DOI: 10.23977/acss.2024.080708
ISSN 2371-8838 Vol. 8 Num. 7

A Frequency Decomposition and Gaussian-Based Enhancement Network for Infrared and Visible Image Fusion

Ting Wang[1,a], Hao Song[1,b], Xuanyu Liao[1,c], Chengjiang Zhou[1,d,*]

[1]School of Information Science and Technology, Yunnan Normal University, Kunming, China
[a]tingwang0202@163.com, [b]songhao00711@163.com, [c]xuanyuleo@163.com,
[d]chengjiangzhou@foxmail.com
[*]Corresponding author

Keywords: Feature enhancement, Frequency decomposition, Image fusion, Generative adversarial network

Abstract: The purpose of infrared image and visible image fusion is to preserve information in different modalities. In order to solve the redundancy of modal frequency domain information extraction and feature mapping, we propose a frequency decomposition and Gaussian-Based enhancement network for infrared and visible image fusion. Firstly, we design a frequency decomposition convolution, which divides the feature map to realize the independent modeling of different frequency information, so as to extract the deep-level features more accurately. In addition, we design enhancement module combined with Gaussian filter to enhance the feature expression and optimize the loss function. Finally, we introduce dual-discriminators to refine the differentiation of infrared and visible images, significantly enhancing global information expression and detail presentation in fused image. Experimental outcomes demonstrate that our fusion method can effectively integrate the dominant information of the two images. Notably, our method outperforms other advanced fusion algorithms by enhancing the performance of object detection tasks, particularly in terms of improving the accuracy of detecting cars and pedestrians.

1. Introduction

Visible images can retain rich texture details, while infrared images are able to effectively capture thermal radiation information in complex environments. With the advancement of image fusion technology, the information of these two model images has achieved complementary advantages, combining the target information from infrared images with the detail information from visible images, which improves the comprehensive performance of the image. However, most existing fusion algorithms focus on optimizing the fusion quality of images through complex network architectures and mathematical transformations, while ignoring the adaptability to specific tasks, which makes the fusion process difficult to apply to high-level visual tasks. Despite demonstrating certain effectiveness, existing fusion methods still face numerous problems.

1.1. A single discriminator network

The single discriminator network uses only one discriminator to fuse images from one source image. This single input will cause information imbalance in the fusion result, which will be biased towards a certain type of source image and easily cause information loss.

1.2. The feature mapping has redundant information

During the feature extraction process of images through convolution, the feature mapping often captures similar image information, especially in multiple convolution layers, the network often extracts similar features repeatedly. Due to the local receptive field of the convolution kernel, the basic information in the image appear repeatedly in the channel, which leads to the accumulation of redundant information.

To overcome the limitations of current challenges, we propose a new feature decomposition and dual-modality discriminator network for infrared and visible image fusion. In our proposed approach, the entire network framework uses frequency decomposition convolution (FDConv) to extract features across various ranges. However, in response to the modal differences in different images, we designed a discriminator with high-frequency and low-frequency enhancement to address the information imbalance problem caused by a single discriminator. Finally, the fusion results are balanced by content loss, gradient loss and similarity metric (SSIM) loss. We adopt a loss mechanism based on the relativistic average discriminator, which aims to train the discriminator to compare the authenticity of two images rather than simply distinguishing whether an image is real or fake. The contributions of this article mainly include:

1.3. A new fusion network

We propose a new feature decomposition and Gaussian-Based enhancement network, aiming to integrate and enhance information from different image frequencies.

1.4. Frequency decomposition module

We designed the frequency decomposition convolution (FDConv), which includes three modules: DeConv, EnConv and MeConv. The frequency decomposition and fusion of input features are performed through convolution, pooling and upsampling operations, aiming to reduce unnecessary information during the feature mapping process and help us obtain deep feature semantic information.

1.5. Gaussian-Based enhancement module

We designed frequency enhancement and frequency decomposition modules based on Gaussian filter to effectively retain the gradient information of infrared and visible images while enhancing the overall contrast of the image. The frequency enhancement module enhances infrared and visible images respectively, reducing the loss of features in the network.

