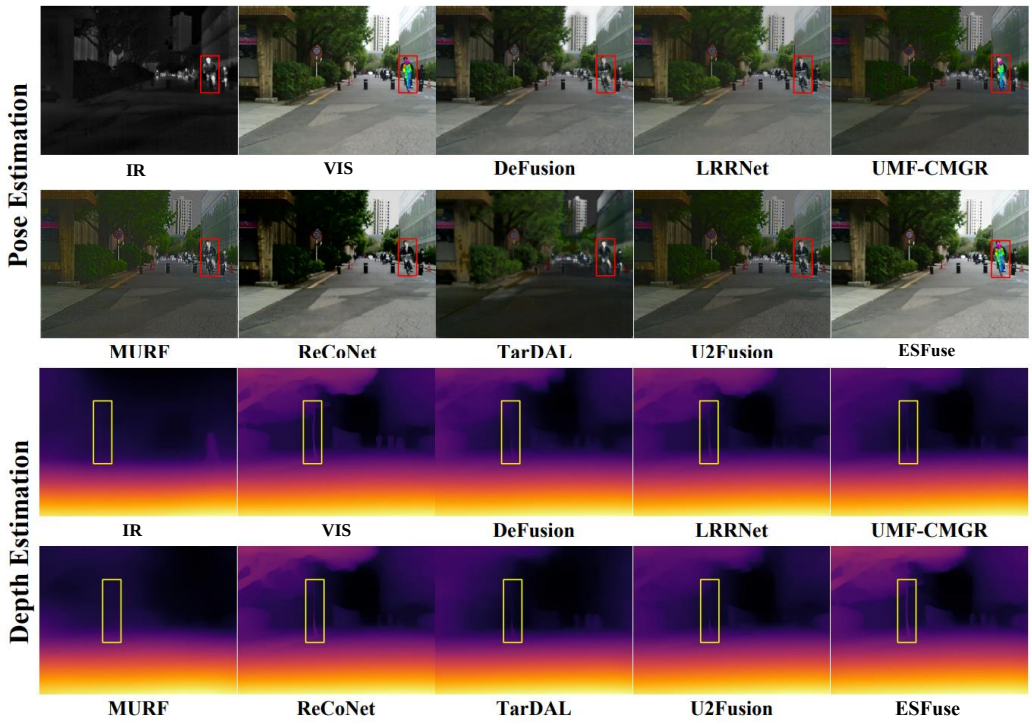

Figure 6. Visualization results of pose estimation and depth estimation for infrared images, visible images, and different fusion images on MSRS dataset.

The visualization results for pose estimation and depth estimation are presented in Figure 7. In the human pose estimation task, inaccuracies in the spatial structure of figures obtained by other methods degrade the quality of the fused images, resulting in incomplete or inaccurate outcomes. This can lead to confusion and incorrect conclusions in applications such as action recognition and video surveillance. In the depth estimation task, the performance results of fused images obtained through different methods exhibited significant discrepancies, attributable to variations in fusion accuracy, which caused varying degrees of visual differences. In contrast, the stabilized fusion achieved by our proposed method offers a robust representation of the objects within the image.

The results of the quantitative comparison are presented in Tables 2 and 3. To ensure a fair assessment, we retrained the segmentation network [62] on the MFNet dataset. It is important to note that since the MSRS dataset only provides segmented labels, we retrained YOLOv5 on the M3FD dataset for a quantitative comparison of target detection. Segmentation performance is measured using pixel intersection and union (IoU), while mean average precision (mAP) is utilized to evaluate detection performance, where AP@0.5 indicates the mAP value for an IoU threshold of 0.5.

