remote sensing

MDPI

Article

Infrared–Visible Image Fusion through Feature-Based Decomposition and Domain Normalization

Weiyi Chen, Lingjuan Miao, Yuhao Wang *, Zhiqiang Zhou and Yajun Qiao

School of Automation, Beijing Institute of Technology, Beijing 100081, China; 3220195102@bit.edu.cn (W.C.); miaolingjuan@bit.edu.cn (L.M.); zhzhzhou@bit.edu.cn (Z.Z.); yajun@bit.edu.cn (Y.Q.)
* Correspondence: wayhao@bit.edu.cn

Abstract: Infrared–visible image fusion is valuable across various applications due to the complementary information that it provides. However, the current fusion methods face challenges in achieving high-quality fused images. This paper identifies a limitation in the existing fusion framework that affects the fusion quality: modal differences between infrared and visible images are often overlooked, resulting in the poor fusion of the two modalities. This limitation implies that features from different sources may not be consistently fused, which can impact the quality of the fusion results. Therefore, we propose a framework that utilizes feature-based decomposition and domain normalization. This decomposition method separates infrared and visible images into common and unique regions. To reduce modal differences while retaining unique information from the source images, we apply domain normalization to the common regions within the unified feature space. This space can transform infrared features into a pseudo-visible domain, ensuring that all features are fused within the same domain and minimizing the impact of modal differences during the fusion process. Noise in the source images adversely affects the fused images, compromising the overall fusion performance. Thus, we propose the non-local Gaussian filter. This filter can learn the shape and parameters of its filtering kernel based on the image features, effectively removing noise while preserving details. Additionally, we propose a novel dense attention in the feature extraction module, enabling the network to understand and leverage inter-layer information. Our experiments demonstrate a marked improvement in fusion quality with our proposed method.

Keywords: infrared and visible image fusion; unified feature space; dynamic instance normalization; non-local Gaussian filter; dense attention

Citation: Chen, W.; Miao, L.; Wang, Y.; Zhou, Z.; Qiao, Y. Infrared–Visible Image Fusion through Feature-Based Decomposition and Domain Normalization. *Remote Sens.* **2024**, *16*, 969. https://doi.org/10.3390/rs16060969

Academic Editors: Stefano Mattoccia, Piotr Kaniewski, Fabio Tosi

Received: 26 January 2024
Revised: 29 February 2024
Accepted: 5 March 2024
Published: 10 March 2024

1. Introduction

Recently, infrared and visible image fusion (IVIF) has gained considerable attention, owing to its extensive applications in various fields [1–3]. Single-modal images typically contain limited scene information and cannot fully reflect the true environment. Therefore, fusing information from different imaging sensors helps to enhance the informational richness of the images. Infrared and visible images have strong complementarity, i.e., infrared cameras capture thermal radiation but may not provide detailed information, while visible images are not sufficient in detecting hidden objects. Due to the complementarity and advantages of these two modalities, IVIF is widely applied in fields such as nighttime driving, military operations, and object detection.

In recent years, researchers have proposed various methods for IVIF, which can be categorized into traditional and deep learning-based methods. Traditional methods aim to design optimal representations across modalities and formulate fusion weights. These methods include multi-scale decomposition (MSD)-based methods [3–6], other transformation-based methods [7–9], and saliency-based methods [10–12]. The advancements of deep learning have significantly accelerated the evolution of IVIF. Researchers have proposed sophisticated modules or structures [13–18] for the integration of features

from both infrared and visible images. Autoencoders [19–24] have also been introduced into the IVIF process due to their powerful feature extraction capabilities. Additionally, generative adversarial networks (GANs) [25–28] have been employed to enhance the fusion performance. However, existing research often neglects the differences between infrared and visible images, as well as the noise present in source images.

There are still some challenges that need to be tackled. Firstly, there is a significant difference between infrared and visible images. This difference leads to the inconsistent fusion of features when they come from these different sources. As a result, the quality of the fusion results is often affected. The differences between the infrared and visible modalities can be attributed to variations in wavelength, sources of radiation, and acquisition sensors. These modal differences lead to variations in images, such as texture, luminance, contrast, etc., subsequently affecting the fusion quality. Although decomposition representation-based methods can reduce the impact of modal differences, they often require complex decomposition and fusion rules. Secondly, low luminance may result in noisy source images. These images often impact the performance of image fusion, leading to suboptimal results. Thirdly, many methods neglect essential information from the middle layers, which are crucial in the fusion process. While dense connections [22] have been introduced into the fusion network, these connections lead to higher computational costs.

To address these challenges, we propose a novel method (UNIFusion) for IVIF, which includes cosine similarity-based image decomposition, a unified feature space, and dense attention for feature extraction. To obtain high-quality fused images, our method reduces the differences between infrared and visible features through the unified feature space, while also preserving their unique information. We first decompose the infrared and visible images into common and unique regions, respectively. Then, the features extracted from common regions are fed into the unified feature space to obtain fused features without modal differences. Specifically, we first obtain unique and common regions based on the cosine similarity between the embedded features of infrared–visible images. The unique regions contain private information that should be preserved in the fusion process, while the common regions in both infrared and visible images contain similar content. Secondly, to obtain fusion results with more information, we design a unified feature space to eliminate the differences between common features. In the space, infrared features are transformed to the pseudo-visible domain, thereby eliminating the differences between modalities. Thirdly, we propose a dense attention to enhance the feature extraction capabilities of the encoder, particularly focusing on improving the model's ability to capture important information from the input data. By applying an attention weight across all layers of the encoder, this method ensures that the model focuses on important features, which helps the model to perform fusion tasks better. Moreover, we propose the non-local Gaussian filter to enhance the fusion results. This filter can learn the shape and kernel parameters, enabling it to remove noise while retaining details.

As demonstrated in Figure 1, our method outperforms current fusion algorithms like FusionGAN [26], PMGI [29], and U2Fusion [15]. It is apparent that we can obtain better results through the unified feature space. Even the current state-of-the-art methods for IVIF cannot obtain satisfactory fused images. For example, FusionGAN generates blurred fused images, while PMGI and U2Fusion lead to fusion artifacts. Conversely, our method can improve the fusion performance by fusing multi-modal features in a consistent space.

The main contributions of this paper are summarized as follows.

- To eliminate the modal difference, we propose a domain normalization method based on the unified feature space, which enables the transformation of infrared features to the pseudo-visible domain, ensuring that all features are fused within the same domain and minimizing the impact of modal differences during the fusion process.
- We propose a feature-based image decomposition method that separates images into common and unique regions based on the cosine similarity. This approach eliminates the need to manually craft intricate decomposition algorithms, offering an adaptive solution that simplifies the process.

- We design a dense attention to allow the encoder to focus on more relevant features while ignoring redundant or irrelevant ones. Moreover, the Non-local Gaussian filter is incorporated into the fusion network to reduce the impact of noisy images on the fusion results.

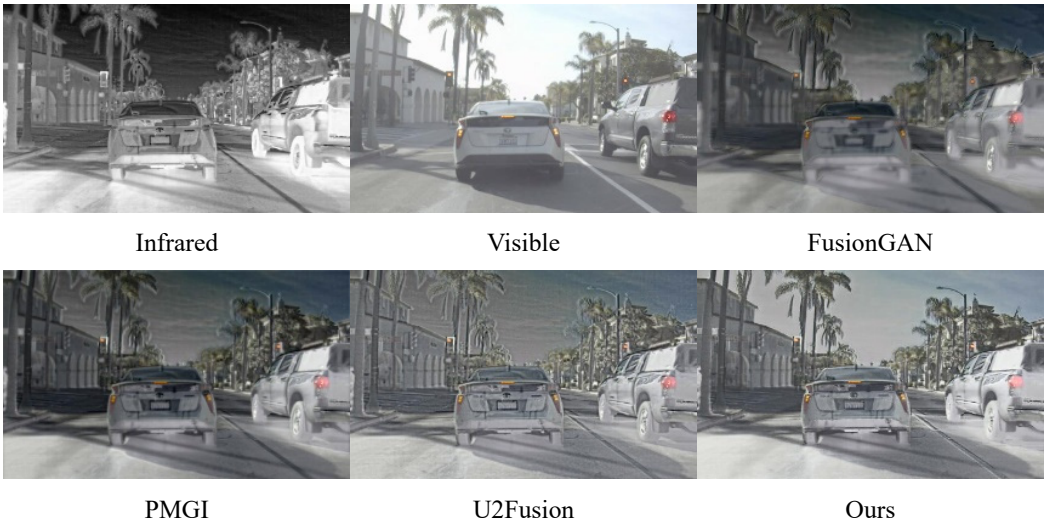


Figure 1. A comparison of the fused images generated by our UNIFusion and other state-of-the-art fusion methods.

2. Related Works

In this section, we review various IVIF methods, categorizing them into traditional, AE-based, and GAN-based approaches. Additionally, related works on image-to-image translation are briefly presented to obtain a deeper understanding of the proposed models.

2.1. Traditional-Based Methods

In the study of traditional methods for IVIF, various techniques have been proposed, which include multi-scale decomposition, saliency detection, etc. Multi-scale decomposition methods [4,5,7] decompose and reconstruct the features of infrared and visible image at various levels to better fuse details, structures, etc. These approaches align the process of scale information with the human visual system. Saliency detection methods [10–12] can enhance the fusion performance on important targets by assigning higher weights to salient regions or objects. Sparse representation techniques [30] use dictionaries learned from a large set of images to encode and preserve essential information from the source images during the fusion process. These traditional approaches provide a foundation for IVIF, which can retain the image details and improve the visual effect.

2.2. CNN-Based Methods

The introduction of convolutional neural networks (CNN) has revolutionized the field of infrared and visible image fusion (IVIF). Specifically, Liu et al. [13] were pioneers in this area, applying a Siamese CNN structure to effectively generate a weight map from the source images. Over time, the architectures of CNNs in IVIF have continuously evolved. Early CNN architectures included single-branch and dual-branch configurations. For instance, Li et al. [14] incorporated residual connections to enhance the fusion capabilities. Xu et al. [31] developed a multi-scale unsupervised network based on joint attention mechanisms, significantly improving the detail preservation in the fused images. Moreover, the research by Ma et al. [17] presents a fusion technique anchored in the Transformer framework, equipped with an attention module to integrate global information. Alongside this, the impact of the lighting conditions in fusion tasks is noteworthy. PIAFusion [18] tries to improve the fusion performance based on an illumination-aware module, but its model is not successful in handling complex lighting scenarios.

2.3. Autoencoder-Based Methods

Autoencoders are effective in infrared–visible image fusion as they are adept at encoding and decoding image features. This capability is essential to effectively fuse infrared and visible information. Li et al. introduced the DenseFuse method [22], which marked a significant advancement in IVIF tasks. This approach efficiently fuses visible and infrared images, paving the way for further research and development in this area. After the introduction of DenseFuse, AE-based methods for IVIF received significant development, which can be categorized as single-branch-based methods [19,20] and dual-branch-based methods [21–24]. The advancements of autoencoders have played a crucial role in improving both the efficiency and performance of the image fusion process. Additionally, the introduction of innovative modules has significantly enhanced the quality of the fused images. These modules include residual connections, channel attention, and self-attention.