2. Related work

Research methods are mainly divided into two categories: traditional and deep learning. Traditional methods include multi-scale transformation [1-3], sparse representation [4-6], subspace analysis[7-8], saliency detection[9-10]. These methods use unified transformation for all images, ignoring feature differences between images, which may cause information loss, and their fusion strategies are not refined enough, hindering performance optimization. In addressing the issues with traditional methods, various new image fusion methods have been proposed.

In 2018, Li et al.[11] presented a novel method, termed as DenseFuse, which is based on auto-encoder framework. The encoding network of this fusion method consists of a fusion layer and a dense block. The use of dense block structure effectively captures the deep features and improves the image quality. However, these methods have relatively simple processing mechanisms for differences between modalities and are difficult to fuse complementary information from different modalities, which may lead to the fusion result being overly dependent on a certain modality or losing critical features.

DeepFuse[12] uses feature extraction layers in order to extract common low-frequency features from each input image. The fusion layer then fuses these features to generate a fused feature map, and finally the fused features are passed through the reconstruction layer to obtain the final fused image. U2Fusion[13] adopts the pretrained VGG-16 network for feature extraction, and the adaptive degrees allows the network to be trained. Sea Fusion[14] proposes a semantic-aware image fusion framework. In terms of the overall framework design, SeAFusion uses classic dual-branch feature extraction, and then reconstructs the spliced and fused image. This type of fusion method uses convolutional neural networks to achieve end-to-end feature extraction.

The image fusion problem is modeled as a confrontational game problem between the generator and the discriminator to estimate the probability distribution of the target, thus implicitly completing the three steps of feature extraction, feature fusion and image reconstruction. FusionGAN[15] is the first time that GAN is used to solve the image fusion method, but this model only uses one discriminator and only uses the original visible image as the input of the discriminator, so the output result is biased towards infrared images. Later, Han et al.[16] proposed a generative adversarial network based on dual discriminators for image fusion, which mainly solved the problem that the images generated by a single discriminator are closer to visible images. GANMcC[17] proposed an end-to-end fusion model based on generative adversarial networks with multi-classification constraints. This fusion method preserves contrast and texture details, solving the fusion imbalance problem of existing methods. Song et al.[18] presented a triple-discriminator generative adversarial network, effectively boosting the prominence of infrared targets while maintaining the fine details of visible images.

3. Method

3.1. Frequency Decomposition Convolution

To solve the problem of feature mapping information redundancy, we designed frequency decomposition convolution (FDConv), which maps the input features to the frequency domain through three modules: DeConv, EnConv, and MeConv.

X represents the input feature, the superscripts H and L represent the high-frequency feature and low-frequency feature respectively, and Y represents the output feature. The DeConv module processes features of different frequencies through pooling and convolution operations, decomposing the input feature X into high-frequency features $X^H$and low-frequency features X^L. In the EnConv module, the high-frequency feature X^H is decomposed into X^{H2H} and X^{H2L} through convolution and pooling operations, the high-frequency feature X^L is decomposed into X^{L2H} and X^{L2L} through convolution and upsampling operations.

3.2. Gaussian-Based Enhancement Module.

We uniquely integrated Gaussian filter technology to thoughtfully design the frequency enhancement and decomposition module. The network architecture of our generator is shown in Figure 1. We separately enhanced the low-frequency (L_{en}) and high-frequency ($\mathrm{H_{en}}$) components of both infrared and visible images. Gau (x,y) represents Gaussian filter, which is defined in Equation (4), and I(x,y) represents the input two-dimensional image. The specific implementation is shown in Equation (4) and Equation (4).