From the results, it is evident that infrared (IR) images perform well in detecting people, as indicated by both the mean Intersection over Union (mIoU) and average precision at a 0.5 IoU threshold (AP@0.5). This suggests that IR images offer the detector ample semantic information regarding salient targets, such as individuals. However, the detection results for cars in infrared images are disappointing. Conversely, visible images supply the detector with substantial semantic information about vehicles. Our ESFuse effectively achieves intensity preservation and texture retention through the guidance of the detail injection module (DIM) and multiscale reconstruction, successfully balancing complementary and irrelevant information between the source images. Consequently, as illustrated in Tables 2 and 3, our algorithm achieves the highest IoU in semantic segmentation across nearly all categories, ranks first in mIoU, and also demonstrates higher accuracy in target detection. In summary, the proposed method opens up new possibilities for applications such as augmented reality (AR), autonomous driving, and coastline safety monitoring.

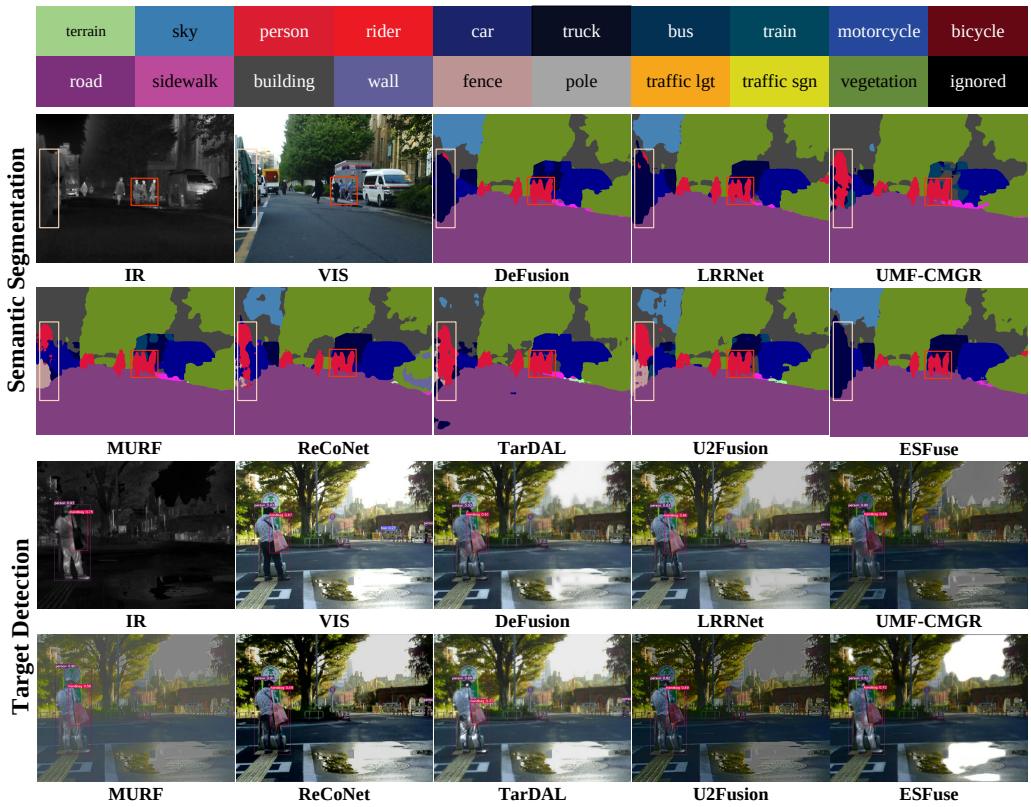

Figure 7. Visualization results of semantic segmentation and target detection for infrared images, visible images, and different fusion images on MSRS dataset.

Table 2. The detection accuracy of the comparison methods and ESFuse on various categories in the M3FD dataset. **Bold** is the best, red is the second, blue is the third.