Autoencoder-based methods can significantly enhance the fusion performance due to their strong capacity for feature extraction and reconstruction. This ability allows for the more comprehensive fusion of source image information, leading to superior fusion results.

2.4. GAN-Based Methods

In the IVIF task, generative adversarial networks (GANs) have been employed to generate fused images that contain rich information from the source images. Liao et al. [25] leveraged the powerful generative capabilities of GANs to produce realistic and information-rich fused images, demonstrating the advantages of GAN-based methods in infrared and visible image fusion. Furthermore, Xu et al. [27] developed a conditional GAN featuring dual discriminators, each trained on infrared and visible images. This approach effectively balances features from both types of images, thereby enhancing the fusion performance.

The architectural innovation in GAN-based methods is noteworthy. Researchers have experimented with multiple discriminators to improve the fusion performance. For example, Song et al. [28] introduced a novel GAN-based method with a triple discriminator for IVIF, which produces detailed fused images. In addition, researchers are focusing on the design of loss functions and architectures. For example, Li et al. [32] and Yuan et al. [33] used the Wasserstein distance and group convolution in GAN architectures, respectively, which led to better fusion results.

2.5. Image-to-Image Translation Methods

The objective of image-to-image (I2I) translation is to convert an image from a source domain to a target domain, ensuring that the essential characteristics of the input image are retained. Various generative adversarial network (GAN)-based frameworks have been proposed to align the output image distribution with that of the target domain. For instance, in 2016, Isola et al. introduced Pix2Pix [34], a conditional GAN model capable of translating images across domains using paired training data. Subsequently, Pix2PixHD [35] was developed to address high-resolution image translation. However, a significant challenge with these paired I2I translation methods is their dependence on paired datasets, which can be challenging and expensive to acquire, and sometimes even unattainable. Consequently, various approaches [36–39] have been explored to overcome the limitation for paired datasets. For instance, Bousmalis et al. [40] proposed an I2I translation method based on unsupervised training that applies domain adaptation in the pixel space. In our approach, we design a unified feature space to transform infrared features into the pseudo-visible domain. This ensures that all features exist within the same domain, eliminating the impact of modality differences on the fusion process.

3. Methods

3.1. Overview

Our proposed UNIFusion is an autoencoder structure, which consists of image decomposition, feature extraction, fusion, and reconstruction modules. The feature extraction module is a three-branch network based on dense attention, consisting of encoders E^{ir},

E^{vi}, and E^{u}, which are used to extract unique and unified features. The fusion and reconstruction module is devised to fuse features and generate fusion results, while employing a non-local Gaussian filter to reduce the adverse impact of noise on the fusion quality. The complete architecture is depicted in Figure 2, providing a detailed overview. Specifically, we decompose infrared–visible images into common regions (C^{vi} and C^{ir}) and unique regions (P^{vi} and P^{ir}). The dense attention is leveraged to effectively extract features from the common and unique regions. To eliminate modal differences, we propose the unified feature space to transform infrared features into the pseudo-visible domain. As noisy source images may degrade the fusion quality, we design a non-local Gaussian filter to minimize the impact of noise on the fusion results while maintaining the image details.

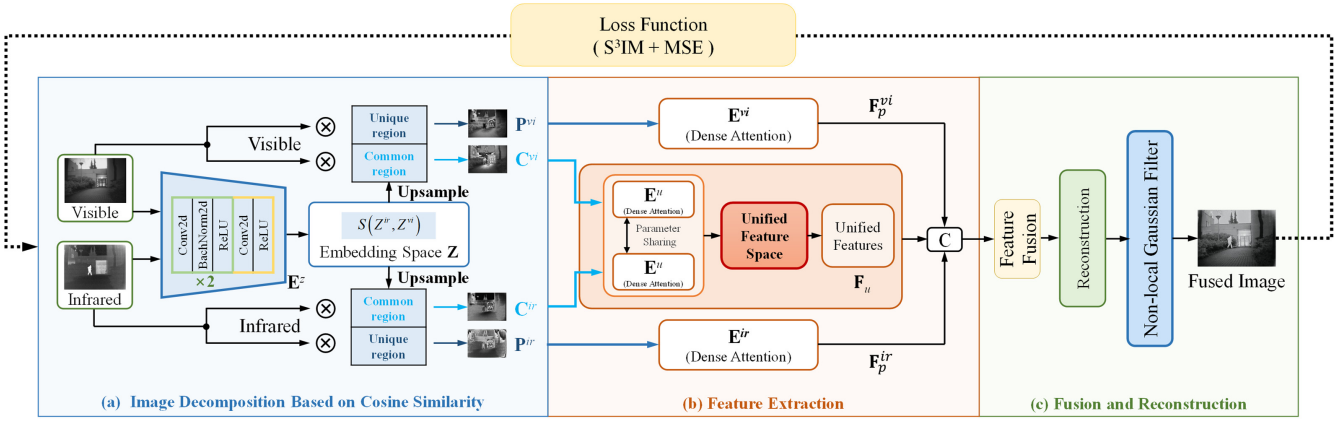


Figure 2. The overall framework of the proposed method. The method consists of (**a**) image decomposition, (**b**) feature extraction module, and (**c**) fusion and reconstruction module. (**a**) decomposes source images into common and unique regions, respectively. (**b**) is a three-branch network, consisting of encoders E^{ir}, E^{vi}, and E^{u}. The encoders based on dense attention are used to extract unique and unified features. (**c**) is devised to fuse features and generate fusion results, while employing a non-local Gaussian filter to reduce the adverse impact of noise on the fusion quality.

During the training phase, we use the S^{3}SIM and MSE loss functions to evaluate the similarity between the fused image and the original inputs. This helps to refine the network parameters.

3.2. Image Decomposition Based on Cosine Similarity

To obtain the common regions (C^{vi} and C^{ir}) and unique regions (P^{vi} and P^{ir}) of the source images, we embed the infrared and visible images into a shared parameter space Z to obtain consistent feature representations. By comparing the similarity of these features using cosine similarity, we can capture the directional similarity of the image features without being affected by the absolute luminance. The size of the feature map is $h \times w$ and the dimension is d, which leads to the definitions (1) and (2) for feature representation. Elements within these feature maps are denoted by the lowercase z, which are vectors in the d-dimensional space. The superscript of z indicates the modality (with vi for visible light and ir for infrared), and its subscript denotes the position of the element. The definitions are shown below:

$$Z^{vi} = \begin{bmatrix} z^{vi}_{1,1} & \cdots & z^{vi}_{1,w} \\ \vdots & \ddots & \vdots \\ z^{vi}_{h,1} & \cdots & z^{vi}_{h,w} \end{bmatrix}_{h \times w}, Z^{vi} \in \mathbb{R}^{d \times h \times w}, \quad (1)$$

$$Z^{ir} = \begin{bmatrix} z_{1,1}^{ir} & \cdots & z_{1,w}^{ir} \\ \vdots & \ddots & \vdots \\ z_{h,1}^{ir} & \cdots & z_{h,w}^{ir} \end{bmatrix}_{h \times w}, Z^{ir} \in \mathbb{R}^{d \times h \times w}, \tag{2}$$

where $z_{i,j}^{vi}$ is the element in the i-th row and j-th column of the visible feature matrix. $z_{i,j}^{ir}$ is the element in the i-th row and j-th column of the infrared feature matrix.

The cosine similarity (denoted as cs in the Equation (3)) is used to decompose infrared and visible images into common and unique regions. This is because the cosine similarity captures the structural similarity between infrared and visible images, which is more important for image fusion than absolute luminance. Two types of masks for source image decomposition are derived by computing the cosine similarity (denoted as c), namely M_c (common mask) and M_p (unique mask), as detailed in Equations (4) and (5):

$$S = cs(Z^{vi}, Z^{ir}) = \begin{bmatrix} cs(z_{1,1}^{vi}, z_{1,1}^{ir}) & \cdots & cs(z_{1,w}^{vi}, z_{1,w}^{ir}) \\ \vdots & \ddots & \vdots \\ cs(z_{h,1}^{vi}, z_{h,1}^{ir}) & \cdots & cs(z_{h,w}^{vi}, z_{h,w}^{ir}) \end{bmatrix}_{h \times w}, \tag{3}$$

$$M_c = \frac{1+S}{2}, \tag{4}$$

$$M_p = \frac{1-S}{2}, \tag{5}$$

where S is the similarity matrix of size $h \times w$, representing the cosine similarity between visible and infrared features. cs is the cosine similarity function. M_c represents the common mask, and $\frac{1+S}{2}$ normalizes the similarity scores to a range [0, 1], where 1 indicates the maximum similarity. M_p is the unique mask, and the transformation $\frac{1-S}{2}$ also normalizes the scores, with 1 indicating the maximum difference.

Next, we upsample the common mask and unique mask to align with the source image size. Element-wise multiplication is performed between the two masks (M_c and M_p) and infrared–visible images (I^{ir} and I^{vi}) to yield four decomposed outcomes (C^{ir}, P^{ir}, C^{vi}, and P^{vi}). The decomposed results are defined as followed, representing infrared–visible common regions and unique regions, respectively:

$$C^{ir} = I^{ir} \times Upsample(M_c), \tag{6}$$

$$P^{ir} = I^{ir} \times Upsample(M_p). \tag{7}$$

$$C^{vi} = I^{vi} \times Upsample(M_c), \tag{8}$$

$$P^{vi} = I^{vi} \times Upsample(M_p), \tag{9}$$

The employment of cosine similarity enables more precise decomposition, ensuring that the common regions and unique regions between the infrared and visible images are captured.

3.3. Dense Attention for Feature Extraction

Although the current fusion methods [15,22] try to utilize skip connection structures to obtain rich features, the differences between multi-scale features are not sufficiently taken into account. Specifically, low-level features capture basic input characteristics, while high-level features are more abstract, representing complex concepts and structures. Dense connections and residual connections concatenate multi-scale features directly, which can

make it challenging for neural networks to differentiate important features, consequently limiting the fusion performance.

To address this limitation, we propose a dense attention-based feature extraction module to obtain multi-scale features, as shown in Figure 3. By inserting attention into every dense connection, the model can learn the significant features and relationships between different layers. Furthermore, as the network depth increases, this attention mechanism helps the model to learn long-range dependencies, improving its generalization and robustness.