$$Gau(x,y) = \frac{1}{2\pi\sigma_x\sigma_y} e^{-\left(\frac{x^2}{2\sigma_x^2}+\frac{y^2}{2\sigma_y^2}\right)}, \tag{1}$$

$$H_{en} = I(x,y) + I(x,y) - Gau(I(x,y)) \tag{2}$$

$$L_{en} = I(x,y) + Gau(I(x,y)). \tag{3}$$

3.3. Discriminator network

To address the modal difference problem of a single discriminator, we use dual discriminators, which input the visible image into the visible discriminator, and the visible image and the fusion image input sum discriminator. We perform low-frequency enhancement and high-frequency enhancement respectively. The network architecture is shown in Figure 1. For the sum discriminator, we perform frequency decomposition on the input image to obtain low-frequency features and high-frequency features. The following two discriminators perform the same operation. Firstly, the DeConv module is decomposed and then added to obtain the basic features (X_b) and detail features (X_d). Then, the low-frequency features (X_L) and high-frequency features(X_H) are obtained by decomposition and addition of the EnConv module. Subsequently, the updated basic features and detail features are further decomposed by this module. Finally, the MeConv module processes the features, and the prediction result is obtained through a 5x5 convolution and a Linear layer.

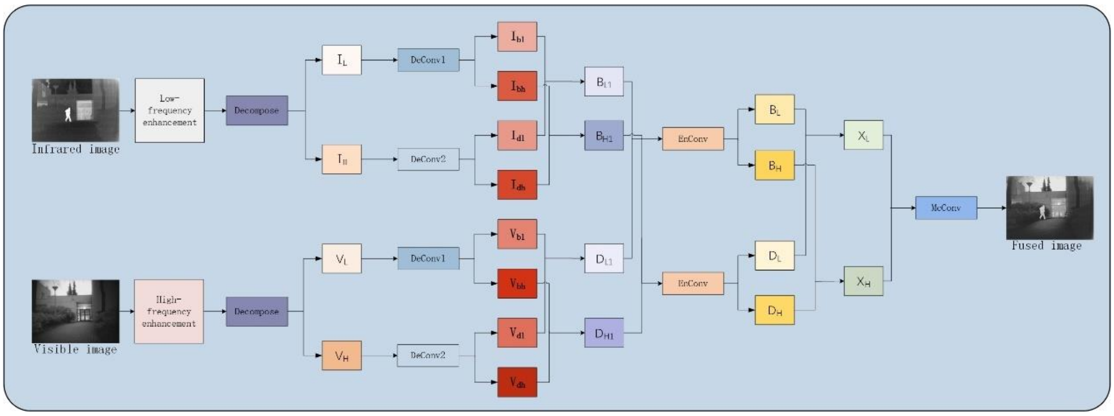

Figure 1: Framework of the fusion network.

3.4. Loss function

The methodology's loss formulation includes the loss for the generator, denoted as L_G, and the loss for the discriminator, denoted as L_D. Subsequently, we will introduce them separately. The generator's loss function consists of three key components: the adversarial loss L_G^{Ra}, the content loss $L_{content}$, and the structural similarity index (SSIM) based measurement loss L_{SSIM}. L_G^{Ra} represents the loss between the generator and RaD. Our generator loss consists of two components, L_G^{Ra}, which includes both $L_{G_v}^{Ra}$ and $L_{G_c}^{Ra}$, representing the contributions of each part. In the description, the adversarial loss between the generator and the visible discriminator is denoted as $L_{G_v}^{Ra}$, while the adversarial loss between the generator and (another) discriminator is represented as $L^{Ra}_{G_s}$. In the Equation (4), the λ_1, λ_2, λ_3 represent the weight proportion of losses in each part. In Equation (5) and (6) , I_v represents the visible image, I_s signifies the combined result of visible and infrared image, whereas I_f stands for the image resulting from the fusion process.

$$L_G = \lambda_1 L^{Ra}_G + \lambda_2 L_{content} + \lambda_3 L_{SSIM} \tag{4}$$

$$L^{Ra}_{G_v} = -E_{I_v}\left[\log\left(1 - D_{Ra}\left(I_{v,}I_f\right)\right)\right] - E_{I_f}\left[log\left(D_{Ra}\left(I_f, I_v\right)\right)\right] \tag{5}$$

$$L^{Ra}_{G_s} = -E_{I_s}\left[log\left(1 - D_{Ra}\left(I_{s,}I_f\right)\right)\right] - E_{I_f}\left[log\left(D_{Ra}\left(I_f, I_s\right)\right)\right] \tag{6}$$

Content loss. Inspired by SeAFusion, the content loss is defined as Equation (7), α_1 and α_2 are tuning parameters. We define the intensity loss of infrared and visible images as Equation (8). ∇ represents the Sobel gradient operator in Equation (9).