Method	Background	Car	Person	Bike	Curve	Car Stop	Color Tone	AP@0.5
Infrared Image	0.944	0.586	0.806	0.184	0.867	0.821	0.000	0.384
Visible Image	0.974	0.873	0.407	0.823	0.660	**0.555**	0.481	0.682
MURF	0.979	0.871	0.739	0.823	0.645	0.515	0.466	0.720
TarDAL	0.982	0.888	0.811	0.827	0.660	0.550	0.464	0.740
U2Fusion	0.981	0.880	0.823	0.815	**0.667**	0.429	0.261	0.694
ReCoNet	0.982	0.886	**0.823**	**0.829**	0.659	0.540	0.462	0.740
Defusion	0.980	0.871	0.820	0.797	0.645	0.357	0.440	0.701
LRRNet	0.971	0.816	0.639	0.754	0.388	0.302	0.300	0.553
UMF-CMGR	0.981	0.884	0.819	0.820	0.656	0.514	0.464	0.734
ESFuse	**0.984**	**0.889**	0.805	0.827	0.659	0.543	**0.484**	**0.741**

Table 3. Comparison of segmentation accuracies of the method ESFuse on different classes of the MSRS dataset. **Bold** is the best, red is the second, blue is the third.

Method	Background	Car	Person	Bike	Curve	Car Stop	Guardrail	mIoU
Visible Image	97.92	86.79	39.97	70.51	53.33	71.84	**85.90**	72.32
Infrared Image	96.14	61.90	70.00	24.46	33.64	20.67	0.06	39.53
LRRNet	98.34	89.09	68.12	69.29	52.02	71.57	81.95	74.77
Defusion	98.46	89.11	73.82	**71.44**	**64.27**	73.21	80.59	76.03
UMF-CMGR	98.17	87.06	70.87	66.00	51.39	68.22	73.59	70.99
ReCoNet	97.56	83.12	56.55	57.38	37.84	55.91	77.91	64.73
U2Fusion	98.42	88.84	72.88	70.92	59.30	72.09	79.15	75.22
TarDAL	98.45	89.50	73.17	69.84	61.49	72.21	80.53	75.60
MURF	98.30	87.88	73.19	68.40	53.37	70.22	75.07	72.67
ESFuse	**98.50**	**89.64**	**74.40**	69.48	60.30	**73.46**	81.76	**76.33**

4.7. Qualitative Results of DIM

To further elucidate the effectiveness of our detail injection module (DIM) in the image fusion task, we present the visualizations of F_{edge} and $\overline{F}_{\text{edge}}$ in Figure 8. Our module successfully identifies key common features across different modalities and assigns higher weights to these regions. Notably, it can effectively recognize significant features even in low-light conditions with insufficient illumination. As illustrated in the right half of Figure 8, our DIM generalizes well to nighttime image fusion tasks. These results convincingly demonstrate the efficacy of the module and contribute to advancing the field of image fusion.

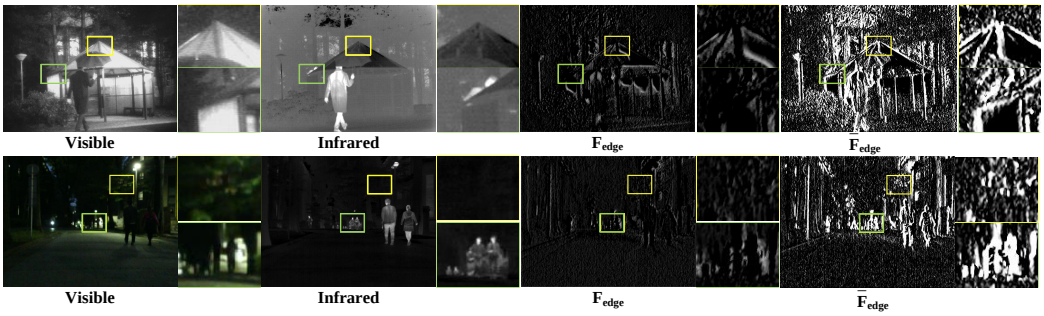

Figure 8. Visualizations of the DIM on infrared and visible images, where F_{edge} is the input of the DIM. For better comparison, two local areas are enlarged in each image.