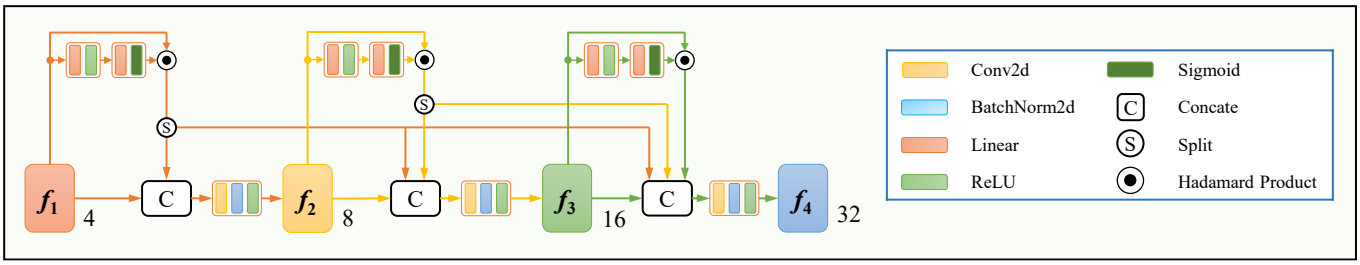


Figure 3. The structure of the feature extraction module based on dense attention.

3.4. Unified Feature Space Based on Dynamic Instance Normalization

We construct the unified feature space to eliminate the difference between infrared and visible features at the multi-scale feature level. The core components of the space include a scale-aware module, shifted patch embedding, and dynamic instance normalization (DIN), as shown in Figure 4. Specifically, the scale-aware module is trained to determine the size and shape of a patch. With the n pairs of scale and size parameters output by this module, shifted patch embedding can divide the feature map into n groups. For each group, it splits the feature map into patches according to the corresponding scale and size. DIN transforms infrared features into a pseudo-visible domain for each patch, which eliminates the differences between infrared and visible images. Subsequently, the learned confidence merges the features from the two modalities to produce the output result.

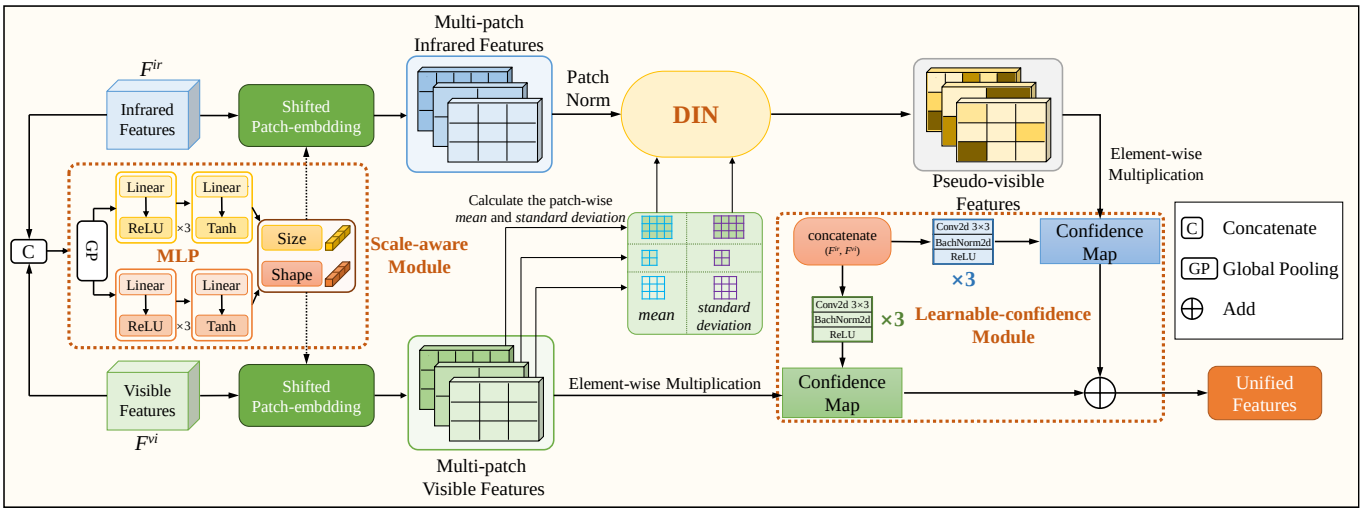


Figure 4. An illustration of the unified feature space based on dynamic instance normalization (DIN).

More specifically, the unified feature space enables the domain transformation from infrared to pseudo-visible, while also being adaptable to multi-scale targets. Dynamic instance normalization (DIN) is the core of the unified feature space, capable of transforming features from infrared features to pseudo-visible, thereby eliminating the difference between the two modalities. Moreover, we employ global pooling to concatenate features in order to enable a multilayer perceptron (MLP) to generate n pairs of size and shape

parameters. The multi-patch embedding module divides the infrared and visible features into n groups along the channel dimension. Within each group, the features are segmented into patches of the same scale, determined by a set of size and shape parameters. Then, DIN transforms the infrared features to the pseudo-visible domain for each patch after shifted patch embedding. For the fusion of infrared and pseudo-visible features, we design a learnable confidence module to learn fusion weights; this method can adjust the fusion weight depending on the image content, compared with the fusion rules of addition, concatenation, and so on.

Although adaptive instance normalization (AdaIN) [41,42] plays a crucial role in image translation tasks, the core idea of AdaIN is to adjust the feature distribution of a content image to match the feature distribution of a target style image, thereby achieving style transfer. This process involves normalizing the features of the content image and then adjusting these normalized features with the statistical data (mean and variance) of the target style image. Through this method, the content image adopts the style characteristics of the style image while retaining its content structure. However, this method is not very precise due to the transformation of the domain at the level of global features. This limitation prevents independent domain transformations for each patch, restricting the effectiveness of domain transformation. To address this, we introduce dynamic instance normalization (DIN), which astutely segments the feature map into distinct subregions, as shown in Figure 5. This segmentation allows for independent domain transformations on each patch, enhancing the adaptability of the process. The DIN function is mathematically represented as

$$\mathrm{DIN(X,\,Y)} = [\mathrm{AdaIN}(x_1, y_1), \mathrm{AdaIN}(x_2, y_2), \ldots, \mathrm{AdaIN}(x_n, y_n)], \tag{10}$$

$$\mathrm{AdaIN}(x, y) = \sigma(y) \cdot \left(\frac{x - \mu(x)}{\sigma(x)} \right) + \mu(y), \tag{11}$$

where both X and Y denote global features, X represents the content input, and Y is the modal attribute input. Both X and Y are segmented into n patches, resulting in patch-wise pairs denoted as (x_i, y_i) for $i = 1, 2, \ldots, n$, where each pair corresponds to matching patches from X and Y. The terms $\mu(x)$ and $\mu(y)$ denote the means of x and y, respectively, while $\sigma(x)$ and $\sigma(y)$ denote their standard deviations.

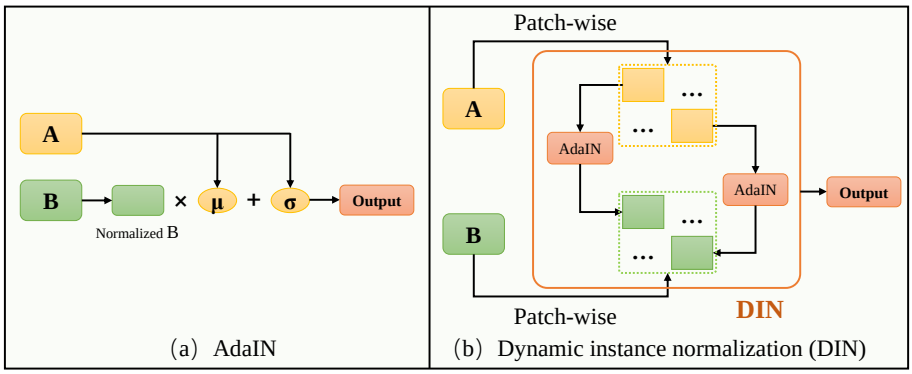


Figure 5. Different domain transformation methods. (**a**) AdaIN performs domain transformation by adjusting the global feature distribution of the content input (denoted as B), making it match the global feature distribution of the modal attribute input (denoted as A). (**b**) DIN, extended from AdaIN, adjusts the feature distribution at the patch-wise level, enabling more detailed domain adaptation.

In particular, we feed the concatenated infrared and visible features into a scale-aware module to obtain the scales and ratios. The shifted patch embedding module separately splits infrared and visible features into n groups and partitions each group of features into patches based on the scale and ratio. Infrared and visible patches can be represented as $\mathrm{X} = [x_1, x_2, \cdots, x_n]$ and $\mathrm{Y} = [y_1, y_2, \cdots, y_n]$, respectively. Applying DIN to each infrared–

visible patch pair, as shown in Equation (10), we transform the infrared features into the pseudo-visible domain at the patch level. Then, we multiply them element-wise with a neural network-derived confidence metric to form the final fusion features. We obtain the final unified features by fusing pseudo-visible and visible features based on the learnable-confidence module.

3.5. Hierarchical Decoder for Fusion and Reconstruction

The hierarchical decoder does not only allow us to fuse infrared–visible features and generate fused images, but is also robust to the noise contained in source images and enhances the clarity of the fusion result. In this paper, we propose a multi-stage decoder to achieve more refined fusion, which can be divided into fusion, reconstruction, and enhancement stages.

The specific design of the hierarchical decoder is shown in Figure 6. We deploy two convolutional layers to fuse unified and unique features, receptively, in order to retain more infrared–infrared information. Then, in the reconstruction, we propose a novel module to learn the fusion strategy and obtain refined features. As every scale feature is vital to the fusion task, we not only insert a nest connection to learn the fusion strategy, but also propose a direct connection to output multi-scale features. Specifically, in the proposed architecture, features are reconstructed to match the size of the input image through a series of convolutional or transposed convolutional layers. These reconstructed features are then propagated to subsequent layers. In the final enhancement stage, we employ two distinct sets of convolutional layers to obtain a guidance feature used to obtain the filter parameters and preliminary fused images. Subsequently, we utilize a cascade of three convolutional layers to derive two-dimensional positional offsets and non-local Gaussian kernels.