$$L_{content} = \alpha_1 L_{int} + \alpha_2 L_{grad} \tag{7}$$

$$L_{int} = \frac{1}{HW}\left\|I_f - max\left(I_i,\ \ I_v\right)\right\| \tag{8}$$

$$L_{grad} = \frac{1}{HW}\left\|\left|\nabla I_f\right| - max\left(\left|\nabla I_i\right|,\ \ \left|\nabla I_v\right|\right)\right\| \tag{9}$$

The structural similarity index measures the loss. SSIM is based on ray similarity (l(x, y)), contrast similarity (c(x, y)), and structural similarity (s(x, y)). The SSIM loss formula is defined as Equation (10).

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\delta_x\delta_y + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\delta_x^2 + \delta_y^2 + c_2)} \tag{10}$$

In our approach, the Structural Similarity Index (SSIM) loss, L_{SSIM}, is formulated as specified in Equation (11). This formulation involves two constants, c_1 and c_2 , which are defined as $c_1 = (0.01 \times L)^2$ and $c_2 = (0.03 \times L)^2$, where L represents the dynamic range of the image pixel values, and 0.01 and 0.03 are the default threshold parameters for these two constants.

$$L_{SSIM} = \eta SSIM\left(I_f,\ \ I_i\right) + SSIM\left(I_f,\ \ I_v\right) \tag{11}$$

In Equation (12), our discriminator loss includes visible discriminator loss $L^{Ra}_{D_v}$ and sum discriminator loss $L^{Ra}_{D_s}$. where $L^{Ra}_{D_v}$ is defined as Equation (13), and $L^{Ra}_{D_s}$ is defined as Equation (14). By evaluating the input images against the fused result, the discriminator iteratively optimizes the loss function, pushing the generator towards achieving higher quality image fusion outcomes.

$$L^{Ra}_D = L^{Ra}_{D_v} + L^{Ra}_{D_s} \tag{12}$$

$$L^{Ra}_{D_v} = -E_{I_v}\left[\log\left(1 - D_{Ra}\left(I_{v,}I_f\right)\right)\right] - E_{I_f}\left[log\left(D_{Ra}\left(I_f, I_v\right)\right)\right] \tag{13}$$

$$L^{Ra}_{D_s} = -E_{I_s}\left[log\left(1 - D_{Ra}\left(I_{s,}I_f\right)\right)\right] - E_{I_f}\left[log\left(D_{Ra}\left(I_f, I_s\right)\right)\right] \tag{14}$$

4. Experiments

4.1. Setup

In the following experiments, we used EN, SD, AG, MI, and SF as evaluation metrics. A qualitative and quantitative comparative analysis was conducted on the TNO dataset. The TNO dataset include various image types such as near-infrared, far-infrared, and thermal-infrared, along with diverse scenarios including nighttime field operations and military activities.

Our entire network was trained using the PyTorch framework on an NVIDIA GeForce RTX 4080 GPU. During the preprocessing stage, we configured both the batch size and training iterations to 16. We employed the AdaBelief optimizer for training our network, initiating the learning rate at 10^{-4}. The stride for each image was set to 14, and each patch had a fixed size of 120 × 120.

4.2. Performance Comparison

To prove the robustness of our method, we used seven advanced fusion methods, including FusionGAN[15], Densefuse[11], IFCNN[19], LRRNet[20], DATFuse[21], IRFS[22], and SemLA[23] for subjective and objective analysis with our method. The above fusion results were evaluated subjectively and objectively.

Qualitative analysis. The qualitative performance is reported in Figure 2. It is worth noting that FusionGAN method has the best effect on texture preservation, and IFCNN method has the best effect on infrared target fusion. However, both methods are biased towards a certain type of source image, while image fusion hopes to provide complementary information to achieve a more comprehensive description of the target. For grass, tree branches, and people's clothing in different scenes in the picture, our method retains texture information and infrared information better than the other seven methods.