4.8. Ablation Studies

To demonstrate the effectiveness of edge refinement, multiscale reconstruction, and the detail injection module (DIM), we conducted ablation experiments on the MSRS dataset. Specifically, for edge refinement, we removed it and replaced the detail map in the subsequent DIM input with the gradient map obtained using the Sobel operator. For the DIM, we excluded it entirely and fed the outputs f and F_{edge} directly into the fusion module instead of the original inputs R and F_{edge}. For multiscale reconstruction, we transformed its parallel three-branch structure into a single-branch structure consisting of four residual blocks.

As shown in Figure 9, the method lacking edge refinement struggles to perceive the information between the source images effectively, resulting in blurred characters. In contrast, our approach utilizing edge refinement yields detailed fusion results, particularly for salient features. Figure 10 illustrates that the method without multiscale reconstruction suffers from insufficient clarity in the figures and increased noise in the fused image. Compared to this method, our approach not only reduces noise but also preserves the details and structure of the image. Without multiscale reconstruction, the salient features and relationships across different layers are not learned effectively, making the reconstruction of key features challenging and leading to a lack of structural detail in the fused results. The DIM emphasizes features within the complementary region by generating attention weights from the gradient map, thereby enhancing the importance of these regions in the source images. As demonstrated in Figure 11, without DIM, details such as leaves and the floor appear less sharp. In contrast, our fusion results exhibit greater richness in detail and clarity.

Table 4 summarizes the results of the ablation experiments. As shown, the removal of the DIM led to a significant reduction in AG, highlighting its crucial role in the fusion process and in preserving rich information. Conversely, the absence of edge refinement and multiscale reconstruction resulted in decreased values for Q_{MI}, Q_S, Q_P, and Q_{CB}, suggesting that these proposed strategies effectively retain more information from the source images. Both qualitative and quantitative results demonstrate that the DIM, edge refinement, and multiscale reconstruction are vital for constructing fully fused images and enhancing image quality.

Figure 9. The visual comparison of ablation with and without the DIM. The yellow boxes are enlarged to highlight the fusion performance on image details.

Figure 10. The visual comparison of ablation with and without edge refinement. The yellow boxes are enlarged to highlight the fusion performance on image details.

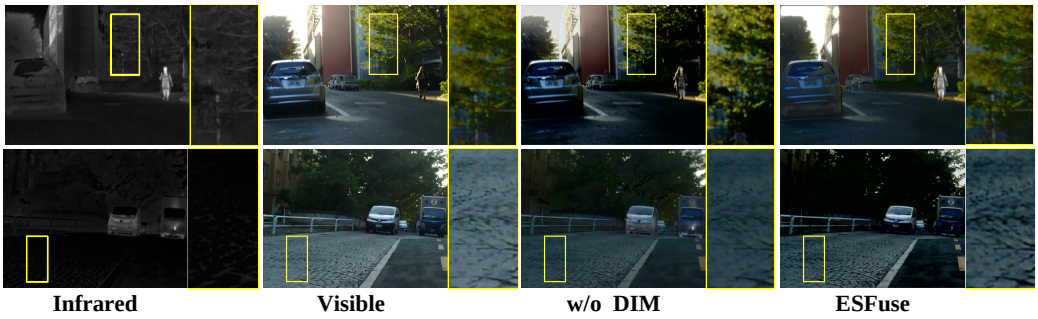

Figure 11. The visual comparison of ablation with and without multiscale reconstruction. The yellow boxes are enlarged to highlight the fusion performance on image details.

Table 4. Ablation experiments for edge refinement, DIM and multiscale reconstruction.