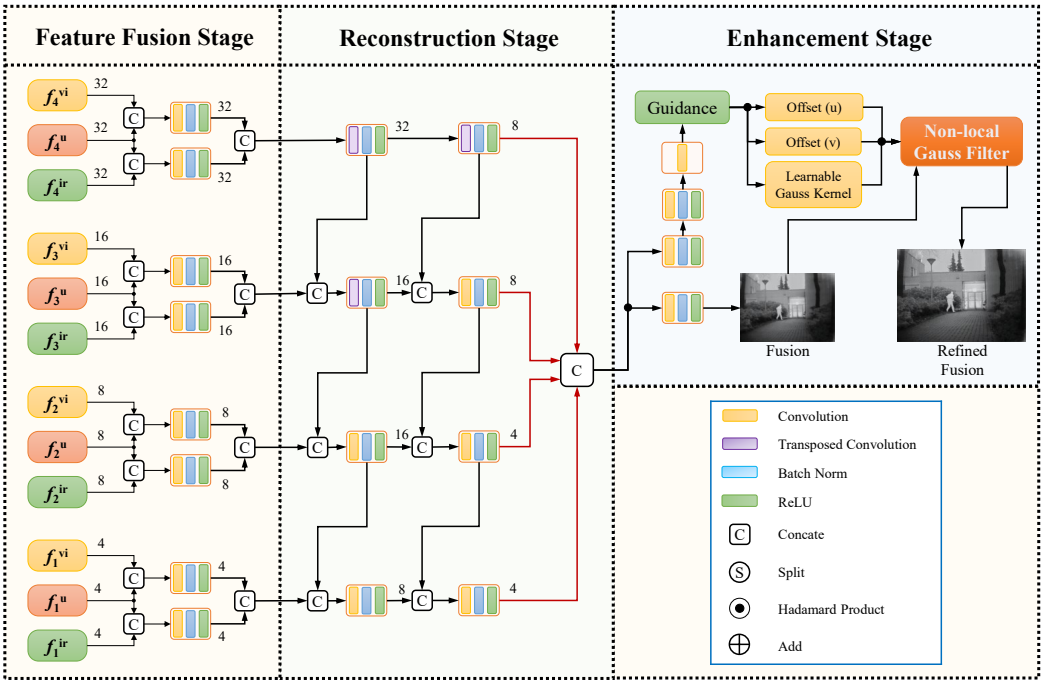


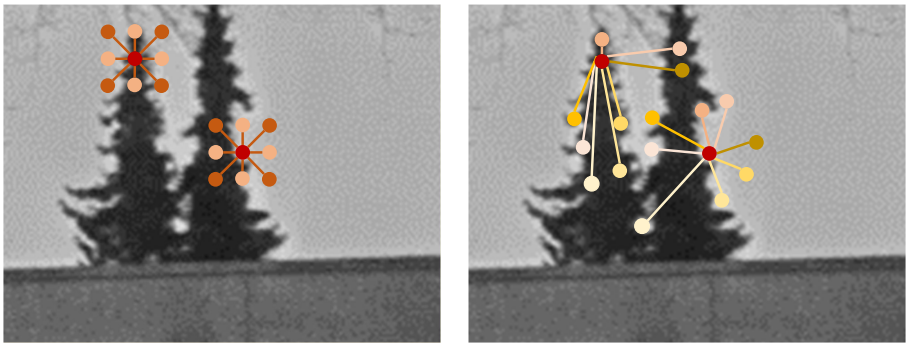
Figure 6. The structure of the hierarchical decoder.

Regarding the non-local Gaussian filter (shown in Figure 7), used for image enhancement, the process involves refining a preliminary fusion result, denoted as f. Here, $f_{i,j}$ represents the value at position (i, j) after an initial fusion step. The refined fusion outcome, $\hat{f}$, is achieved through an advanced filtering technique, mathematically formulated as

$$S_{i,j} = \sum_{n=1}^{N} w_{i,j}^{n}, \tag{12}$$

$$\hat{f}_{i,j} = \sum_{n=1}^{N} \frac{w_{i,j}^{n}}{S_{i,j}} \cdot f_{i+\Delta i_n, j+\Delta j_n}, \tag{13}$$

where $f_{i,j}$ represents the value at position (i, j), and N is the total number of neighbors, with a default value of 9. The term $w_{i,j}^{n}$ denotes learnable Gaussian kernels for the n-th neighbor of the pixel at (i, j). $S_{i,j}$ is the sum of weights for all neighbors, used to normalize the weights such that the sum of weights within the neighborhood equals 1. The terms Δi_n and Δj_n represent the positional offset values for the n-th neighbor, indicating the deviations in the row (vertical) and column (horizontal) directions, respectively, relative to the central pixel (i, j).



(a) Gaussian filter (b) Non-local Gaussian filter

Figure 7. An illustration of the non-local Gaussian filter, which employs a dynamic kernel to enhance the image fusion.

The non-local Gaussian filter enables the adaptive refinement of the fusion process. By dynamically adjusting the offsets and weights based on the local structures of the initial fusion result, the network can achieve a more optimized and contextually aware fusion outcome.

3.6. Loss Function

In this paper, we introduce two types of loss functions to simultaneously preserve crucial information from the source images and enhance the saliency of the fused image. Our loss functions incorporate two key components: the mean squared error (MSE) loss $\mathcal{L}mse$ and the proposed saliency structural similarity index (S^3IM) loss $\mathcal{L}_{s^3im}$. The MSE loss is used to constrain the similarity between the fusion results and the infrared–visible images. This loss focuses on maintaining fidelity to the source images by minimizing pixel-wise differences. Our proposed S^3IM loss aims to emphasize the saliency in the fused image. The total loss is calculated as follows:

$$\mathcal{L}(\theta, D) = \mathcal{L}_{mse}(\theta, D) + \lambda \mathcal{L}_{s^3im}(\theta, D), \tag{14}$$

where θ represents the parameters of the neural network, D represents the training data, and λ is the hyperparameter that balances the two losses.

Due to its efficiency and stability, the mean squared error loss $\mathcal{L}_{mse}$ can provide high accuracy and reliability in many cases. Therefore, we use it to constrain the similarity between the source images I_1, I_2, and the fused image I_f. Its definition is as follows:

$$\mathrm{MSE}(A, B) = \frac{1}{N} \sum_{i=1}^{N} (A_i - B_i)^2, \tag{15}$$

$$\mathcal{L}_{mse}(\theta, D) = \mu_1 \mathrm{MSE}(I_f, I_1) + \mu_2 \mathrm{MSE}(I_f, I_2), \tag{16}$$

where μ_1 and μ_2 are hyperparameters that balance the weights of the two MSE terms in the loss function. This allows the model to adjust the reliance on the visible image and the infrared image according to the needs of the specific task.

The structural similarity index measure (SSIM) [43] is a widely used image quality assessment metric that aims to quantify the perceptual similarity between two images. However, in infrared images, there are pixels with zero or very low intensity values, which means that the corresponding regions do not have objects with thermal radiation. In the fusion process, they should be assigned lower weights. To address this issue, we propose the saliency SSIM (S^3IM). Specifically, S^3IM can adaptively determine the loss weights based on the pixel intensity. We divide the normalized pixel values into three major regions: the low-saliency area, the linear area, and the high-saliency area, as shown in Figure 8.

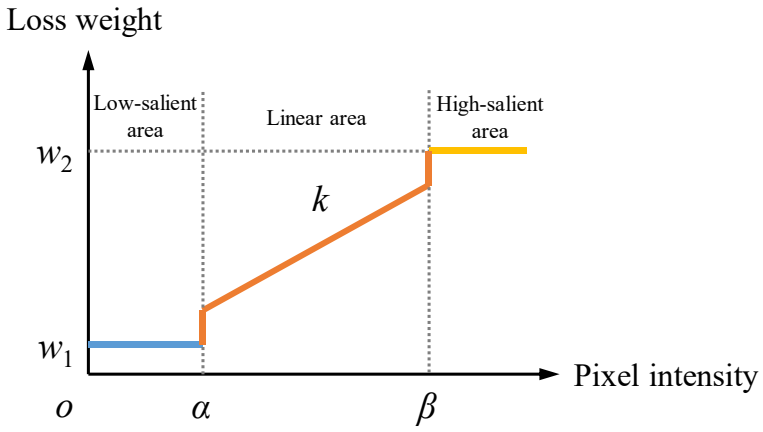


Figure 8. The schematic diagram of the s^3im weight.

The low-saliency area contains pixels with lower intensity values, which typically do not contain target information. When calculating the loss, they should be assigned a very low weight. The high-saliency region contains pixels with high intensity values, indicating objects with high thermal radiation, and they should have higher saliency in the fused image. For the remaining pixels, we adopt a linear transformation strategy to determine their loss weights, corresponding to the linear region in Figure 8. In summary, the calculation method is shown as follows:

$$h(x) = \begin{cases} w_1, x < \alpha \\ kx + b, \alpha \leq x \leq \beta \\ w_2, x > \beta \end{cases}, \tag{17}$$

$$\mathcal{L}_{s^3im}(\theta, D) = \varphi\Big[1 - \mathrm{SSIM}(I_f, I_1)\Big] + h(I_2) \cdot \Big[1 - \mathrm{SSIM}(I_f, I_2)\Big], \tag{18}$$

where φ is a hyperparameter used to adjust the weights of the infrared and visible images during the fusion process.

4. Experimental Results

In this section, we describe the experimental setup and the details of the network training. Following this, we perform a comparative analysis of the current fusion methods and carry out generalization experiments to highlight the benefits of our approach. Additionally, we conduct ablation studies to validate the effectiveness of our proposed methods.

4.1. Experimental Settings

We conduct experiments using four publicly available datasets. The M3FD dataset [44] is used for model training, while the TNO [45], RoadScene [15], and VTUAV [46] datasets are used to evaluate the performance of our method. The M3FD dataset contains 300 pairs of infrared and visible images for IVIF, including targets such as people, cars, buses,

motorcycles, trucks, etc. These images were collected under various illuminance conditions and scenarios. The TNO dataset contains multispectral imagery from various military scenarios. The RoadScene dataset includes 221 image pairs featuring roads, vehicles, pedestrians, etc. The VTUAV dataset is used for remote sensing analysis and contains complex backgrounds and moving objects. We selected 20 pairs of infrared–visible images from both the TNO and RoadScene datasets, as well as 10 pairs from the VTUAV dataset, for the evaluation of our approach.

Our UNIFusion is compared with nine current state-of-the-art fusion methods, including a biological vision-based method, i.e., PFF [47]; an autoencoder-based method, i.e., MFEIF [48]; two generative adversarial network -based methods, i.e., FusionGAN [26] and UMF [49]; two convolutional neural network-based methods, i.e., U2Fusion [15], PMGI [29], and RFN [50]; a transformer-based method, i.e., swinfusion [17]; and a high-level task supervision-based method, i.e., PIAFusion [18].

To quantitatively evaluate the fusion performance, we utilize five key metrics: the average gradient (AG) [51], standard deviation (SD) [26], correlation coefficient (CC) [52], spatial frequency [53], and multi-scale structural similarity index (MS-SSIM) [54]. The AG measures the texture richness in the image, while the SD highlights the contrast within the fused image. The SF is indicative of the detail richness and image definition. The CC evaluates the linear relationship between the fusion results and infrared–visible images. MS-SSIM is employed to calculate the structural similarity between images. Generally, higher values in AG, SD, SF, MS-SSIM, and CC denote superior fusion performance.

4.2. Implementation Details

We trained our fusion model using the M3FD fusion dataset, which contains 300 infrared–visible pairs. During training, we randomly cropped the infrared–visible image pairs into multiple 256 $\times$ 256 patches, applied random affine transformations to enhance the model performance, and normalized all images to the [0, 1] range before inputting them into the fusion model. For training, we utilized the Adam optimizer with a batch size of 16. The initial learning rate was set to 5×10^{-4} and was halved every two epochs starting from epoch 30, continuing this reduction until the final epoch at 60. Additionally, we set the parameters of Equations (13)–(16) as follows: $\lambda = 1$, $\mu_1 = 1$, $\mu_2 = 1$, $\alpha = 0.2$, $\beta = 0.7$, $k = 1$, $b = 0$, $w_1 = 0.2$, $w_2 = 2$, $\varphi = 1$. The entire network was trained using the PyTorch 1.8.2 framework on an NVIDIA GeForce GTX 3080 GPU and a 3.69 GHz Intel Core i5-12600KF CPU.