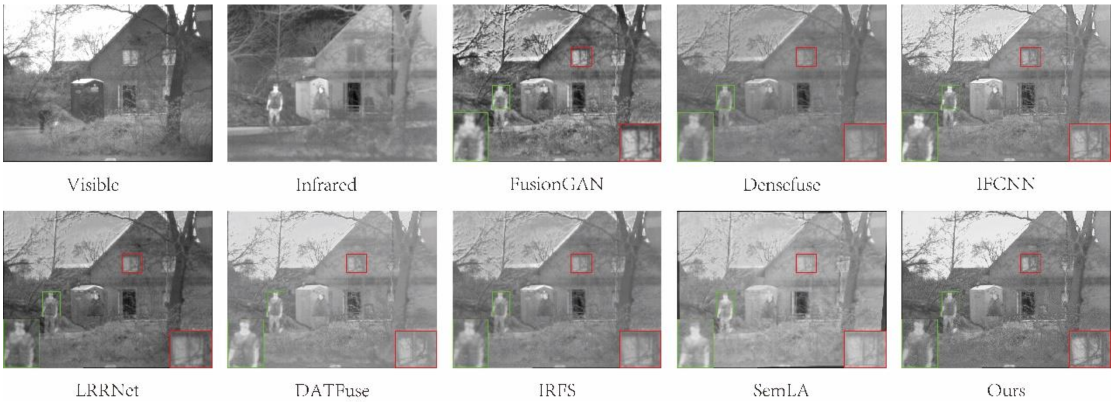

Figure 2: The fusion results of our method compared to other methods on the TNO dataset.

Quantitative analysis. Table 1 demonstrates the quantitative performance of all fusion methods. The experiment show that this method performs best on the four indicators of EN, SD, AG and SF, followed by the MI. It can be seen that our method significantly outperforms other methods in terms of texture details and visual fidelity.

Table 1: The quantitative resultsof our method compared to other methods on the TNO dataset.

Methods	EN	SD	AG	MI	SF
FusionGAN	7.050	37.471	3.899	1.531	8.290
DenseFuse	6.487	26.690	2.727	2.297	6.987
IFCNN	6.872	36.169	5.019	2.465	13.059
LRRNet	6.994	39.375	3.521	2.517	9.112
DATFuse	6.560	29.841	3.745	3.164	10.072
IRFS	6.753	34.169	3.489	2.209	9.745
SemLA	6.832	36.133	3.747	2.147	11.665
Ours	7.061	44.454	6.033	2.798	15.412

4.3. Ablation experiment

The performance of our fusion method relies on dual discriminators, Gaussian modules, and FDConv module.In this section, we conduct a series of ablation studies to verify the effectiveness of the specific module designs. The results of the ablation experiments on the TNO dataset are shown in Table 2.

Table 2: The ablation experiment results on the TNO dataset.

	EN	SD	AG	SF
Ours	7.061	44.454	6.033	15.412
w/o D	6.858	40.991	3.708	9.512
w/o Gau	5.998	24.762	1.699	4.206
w/o Deconv	7.048	42.484	4.794	12.641
w/o Enconv	6.610	34.483	2.713	6.612
w/o Meconv	6.843	46.531	4.398	11.020

5. Conclusion

We propose a new feature decomposition and Gaussian-Based Enhancement network. Considering the potential redundancy in feature mappings, we have devised a frequency decomposition convolution module to manage and reduce it. Employing a frequency enhancement and decomposition strategy based on Gaussian filter, we first conduct low-frequency enhancement on infrared images to highlight their overall structure and thermal radiation characteristics, while applying high-frequency enhancement to visible images to sharpen their detailed textures and edge information. Subsequently, utilizing frequency decomposition techniques, we decompose these enhanced images into distinct frequency components.

However, our method has limitations in the case of inconsistent exposure, which leads to the imbalance of image contrast and affects the visual quality of the image. Therefore, we hope to adjust the image locally and design an adaptive brightness module to perform different processing on the brightness features of different regions.

References

[1] Hu Y, He J, Xu L. Infrared and visible image fusion based on multiscale decomposition with Gaussian and co-occurrence filters[C]//2021 4th International Conference on Pattern Recognition and Artificial Intelligence (PRAI). IEEE, 2021: 46-50.