Methods	Q_{MI}	Q_S	Q_P	Q_{CB}	AG
w/o DIM	0.4696	0.6507	0.3838	0.5040	2.4739
w/o edge refinement	0.5394	0.7668	0.4382	0.5087	3.1434
w/o multiscale reconstruction	0.4994	0.7080	0.4119	0.5287	3.2434
ESFuse	**0.5690**	**0.7714**	**0.47709**	**0.5247**	**3.3181**

5. Conclusions

In this study, we propose a detailed enhanced infrared and visible image fusion (IVIF) network. First, to minimize the feature discrepancies between the different modalities of infrared and visible light while preserving their unique information, we design a deep feature extractor—referred to as the head interpreter—to extract common features from the source images. Subsequently, the two-branch outputs of the head interpreter are superimposed and fed into an edge refinement module, which reveals the feature location

information in the source images of different modalities, enhancing the overall performance in the fusion module. Next, we introduce the detail injection module, which increases the importance of the detailed features identified by the edge refinement module, enabling the fusion model to better recover ambiguous information across different modal features. Finally, the feature maps at varying scales are reconstructed into fused images using a multiscale reconstruction strategy. Both subjective and objective experimental results demonstrate that the proposed ESFuse exhibits exceptional image fusion performance, surpassing existing algorithms. Additionally, it performs well in downstream tasks such as semantic segmentation, target detection, pose estimation, and depth estimation, showcasing the robustness of our proposed approach.

Author Contributions: W.L.: conceptualization, methodology, software, validation, writing—original draft; H.T.: supervision, writing—review and editing; X.C.: data curation, funding acquisition, writing—review and editing; X.L.: resources, formal analysis, supervision, funding acquisition, writing—review and editing. All authors have read and agreed to the published version of the manuscript.

Funding: This study was supported by the Basic and Applied Basic Research of Guangdong Province under Grant 2023A1515140077, the Natural Science Foundation of Guangdong Province under Grant 2024A1515011880, the National Natural Science Foundation of China under Grants 62201149 and 62201151, the Guangdong Higher Education Innovation and Strengthening of Universities Project under Grant 2023KTSCX127, and the Research Fund of Guangdong-HongKong-Macao Joint Laboratory for Intelligent Micro-Nano Optoelectronic Technology under Grant 2020B1212030010.

Data Availability Statement: The source code of the paper is available at https://github.com/lwy12345678/ESFuse (accessed on 10 July 2024).

Conflicts of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Li, X.; Li, X.; Ye, T.; Cheng, X.; Liu, W.; Tan, H. Bridging the Gap Between Multi-Focus and Multi-Modal: A Focused Integration Framework for Multi-Modal Image Fusion. In Proceedings of the Winter Conference on Applications of Computer Vision WACV, Waikoloa, HI, USA, 3–8 January 2024; pp. 1628–1637.
2. Sagar, A.S.; Chen, Y.; Xie, Y.; Kim, H.S. MSA R-CNN: A comprehensive approach to remote sensing object detection and scene understanding. *Expert Syst. Appl.* **2024**, *241*, 122788. [CrossRef]
3. Xia, C.; Wang, J.; Ge, B. MLBSNet: Mutual Learning and Boosting Segmentation Network for RGB-D Salient Object Detection. *Electronics* **2024**, *13*, 2690. [CrossRef]
4. Wang, J.; Wang, J. MHDNet: A Multi-Scale Hybrid Deep Learning Model for Person Re-Identification. *Electronics* **2024**, *13* 1435. [CrossRef]
5. Li, H.; Wu, X.; Kittler, J. MDLatLRR: A Novel Decomposition Method for Infrared and Visible Image Fusion. *IEEE Trans. Image Process.* **2020**, *29*, 4733–4746. [CrossRef]
6. Jian, L.; Yang, X.; Zhou, Z.; Zhou, K.; Liu, K. Multi-scale image fusion through rolling guidance filter. *Futur. Gener. Comput. Syst.* **2018**, *83*, 310–325. [CrossRef]
7. Hait, E.; Gilboa, G. Spectral total-variation local scale signatures for image manipulation and fusion. *IEEE Trans. Image Process.* **2018**, *28*, 880–895. [CrossRef] [PubMed]
8. Wang, J.; Peng, J.; Feng, X.; He, G.; Fan, J. Fusion method for infrared and visible images by using non-negative sparse representation. *Infrared Phys. Technol.* **2014**, *67*, 477–489. [CrossRef]
9. Zong, J.; Qiu, T. Medical image fusion based on sparse representation of classified image patches. *Biomed. Signal Process. Control.* **2017**, *34*, 195–205. [CrossRef]
10. Paramanandham, N.; Rajendiran, K. Multi sensor image fusion for surveillance applications using hybrid image fusion algorithm. *Multimed Tools.* **2018**, *77*, 12405–12436. [CrossRef]
11. Yang, Y.; Que, Y.; Huang, S.; Lin, P. Multiple Visual Features Measurement With Gradient Domain Guided Filtering for Multisensor Image Fusion. *IEEE Trans. Instrum. Meas* **2017**, *66*, 691–703. [CrossRef]
12. Zhang, B.; Lu, X.; Pei, H.; Zhao, Y. A fusion algorithm for infrared and visible images based on saliency analysis and non-subsampled Shearlet transform. *Infrared Phys. Technol.* **2015**, *73*, 286–297. [CrossRef]
13. Xu, H.; Yuan, J.; Ma, J. MURF: Mutually Reinforcing Multi-modal Image Registration and Fusion. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 12148–12166. [CrossRef] [PubMed]
14. Liu, J.; Fan, X.; Huang, Z.; Wu, G.; Liu, R.; Zhong, W.; Luo, Z. Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection. In Proceedings of the Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 5802–5811.