4.3. Fusion Performance Analysis

In this section, we conduct a comprehensive qualitative and quantitative analysis to illustrate the advantages of our UNIFusion, comparing our method with nine state-of-the-art (SOTA) fusion approaches. In addition, we test the performance of our UNIFusion across various illumination scenarios within the VTUAV dataset.

4.3.1. Qualitative Results

The visualized comparisons of our UNIFusion with the nine SOTA methods are provided in Figures 9–11. Figures 9 and 10 present the fusion results of the different methods on the TNO and RoadScene datasets, respectively, while Figure 11 shows the color fusion results. Moreover, we evaluate our model's performance with remote sensing data collected under normal and low-light conditions, as shown in Figure 11. In our approach, we effectively transform infrared features into the pseudo-visible domain, resulting in fused images that maintain superior visual perception. This transformation process enhances the fusion of infrared and visible information, yielding more natural and clearer fusion results. Notably, our image decomposition method plays a crucial role in preserving unique information from multiple modalities, thereby highlighting salient objects in the fused images.

In Figure 9, it can be seen that FusionGAN, PMGI, RFN, U2Fusion, and UMF generate fusion results with less information and lower brightness (see the red boxes), which contain more infrared information and do not fully fuse visible image. The objects in MFEIF and PIAFusion are not salient and therefore not easily observed (see the orange boxes in Figure 9). SwinFusion suffers from overexposure and oversmoothing, resulting in some details not being clear enough (see the orange boxes in Figure 9). Although PFF can fuse more details, the results of this method contain noise (see the yellow boxes in Figure 9). On the contrary, our fused images can fuse more information through the unified feature space, which leads to rich details and structures (see the red boxes in Figure 9). Our UNIFusion can also obtain better fusion performance on small objects (see the orange boxes in Figure 9). Moreover, the results generated from our method are clear and contain less noise due to the non-local Gaussian filter (see the orange boxes in Figure 9).

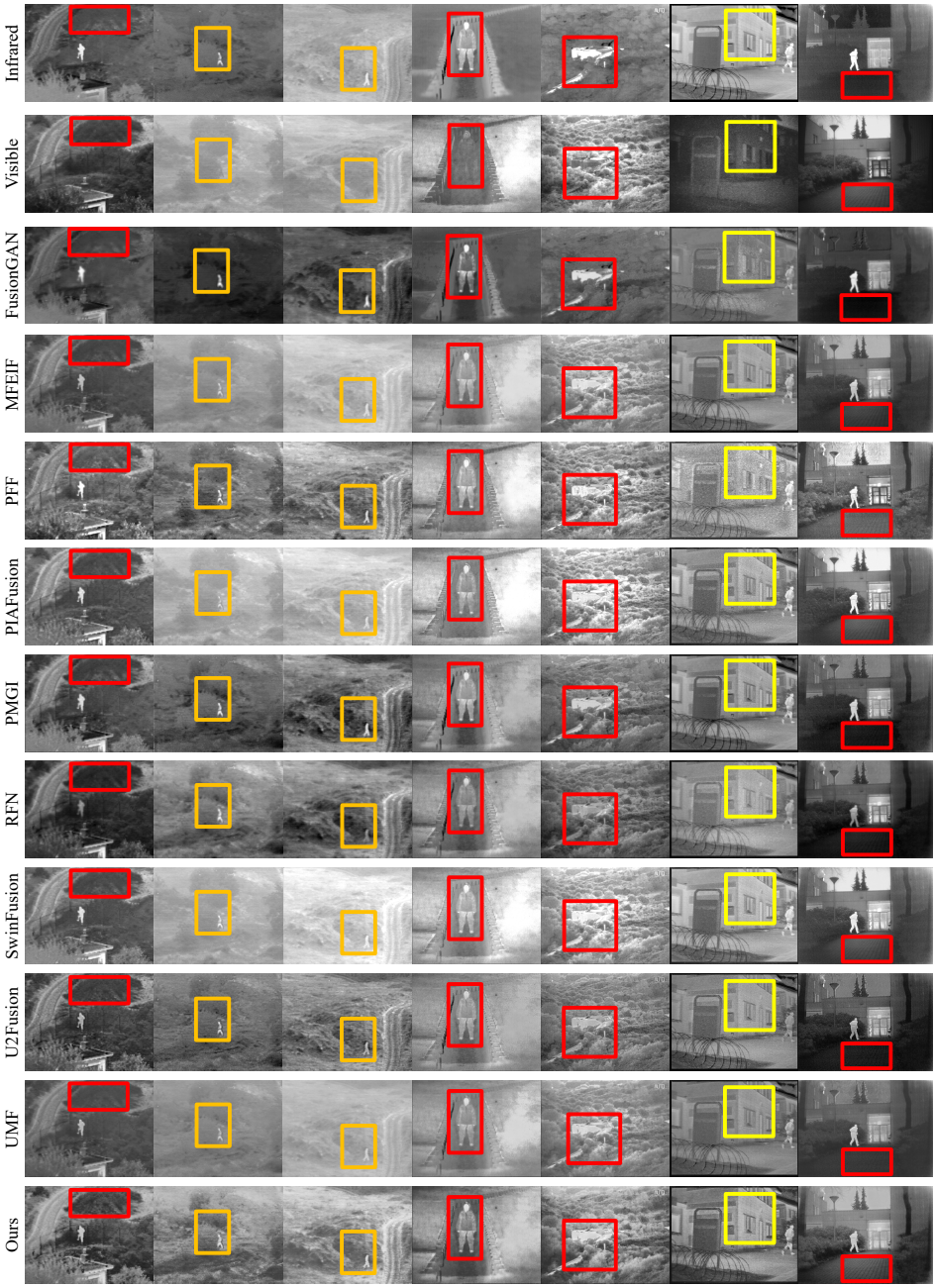


Figure 9. Qualitative comparison of the fused images from various methods on the TNO dataset.

Figures 10 and 11 show more fused images on the RoadScene dataset. In the red boxes, it can be seen that the fused images obtained from PFF contain more visible information and lees infrared information. In the fusion results obtained by FusionGAN, PMGI, and RFN, the overall brightness of the image is relatively low, leading to objects in the fused image that are not salient (see the red boxes). FusionGAN, PMGI, and RFN generate fusion results with low overall brightness, resulting in less salient objects (see the red boxes). Although MFEIF, PIAFusion, SwinFusion, and UMF produce brighter fusion results, their results appear less contrasted in Figures 10 and 11. In the orange boxes of Figure 11, the fusion result from PIAFusion and SwinFusion exhibits blurry details for the cloud, and the results of UMF and U2Fusion are unable to successfully process object edges (see the edge of the tree in orange boxes). In comparison, our method can achieve superior fusion performance in both day and night conditions. The fusion results obtained by our UNIFusion can effectively integrate the source information from infrared and visible images, and it exhibits better performance on the edges of the target.

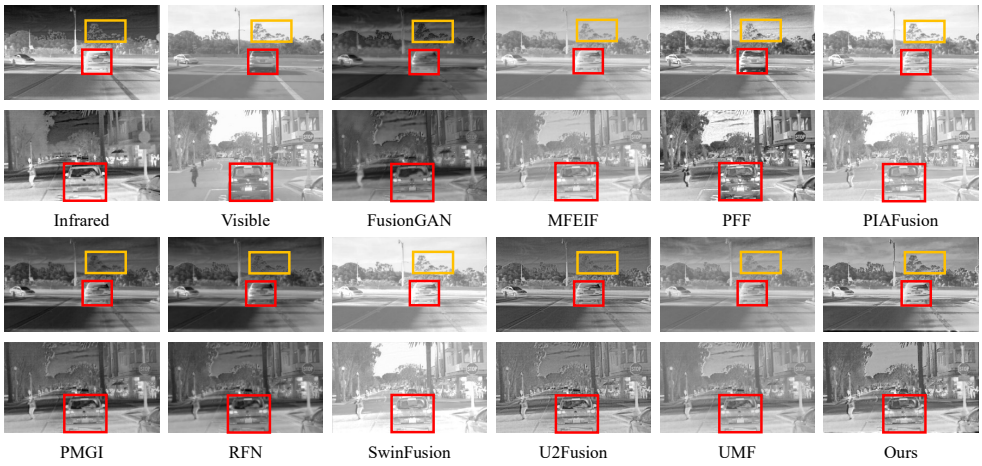


Figure 10. Qualitative comparison of the fused images from various methods on the RoadScene dataset.

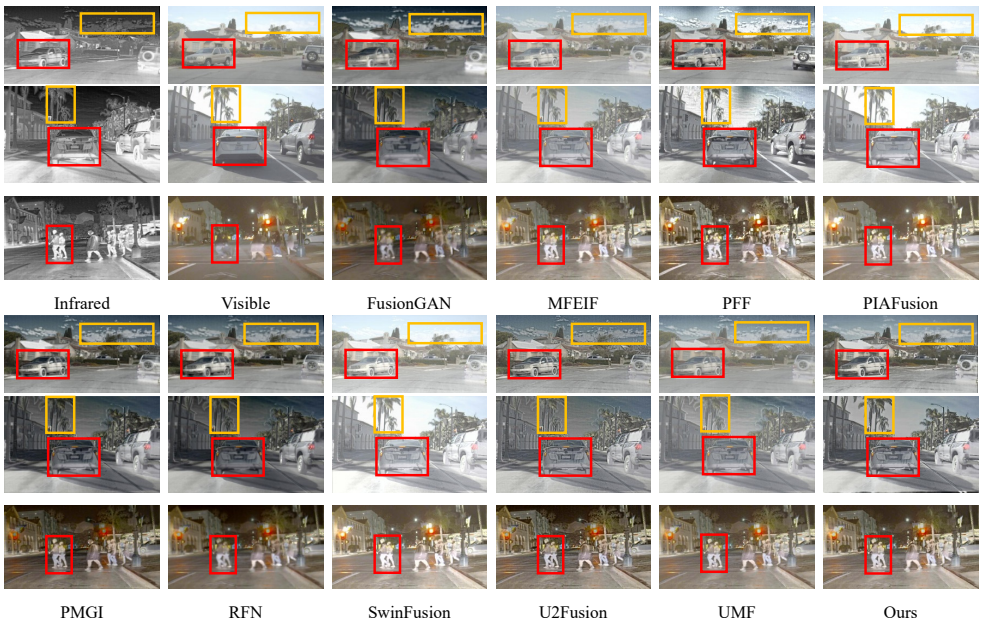
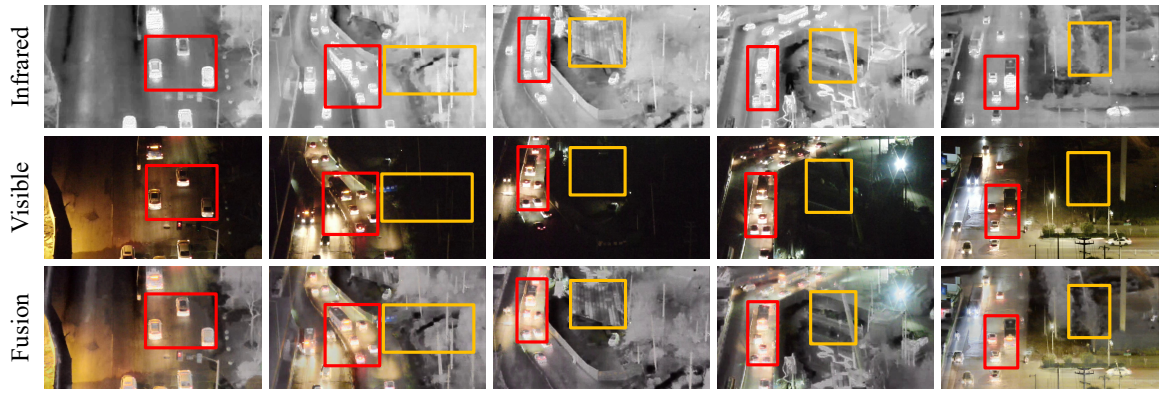


Figure 11. Qualitative comparison of the color fused images from various methods on the RoadScene dataset.