[2] Yang C, He Y, Sun C, et al. Infrared and visible image fusion based on QNSCT and Guided Filter[J]. Optik, 2022, 253: 168592.

[3] Li L, Ma H, Jia Z, et al. A novel multiscale transform decomposition based multi-focus image fusion framework[J]. Multimedia Tools and Applications, 2021, 80: 12389-12409.

[4] Tan J, Zhang T, Zhao L, et al. Multi-focus image fusion with geometrical sparse representation[J]. Signal Processing: Image Communication, 2021, 92: 116130.

[5] Xing C, Wang M, Dong C, et al. Using Taylor expansion and convolutional sparse representation for image fusion [J]. Neurocomputing, 2020, 402: 437-455.

[6] Li H, Wu X J, Kittler J. MDLatLRR: A novel decomposition method for infrared and visible image fusion[J]. IEEE Transactions on Image Processing, 2020, 29: 4733-4746.

[7] Bavirisetti D P, Xiao G, Liu G. Multi-sensor image fusion based on fourth order partial differential equations[C]//2017 20th International conference on information fusion (Fusion). IEEE, 2017: 1-9.

[8] Fu Z, Wang X, Xu J, et al. Infrared and visible images fusion based on RPCA and NSCT[J]. Infrared Physics & Technology, 2016, 77: 114-123.

[9] Chen J, Wu K, Cheng Z, et al. A saliency-based multiscale approach for infrared and visible image fusion[J]. Signal Processing, 2021, 182: 107936.

[10] Yang Y, Zhang Y, Huang S, et al. Infrared and visible image fusion using visual saliency sparse representation and detail injection model[J]. IEEE Transactions on Instrumentation and Measurement, 2020, 70: 1-15.

[11] Li H, Wu X J. DenseFuse: A fusion approach to infrared and visible images[J]. IEEE Transactions on Image Processing, 2018, 28(5): 2614-2623.

[12] Ram Prabhakar K, Sai Srikar V, Venkatesh Babu R. Deepfuse: A deep unsupervised approach for exposure fusion with extreme exposure image pairs[C]//Proceedings of the IEEE international conference on computer vision. 2017: 4714-4722.

[13] Xu H, Ma J, Jiang J, et al. U2Fusion: A unified unsupervised image fusion network[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 44(1): 502-518.

[14] 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[J]. Information Fusion, 2022, 82: 28-42.

[15] Ma J, Yu W, Liang P, et al. FusionGAN: A generative adversarial network for infrared and visible image fusion[J]. Information fusion, 2019, 48: 11-26.

[16] Ma J, Xu H, Jiang J, et al. DDcGAN: A dual-discriminator conditional generative adversarial network for multi-resolution image fusion[J]. IEEE Transactions on Image Processing, 2020, 29: 4980-4995.

[17] Ma J, Zhang H, Shao Z, et al. GANMcC: A generative adversarial network with multiclassification constraints for infrared and visible image fusion[J]. IEEE Transactions on Instrumentation and Measurement, 2020, 70: 1-14.

[18] Song A, Duan H, Pei H, et al. Triple-discriminator generative adversarial network for infrared and visible image fusion[J]. Neurocomputing, 2022, 483: 183-194.

[19] Zhang Y, Liu Y, Sun P, et al. IFCNN: A general image fusion framework based on convolutional neural network[J]. Information Fusion, 2020, 54: 99-118.

[20] Li H, Xu T, Wu X J, et al. Lrrnet: A novel representation learning guided fusion network for infrared and visible images[J]. IEEE transactions on pattern analysis and machine intelligence, 2023, 45(9): 11040-11052.

[21] Tang W, He F, Liu Y, et al. DATFuse: Infrared and visible image fusion via dual attention transformer[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 33(7): 3159-3172.

[22] Wang D, Liu J, Liu R, et al. An interactively reinforced paradigm for joint infrared-visible image fusion and saliency object detection[J]. Information Fusion, 2023, 98: 101828.

[23] Xie H, Zhang Y, Qiu J, et al. Semantics lead all: Towards unified image registration and fusion from a semantic perspective[J]. Information Fusion, 2023, 98: 101835.