15. Liu, Y.; Chen, X.; Wang, Z.; Wang, Z.J.; Ward, R.K.; Wang, X. Deep learning for pixel-level image fusion: Recent advances and future prospects. *Inf. Fusion* **2018**, *42*, 158–173. [CrossRef]
16. Fu, Q.; Fu, H.; Wu, Y. Infrared and Visible Image Fusion Based on Mask and Cross-Dynamic Fusion. *Electronics* **2023**, *12*, 4342 [CrossRef]
17. Zhang, Y.; Zhai, B.; Wang, G.; Lin, J. Pedestrian Detection Method Based on Two-Stage Fusion of Visible Light Image and Thermal Infrared Image. *Electronics* **2023**, *12*, 3171 [CrossRef]
18. Li, H.; Wu, X.J.; Kittler, J. RFN-Nest: An end-to-end residual fusion network for infrared and visible images. *Inf. Fusion* **2021**, *73*, 72–86. [CrossRef]
19. Ma, J.; Tang, L.; Xu, M.; Zhang, H.; Xiao, G. STDFusionNet: An infrared and visible image fusion network based on salient target detection. *IEEE Trans. Instrum. Meas.* **2021**, *70*, 1–13. [CrossRef]
20. Zhao, Z.; Xu, S.; Zhang, J.; Liang, C.; Zhang, C.; Liu, J. Efficient and model-based infrared and visible image fusion via algorithm unrolling. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *32*, 1186–1196. [CrossRef]
21. Jian, L.; Yang, X.; Liu, Z.; Jeon, G.; Gao, M.; Chisholm, D. SEDRFuse: A symmetric encoder–decoder with residual block network for infrared and visible image fusion. *IEEE Trans. Instrum. Meas.* **2020**, *70*, 1–15. [CrossRef]
22. Ma, J.; Yu, W.; Liang, P.; Li, C.; Jiang, J. FusionGAN: A generative adversarial network for infrared and visible image fusion. *Inf. Fusion* **2019**, *48*, 11–26. [CrossRef]
23. Li, X.; Liu, W.; Li, X.; Tan, H. Physical Perception Network and an All-weather Multi-modality Benchmark for Adverse Weather Image Fusion. *arXiv* **2024**, arXiv:2402.02090.
24. Li, H.; Xu, T.; Wu, X.J.; Lu, J.; Kittler, J. LRRNet: A Novel Representation Learning Guided Fusion Network for Infrared and Visible Images. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 11040–11052. [CrossRef] [PubMed]
25. Geirhos, R.; Rubisch, P.; Michaelis, C.; Bethge, M.; Wichmann, F.A.; Brendel, W. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv* **2018**, arXiv:1811.12231.
26. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
27. Li, X.; Li, X.; Tan, H.; Li, J. SAMF: Small-area-aware multi-focus image fusion for object detection. In Proceedings of the ICASSP 2024—2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Republic of Korea, 14–19 April 2024; IEEE: Piscataway, NJ, USA, 2024; pp. 3845–3849.
28. Li, X.; Zhou, F.; Tan, H.; Zhang, W.; Zhao, C. Multimodal medical image fusion based on joint bilateral filter and local gradient energy. *Inf. Sci.* **2021**, *569*, 302–325. [CrossRef]
29. Quan, S.; Qian, W.; Guo, J.; Zhao, H. Visible and infrared image fusion based on curvelet transform. In Proceedings of the 2nd International Conference on Systems and Informatics (ICSAI 2014), Shanghai, China, 15–17 November 2014; pp. 828–832.