To assess the generalization of our method and its performance in low-light conditions, we conducted experiments on the VTUAV dataset. Figure 12 displays our fusion results, with Figure 12a showing the fusion results under normal-light conditions, and Figure 12b showcasing the fusion results under low-light conditions. In the normal-light scene (see the red boxes in Figure 12a), the infrared images display high thermal contrast, which our algorithm effectively integrates with the visible spectrum images, known for their rich contextual details. The resulting fusion images demonstrate the algorithm's proficiency in synthesizing the distinct attributes of each spectrum to enhance the overall image quality. Under low-light conditions (see the red boxes in Figure 12b), where visible images suffer from limited visibility, our algorithm leverages infrared imaging to accentuate thermal details otherwise obscured by darkness. The fusion process yields images that not only retain the luminance from visible light but also highlight thermal aspects, thus improving the interpretability of the scene in suboptimal lighting.

(a) Normal-light scene



(b) Low-light scene

Figure 12. Fused images in normal and low-light scenes on the VTUAV dataset. The orange boxes show our fusion results in very low-light areas.

We evaluate the performance of our method using remote sensing data that include natural environments, urban landscapes, and beach scenes. Figure 13 shows our fused images in these environments. Our fusion method effectively integrates valuable information from the source images, achieving satisfactory results in terms of illumination, detail, and structural integrity. The fused images across the first, second, and third columns exhibit our method's capability to successfully fuse infrared and visible data, enhancing the clarity in details and structures, as highlighted in the red boxes. Moreover, our approach excels at retaining essential features while disregarding irrelevant information, as seen in the urban and beach scenes of the fourth and fifth columns, respectively. Despite the visible images in the fourth and fifth columns being somewhat dark and containing some details, our fusion outcome maintains these details without being affected by the abnormal

illumination of the visible image. Our method is robust in preserving critical information across diverse scenes and lighting conditions.

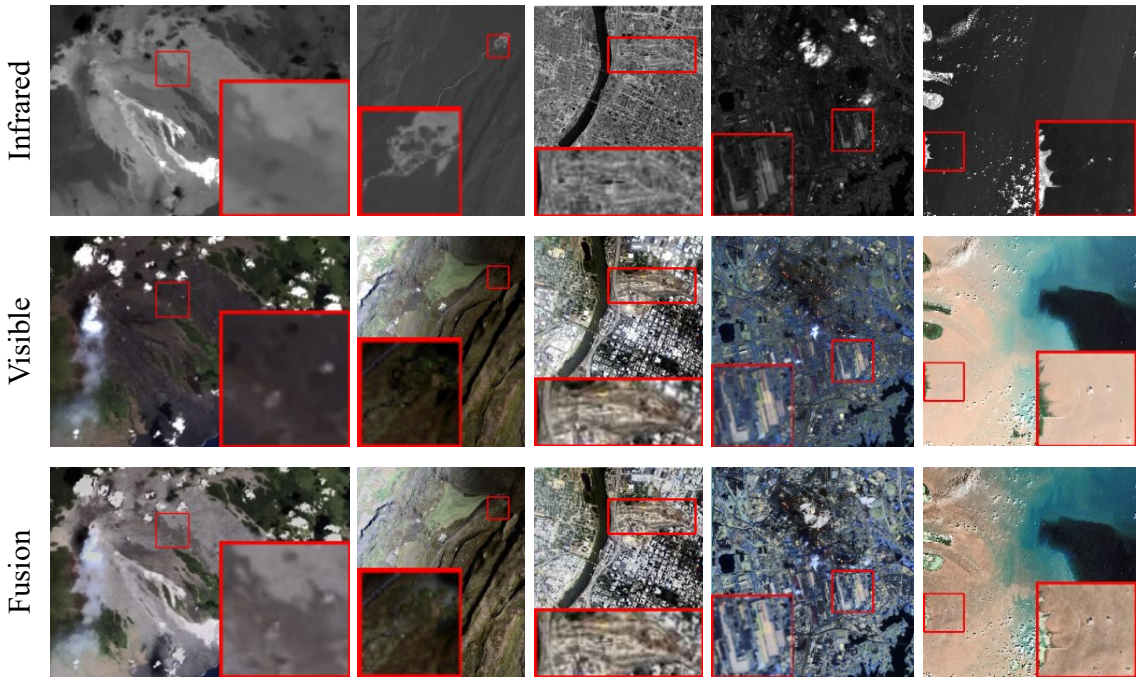


Figure 13. Fusion results in remote sensing imagery. The red boxes are enlarged to highlight the fusion performance on image details.

4.3.2. Qualitative Results

Figures 14 and 15 provide a quantitative comparison between our method and the state-of-the-art (SOTA) methods on the TNO and RoadScene datasets, respectively. The average metric values for these methods are summarized in Tables 1 and 2, respectively. Our method stands out in terms of overall performance. On the TNO dataset, our UNIFusion obtains better performance with the highest average values of SD and CC, indicating the effective integration of information from the source images while preserving the rich details in the fused images. Additionally, our method achieves the second-best results in AG and MS-SSIM, coming close to the top performer. This demonstrates our method's capability to integrate detailed information from source images effectively. In the RoadScene dataset's results, our method obtains remarkably high scores in AG, SD, and CC, further confirming its outstanding overall performance. While PFF achieves the best metrics in AG and SF by incorporating the characteristics of the human visual system, it relies on complex decomposition algorithms and faces challenges in preserving the rich information from the source images.

Table 1. Quantitative analysis on the TNO dataset. The best results are highlighted in red, the second-best in pink, and the third-best in orange.

Methods	AG	SD	SF	MS-SSIM	CC
FusionGAN [26]	3.41 ± 1.27	30.73 ± 6.10	4.32 ± 1.26	0.754 ± 0.10	0.761 ± 0.10
MFEIF [48]	4.24 ± 1.90	34.85 ± 8.26	4.86 ± 1.39	**0.914 ± 0.03**	0.771 ± 0.13
PFF [47]	**10.02 ± 4.40**	40.52 ± 7.24	**8.76 ± 1.61**	0.782 ± 0.09	0.722 ± 0.13
PIAFusion [18]	6.69 ± 3.27	**41.95 ± 11.48**	6.77 ± 1.73	0.860 ± 0.06	0.752 ± 0.13
PMGI [29]	4.86 ± 1.43	39.12 ± 4.04	5.51 ± 1.30	0.912 ± 0.07	0.750 ± 0.13
RFN [50]	3.40 ± 1.11	**43.89 ± 9.63**	4.24 ± 1.16	0.896 ± 0.05	**0.780 ± 0.15**
SwinFusion [17]	6.52 ± 2.90	39.74 ± 10.88	6.72 ± 1.62	0.890 ± 0.06	0.758 ± 0.13
U2Fusion [15]	**6.91 ± 2.20**	40.06 ± 7.42	**7.11 ± 1.42**	**0.931 ± 0.03**	**0.779 ± 0.14**
UMF [49]	4.63 ± 1.87	32.60 ± 6.83	5.15 ± 1.41	0.896 ± 0.07	0.768 ± 0.14
Ours	**8.24 ± 3.58**	**45.20 ± 4.70**	**7.06 ± 0.65**	**0.928 ± 0.03**	**0.795 ± 0.14**

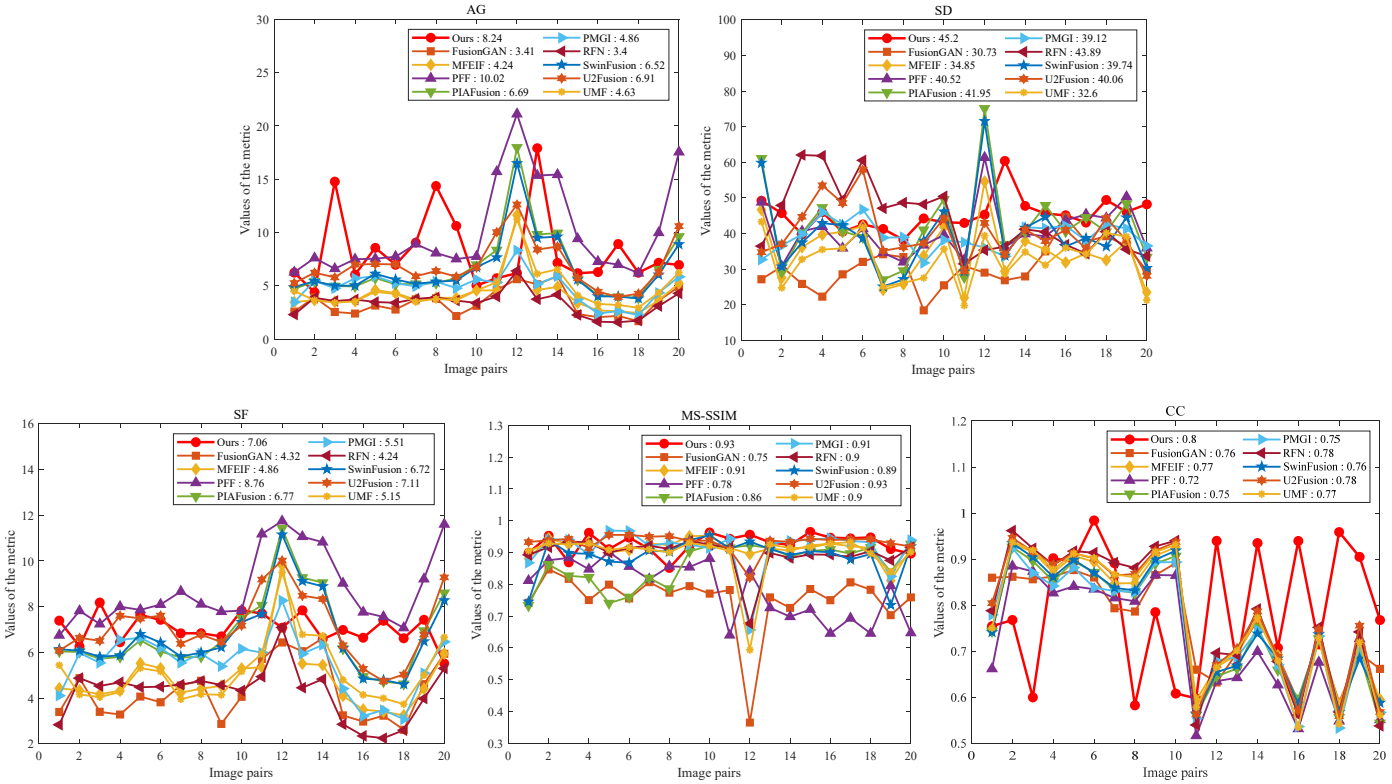


Figure 14. Comparative analysis of nine state-of-the-art methods using five metrics on the TNO dataset.

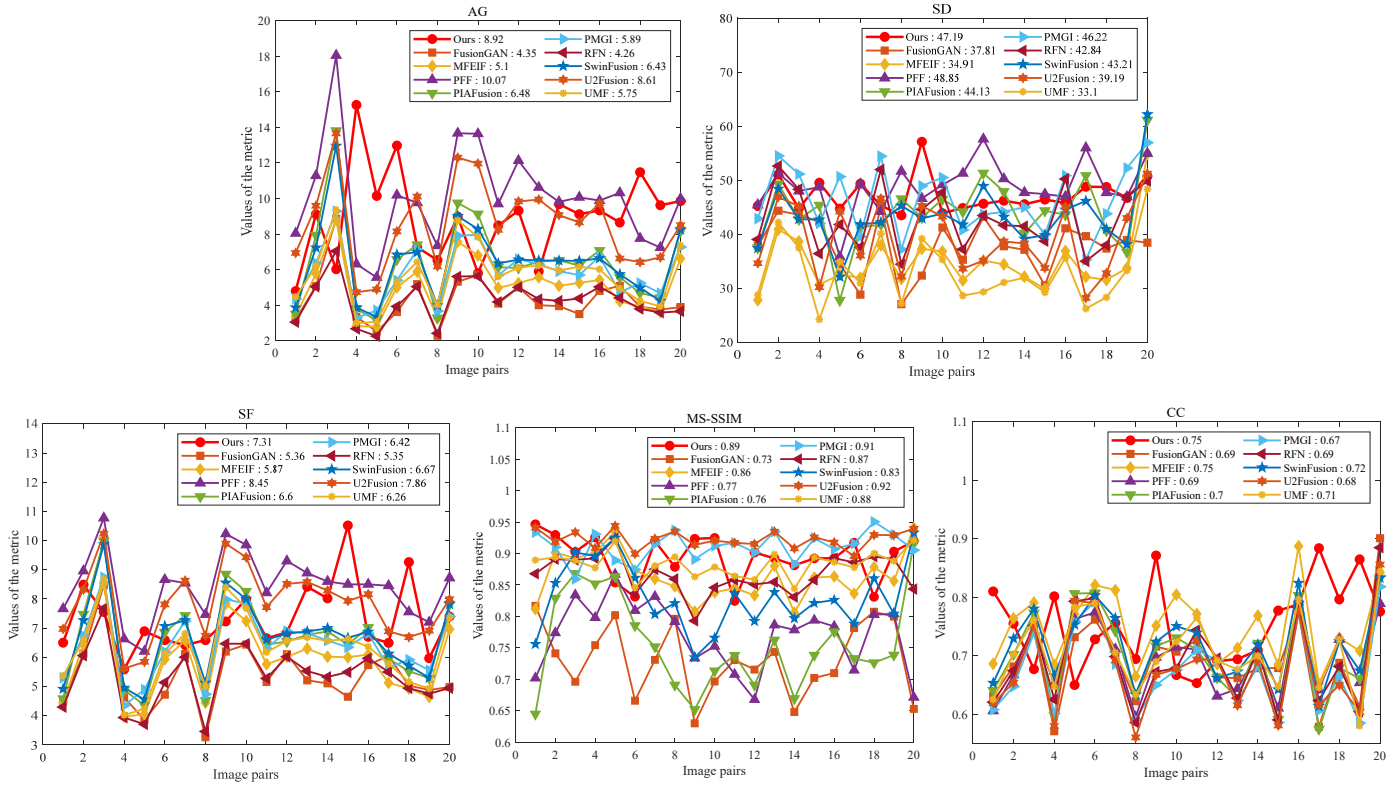


Figure 15. Comparative analysis of nine state-of-the-art methods using five metrics on the RoadScene dataset.

Table 2. Quantitative analysis on the RoadScene dataset. The best results are highlighted in red, the second-best in pink, and the third-best in orange.

Methods	AG	SD	SF	MS-SSIM	CC
FusionGAN [26]	4.35 ± 1.41	37.81 ± 5.17	5.36 ± 1.11	0.731 ± 0.06	0.692 ± 0.07
MFEIF [48]	5.1 ± 1.58	34.91 ± 5.77	5.87 ± 1.24	0.864 ± 0.04	0.750 ± 0.06
PFF [47]	**10.07 ± 2.86**	**48.85 ± 4.64**	**8.45 ± 1.13**	0.770 ± 0.05	0.692 ± 0.06
PIAFusion [18]	6.48 ± 2.55	44.13 ± 6.70	6.60 ± 1.49	0.757 ± 0.08	0.701 ± 0.08
PMGI [29]	5.89 ± 1.58	46.22 ± 6.25	6.42 ± 1.14	0.911 ± 0.02	0.674 ± 0.07
RFN [50]	4.26 ± 1.18	42.84 ± 5.85	5.35 ± 1.02	0.867 ± 0.03	0.692 ± 0.08
SwinFusion [17]	6.43 ± 2.22	43.21 ± 5.88	6.67 ± 1.34	0.831 ± 0.05	0.718 ± 0.06
U2Fusion [15]	8.61 ± 2.39	39.19 ± 6.59	7.86 ± 1.23	**0.923 ± 0.01**	0.678 ± 0.08
UMF [49]	5.75 ± 1.72	33.10 ± 6.06	6.26 ± 1.24	0.883 ± 0.02	0.707 ± 0.07
Ours	8.92 ± 2.51	47.19 ± 3.24	7.31 ± 1.19	0.895 ± 0.02	**0.752 ± 0.07**

4.4. Ablation Study

We conducted experiments to analyze the effectiveness of the proposed method for infrared and visible image fusion. The fusion results with and without the unified feature space (UFS), non-local Gaussian filter (NGF), and dense attention (DA) were compared in the experiments. Figure 16 shows the fused images with and without UFS. It can be seen that the method without UFS generates blurred text on the signboard (see the red boxes in the first row of Figure 16) and does not sufficiently retain the information from the source images. In contrast, our method with UFS produces a detailed fusion result, particularly with much clearer text. From the second row of Figure 16, it can be observed that our method can retain more details of the car compared with the method without UFS. Furthermore, the red boxes in the first row of Figure 16 show that our method generates clearer edges on the signboard, indicating that the unified feature space (UFS) effectively fuses information from different modalities, thereby achieving high fusion performance. In the absence of NGF, there is an increase in noise within the fused image (see the red box in Figure 17). Compared with the method without NGF, our method not only removes more noise but also preserves image details and structures. We propose the dense attention-based feature extraction module to obtain multi-scale features, which can learn the significant features and relationships between different layers. Without dense attention, the extraction of key features becomes challenging, resulting in fusion outcomes that are lacking in detail. In Figure 18, without dense attention (DA), features such as the clouds in the sky and people on the grass appear less prominent and blurred. In contrast, our fusion results are richer in detail and clarity.

Figure 16. The fused images with and without the unified feauture space (UFS). The red boxes are enlarged to highlight the fusion performance on image details.

Figure 17. The fused images with and without the non-local Gaussian filter (NGF). The red boxes are enlarged to highlight the fusion performance on image details.

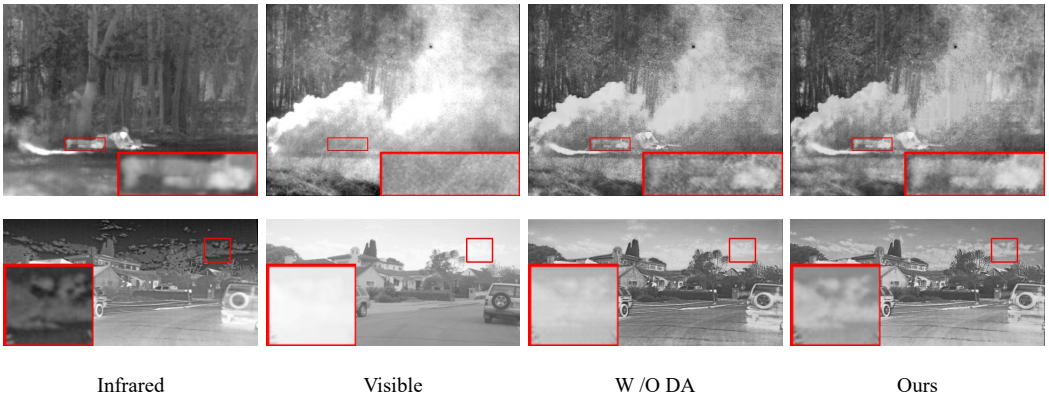


Figure 18. The fused images with and without the dense attention (DA). The red boxes are enlarged to highlight the fusion performance on image details.

We selected three representative metrics to demonstrate the effectiveness of each module: AG, MS-SSIM, and CC. AG indicates that the image contains rich information, while MS-SSIM and CC suggest that the fusion results retain substantial content from the source images. Table 3 presents the comparison results, which demonstrate that each component influences the overall performance. The removal of UFS lead to a marked decrease in AG, indicating its vital role in the fusion process and in maintaining rich information. The absence of NGF and DA leads to a decrease in MS-SSIM, as shown in Table 3, which shows that our proposed NGF and DA are capable of retaining more information from the source image. The absence of DA leading to a significant decrease in MS-SSIM indicates that DA captures essential features, thereby enriching the fusion results with more details from the source images. Both the qualitative and quantitative results demonstrate that the UFS, NGF, and DA are effective in removing noise while maintaining the information from the source images.

Table 3. The results of the ablation study on the TNO dataset. The best results are highlighted in red.

Methods	AG	MS-SSIM	CC
W/O UFS	7.94 ± 3.24	0.927 ± 0.03	0.79 ± 0.14
W/O NGF	8.10 ± 3.30	0.920 ± 0.03	0.79 ± 0.13
W/O DA	8.05 ± 3.17	0.908 ± 0.04	0.79 ± 0.13
Ours	**8.24 ± 3.58**	**0.930 ± 0.03**	**0.80 ± 0.14**

5. Conclusions

In this paper, we fuse infrared and visible images through feature-based decomposition and domain normalization. This decomposition method separates infrared and visible images into common and unique regions. We apply domain normalization to the common regions within the unified feature space to reduce modal differences while retaining unique information. The domain normalization is achieved by transforming the infrared features into a pseudo-visible domain via the unified feature space based on dynamic instance normalization (DIN). Thus, we create a consistent space for the fusion of information from diverse source images, while eliminating modal differences that affect the fusion process. To effectively extract essential features, we integrate a novel dense attention into the feature extraction process. The dense attention ensures that the network can dynamically capture key information across various layers, thereby improving the overall fusion performance in comparison to existing CNN-based methods, autoencoder-based approaches, and others. As the source images may contain noise, we propose a non-local Gaussian filter with learnable filter kernels that depend on the image content. This approach filters out noise while preserving the image details and structure. The experimental results indicate that our method can achieve fusion results of higher quality.