30. Liu, S.; Wang, J.; Lu, Y.; Li, H.; Zhao, J.; Zhu, Z. Multi-focus image fusion based on adaptive dual-channel spiking cortical model in non-subsampled shearlet domain. *IEEE Access.* **2019**, *7*, 56367–56388. [CrossRef]
31. Li, J.; Li, X.; Li, X.; Han, D.; Tan, H.; Hou, Z.; Yi, P. Multi-focus image fusion based on multiscale fuzzy quality assessment. *Digit. Signal Process.* **2024**, *153*, 104592. [CrossRef]
32. Li, X.; Zhou, F.; Tan, H.; Chen, Y.; Zuo, W. Multi-focus image fusion based on nonsubsampled contourlet transform and residual removal. *Signal Process.* **2021**, *184*, 108062. [CrossRef]
33. Li, X.; Wan, W.; Zhou, F.; Cheng, X.; Jie, Y.; Tan, H. Medical image fusion based on sparse representation and neighbor energy activity. *Biomed. Signal Process. Control* **2023**, *80*, 104353. [CrossRef]
34. Li, H.; Wu, X.J. DenseFuse: A fusion approach to infrared and visible images. *IEEE Trans. Image Process.* **2018**, *28*, 2614–2623. [CrossRef]
35. Huang, J.; Li, X.; Tan, H.; Yang, L.; Wang, G.; Yi, P. DeDNet: Infrared and visible image fusion with noise removal by decomposition-driven network. *Measurement* **2024**, *237*, 115092. [CrossRef]
36. Ma, J.; Xu, H.; Jiang, J.; Mei, X.; Zhang, X.P. DDcGAN: A dual-discriminator conditional generative adversarial network for multi-resolution image fusion. *IEEE Trans. Image Process.* **2020**, *29*, 4980–4995. [CrossRef] [PubMed]
37. Song, A.; Duan, H.; Pei, H.; Ding, L. Triple-discriminator generative adversarial network for infrared and visible image fusion. *Neurocomputing* **2022**, *483*, 183–194. [CrossRef]
38. Nichol, A.Q.; Dhariwal, P. Improved denoising diffusion probabilistic models. In Proceedings of the International Conference on Machine Learning, PMLR, Virtual, 18–24 July 2021; pp. 8162–8171.
39. Dhariwal, P.; Nichol, A. Diffusion models beat gans on image synthesis. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 8780–8794.
40. Zhao, Z.; Bai, H.; Zhu, Y.; Zhang, J.; Xu, S.; Zhang, Y.; Zhang, K.; Meng, D.; Timofte, R.; Van Gool, L. DDFM: Denoising diffusion model for multi-modality image fusion. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–6 October 2023; pp. 8082–8093.
41. Xu, Y.; Li, X.; Jie, Y.; Tan, H. Simultaneous Tri-Modal Medical Image Fusion and Super-Resolution using Conditional Diffusion Model. *arXiv* **2024**, arXiv:2404.17357.
42. Chung, H.; Kim, J.; Mccann, M.T.; Klasky, M.L.; Ye, J.C. Diffusion posterior sampling for general noisy inverse problems. *arXiv* **2022**, arXiv:2209.14687.
43. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 3–19.
44. Zhao, H.; Gallo, O.; Frosio, I.; Kautz, J. Loss functions for image restoration with neural networks. *IEEE Trans. Comput. Imaging* **2016**, *3*, 47–57. [CrossRef]

45. Tang, L.; Yuan, J.; Zhang, H.; Jiang, X.; Ma, J. PIAFusion: A progressive infrared and visible image fusion network based on illumination aware. *Inf. Fusion* **2022**, *83*, 79–92. [CrossRef]
46. Xu, H.; Ma, J.; Le, Z.; Jiang, J.; Guo, X. Fusiondn: A unified densely connected network for image fusion. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 12484–12491.
47. Toet, A. The TNO multiband image data collection. *Data Brief* **2017**, *15*, 249–251. [CrossRef]
48. Qu, G.; Zhang, D.; Yan, P. Information measure for performance of image fusion. *Electron. Lett.* **2002**, *38*, 1. [CrossRef]
49. Hossny, M.; Nahavandi, S.; Creighton, D. Comments on 'Information measure for performance of image fusion'. *Electron. Lett.* **2008**, *44*, 1066–1067.
50. Zhao, J.; Laganiere, R.; Liu, Z. Performance assessment of combinative pixel-level image fusion based on an absolute feature measurement. *Int. J. Innov. Comput. Inf. Control* **2007**, *3*, 1433–1447.
51. Liu, Z.; Forsyth, D.S.; Laganière, R. A feature-based metric for the quantitative evaluation of pixel-level image fusion. *Comput. Vis. Image Underst.* **2008**, *109*, 56–68. [CrossRef]
52. Chen, Y.; Blum, R.S. A new automated quality assessment algorithm for image fusion. *Image Vis. Comput.* **2009**, *27*, 1421–1432. [CrossRef]
53. Piella, G.; Heijmans, H. A new quality metric for image fusion. In Proceedings of the 2003 International Conference on Image Processing (Cat. No. 03CH37429), Barcelona, Spain, 14–17 September 2003; IEEE: Piscataway, NJ, USA, 2003; Volume 3, pp. III–173.
54. Xu, H.; Ma, J.; Jiang, J.; Guo, X.; Ling, H. U2Fusion: A unified unsupervised image fusion network. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *44*, 502–518. [CrossRef] [PubMed]
55. Huang, Z.; Liu, J.; Fan, X.; Liu, R.; Zhong, W.; Luo, Z. Reconet: Recurrent correction network for fast and efficient multi-modality image fusion. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; Springer: Berlin/Heidelberg, Germany, 2022; pp. 539–555.
56. Liang, P.; Jiang, J.; Liu, X.; Ma, J. Fusion from decomposition: A self-supervised decomposition approach for image fusion. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; Springer: Berlin/Heidelberg, Germany, 2022; pp. 719–735.
57. Wang, D.; Liu, J.; Fan, X.; Liu, R. Unsupervised Misaligned Infrared and Visible Image Fusion via Cross-Modality Image Generation and Registration. In Proceedings of the International Joint Conference on Artificial Intelligence, Vienna, Austria, 23–29 July 2022.
58. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 801–818.
59. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.
60. Cao, Z.; Simon, T.; Wei, S.E.; Sheikh, Y. Realtime multi-person 2d pose estimation using part affinity fields. In Proceedings of the Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7291–7299.
61. Ranftl, R.; Lasinger, K.; Hafner, D.; Schindler, K.; Koltun, V. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *44*, 1623–1637. [CrossRef]
62. Tang, L.; Yuan, J.; Ma, J. Image fusion in the loop of high-level vision tasks: A semantic-aware real-time infrared and visible image fusion network. *Inf. Fusion* **2022**, *82*, 28–42. [CrossRef]