Author Contributions: Conceptualization, W.C. and Y.W.; methodology, Y.W. and W.C.; software, Y.W.; validation, Y.W.; formal analysis, W.C., Z.Z., and L.M.; investigation, Y.W. and Y.Q.; resources, L.M. and Z.Z.; data curation, W.C., L.M., and Z.Z.; writing—original draft preparation, Y.W.; writing—review and editing, W.C. and Y.W.; visualization, Y.W. and Y.Q.; supervision, L.M. and Z.Z.; project administration, W.C and Y.W. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Natural Science Foundation of China under Grant 62173040 and Grant 62071036.

Data Availability Statement: The source code of the paper is available at https://github.com/wyhlaowang/DNFusion (accessed on 28 February 2024).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Ma, J.; Ma, Y.; Li, C. Infrared and visible image fusion methods and applications: A survey. *Inf. Fusion* **2019**, *45*, 153–178. [CrossRef]
2. Yang, Y.; Zhang, Y.; Huang, S.; Zuo, Y.; Sun, J. Infrared and visible image fusion using visual saliency sparse representation and detail injection model. *IEEE Trans. Instrum. Meas.* **2020**, *70*, 5001715. [CrossRef]
3. Hait, E.; Gilboa, G. Spectral total-variation local scale signatures for image manipulation and fusion. *IEEE Trans. Image Process.* **2018**, *28*, 880–895. [CrossRef]
4. Vishwakarma, A.; Bhuyan, M.K. Image fusion using adjustable non-subsampled shearlet transform. *IEEE Trans. Instrum. Meas.* **2018**, *68*, 3367–3378. [CrossRef]
5. Zhou, Z.; Wang, B.; Li, S.; Dong, M. Perceptual fusion of infrared and visible images through a hybrid multi-scale decomposition with Gaussian and bilateral filters. *Inf. Fusion* **2016**, *30*, 15–26. [CrossRef]
6. Zhou, Z.; Dong, M.; Xie, X.; Gao, Z. Fusion of infrared and visible images for night-vision context enhancement. *Appl. Opt.* **2016**, *55*, 6480–6490. [CrossRef]
7. Li, H.; Wu, X.J.; Kittler, J. MDLatLRR: A novel decomposition method for infrared and visible image fusion. *IEEE Trans. Image Process.* **2020**, *29*, 4733–4746. [CrossRef]
8. Bavirisetti, D.P.; Dhuli, R. Fusion of infrared and visible sensor images based on anisotropic diffusion and Karhunen-Loeve transform. *IEEE Sens. J.* **2015**, *16*, 203–209. [CrossRef]
9. Cvejic, N.; Bull, D.; Canagarajah, N. Region-based multimodal image fusion using ICA bases. *IEEE Sens. J.* **2007**, *7*, 743–751. [CrossRef]
10. Wan, T.; Canagarajah, N.; Achim, A. Segmentation-driven image fusion based on alpha-stable modeling of wavelet coefficients. *IEEE Trans. Multimed.* **2009**, *11*, 624–633. [CrossRef]
11. Han, J.; Pauwels, E.J.; De Zeeuw, P. Fast saliency-aware multi-modality image fusion. *Neurocomputing* **2013**, *111*, 70–80. [CrossRef]
12. Ellmauthaler, A.; da Silva, E.A.; Pagliari, C.L.; Neves, S.R. Infrared-visible image fusion using the undecimated wavelet transform with spectral factorization and target extraction. In Proceedings of the 2012 19th IEEE International Conference on Image Processing, Orlando, FL, USA, 30 September–3 October 2012; IEEE: Piscataway, NJ, USA, 2012; pp. 2661–2664.

13. Liu, Y.; Chen, X.; Cheng, J.; Peng, H.; Wang, Z. Infrared and visible image fusion with convolutional neural networks. *Int. J. Wavelets Multiresolution Inf. Process.* **2018**, *16*, 1850018. [CrossRef]
14. Li, Q.; Han, G.; Liu, P.; Yang, H.; Chen, D.; Sun, X.; Wu, J.; Liu, D. A multilevel hybrid transmission network for infrared and visible image fusion. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 1–14. [CrossRef]
15. 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]
16. 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*, 5002215. [CrossRef]
17. Ma, J.; Tang, L.; Fan, F.; Huang, J.; Mei, X.; Ma, Y. SwinFusion: Cross-domain long-range learning for general image fusion via swin transformer. *IEEE/CAA J. Autom. Sin.* **2022**, *9*, 1200–1217. [CrossRef]
18. 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]
19. Lebedev, M.; Komarov, D.; Vygolov, O.; Vizilter, Y.V. Multisensor image fusion based on generative adversarial networks. In Proceedings of the Image and Signal Processing for Remote Sensing XXV, Strasbourg, France, 9–11 Septmeber 2019; SPIE: Bellingham, WA, USA, 2019; Volume 11155, pp. 565–574.
20. Cui, Y.; Du, H.; Mei, W. Infrared and visible image fusion using detail enhanced channel attention network. *IEEE Access* **2019**, *7*, 182185–182197. [CrossRef]
21. Li, Y.; Wang, J.; Miao, Z.; Wang, J. Unsupervised densely attention network for infrared and visible image fusion. *Multimed. Tools Appl.* **2020**, *79*, 34685–34696. [CrossRef]
22. Li, H.; Wu, X.J. DenseFuse: A fusion approach to infrared and visible images. *IEEE Trans. Image Process.* **2018**, *28*, 2614–2623. [CrossRef]
23. Hou, R.; Zhou, D.; Nie, R.; Liu, D.; Xiong, L.; Guo, Y.; Yu, C. VIF-Net: An unsupervised framework for infrared and visible image fusion. *IEEE Trans. Comput. Imaging* **2020**, *6*, 640–651. [CrossRef]
24. Liu, L.; Chen, M.; Xu, M.; Li, X. Two-stream network for infrared and visible images fusion. *Neurocomputing* **2021**, *460*, 50–58. [CrossRef]
25. Liao, B.; Du, Y.; Yin, X. Fusion of infrared-visible images in UE-IoT for fault point detection based on GAN. *IEEE Access* **2020**, *8*, 79754–79763. [CrossRef]
26. 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]
27. 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]
28. 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]
29. Zhang, H.; Xu, H.; Xiao, Y.; Guo, X.; Ma, J. Rethinking the image fusion: A fast unified image fusion network based on proportional maintenance of gradient and intensity. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 12797–12804.
30. 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]
31. Xu, D.; Zhang, N.; Zhang, Y.; Li, Z.; Zhao, Z.; Wang, Y. Multi-scale unsupervised network for infrared and visible image fusion based on joint attention mechanism. *Infrared Phys. Technol.* **2022**, *125*, 104242. [CrossRef]
32. Li, J.; Huo, H.; Li, C.; Wang, R.; Feng, Q. AttentionFGAN: Infrared and visible image fusion using attention-based generative adversarial networks. *IEEE Trans. Multimed.* **2020**, *23*, 1383–1396. [CrossRef]
33. Yuan, C.; Sun, C.; Tang, X.; Liu, R. Flgc-fusion gan: An enhanced fusion gan model by importing fully learnable group convolution. *Math. Probl. Eng.* **2020**, *2020*, 6384831 . [CrossRef]
34. Isola, P.; Zhu, J.Y.; Zhou, T.; Efros, A.A. Image-to-image translation with conditional adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1125–1134.
35. Wang, T.C.; Liu, M.Y.; Zhu, J.Y.; Tao, A.; Kautz, J.; Catanzaro, B. High-resolution image synthesis and semantic manipulation with conditional gans. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8798–8807.
36. Kim, T.; Cha, M.; Kim, H.; Lee, J.K.; Kim, J. Learning to discover cross-domain relations with generative adversarial networks. In Proceedings of the International Conference on Machine Learning, PMLR 2017, Sydney, Australia, 6–11 August 2017; pp. 1857–1865.
37. Liu, M.Y.; Breuel, T.; Kautz, J. Unsupervised image-to-image translation networks. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017; Volume 30.
38. Zhu, J.Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2223–2232.
39. Zhu, J.Y.; Zhang, R.; Pathak, D.; Darrell, T.; Efros, A.A.; Wang, O.; Shechtman, E. Toward multimodal image-to-image translation. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017; Volume 30.

40. Bousmalis, K.; Silberman, N.; Dohan, D.; Erhan, D.; Krishnan, D. Unsupervised pixel-level domain adaptation with generative adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 3722–3731.
41. Huang, X.; Belongie, S. Arbitrary style transfer in real-time with adaptive instance normalization. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 1501–1510.
42. Jing, Y.; Liu, X.; Ding, Y.; Wang, X.; Ding, E.; Song, M.; Wen, S. Dynamic instance normalization for arbitrary style transfer. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 4369–4376.
43. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [CrossRef]
44. 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 IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 5802–5811.
45. Toet, A. The TNO multiband image data collection. *Data Brief* **2017**, *15*, 249–251. [CrossRef] [PubMed]
46. Zhang, P.; Zhao, J.; Wang, D.; Lu, H.; Ruan, X. Visible-thermal UAV tracking: A large-scale benchmark and new baseline. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 8886–8895.
47. Zhou, Z.; Fei, E.; Miao, L.; Yang, R. A perceptual framework for infrared–visible image fusion based on multiscale structure decomposition and biological vision. *Inf. Fusion* **2023**, *93*, 174–191. [CrossRef]
48. Liu, J.; Fan, X.; Jiang, J.; Liu, R.; Luo, Z. Learning a deep multi-scale feature ensemble and an edge-attention guidance for image fusion. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *32*, 105–119. [CrossRef]
49. Di, W.; Jinyuan, L.; Xin, F.; 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 (IJCAI), Vienna, Austria, 23–29 July 2022.
50. 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]
51. Cui, G.; Feng, H.; Xu, Z.; Li, Q.; Chen, Y. Detail preserved fusion of visible and infrared images using regional saliency extraction and multi-scale image decomposition. *Opt. Commun.* **2015**, *341*, 199–209. [CrossRef]
52. Deshmukh, M.; Bhosale, U.; et al. Image fusion and image quality assessment of fused images. *Int. J. Image Process. (IJIP)* **2010**, *4*, 484.
53. Roberts, J.W.; Van Aardt, J.A.; Ahmed, F.B. Assessment of image fusion procedures using entropy, image quality, and multispectral classification. *J. Appl. Remote Sens.* **2008**, *2*, 023522.
54. Ma, K.; Zeng, K.; Wang, Z. Perceptual quality assessment for multi-exposure image fusion. *IEEE Trans. Image Process.* **2015**, *24*, 3345–3356. [CrossRef]