Unsupervised Image Fusion Method based on Feature Mutual Mapping

Dongyu Rao, Xiao-Jun Wu*, Tianyang Xu, Guoyang Chen

Abstract—Deep learning-based image fusion approaches have obtained wide attention in recent years, achieving promising performance in terms of visual perception. However, the fusion module in the current deep learning-based methods suffers from two limitations, i.e., manually designed fusion function, and input-independent network learning. In this paper, we propose an unsupervised adaptive image fusion method to address the above issues. We propose a feature mutual mapping fusion module and dual-branch multi-scale autoencoder. More specifically, we construct a global map to measure the connections of pixels between the input source images. Besides, we design a dual-branch multi-scale network through sampling transformation to extract discriminative image features. We further enrich feature representations of different scales through feature aggregation in the decoding process. Finally, we propose a modified loss function to train the network with efficient convergence property. Through sufficient training on infrared and visible image data sets, our method also shows excellent generalized performance in multi-focus and medical image fusion. Our method achieves superior performance in both visual perception and objective evaluation. Experiments prove that the performance of our proposed method on a variety of image fusion tasks surpasses other state-of-the-art methods, proving the effectiveness and versatility of our approach.

Index Terms—Image fusion, multi-scale autoencoder, adaptive fusion

1 INTRODUCTION

Image fusion is an essential image processing technique that aims to fuse input images with different information into one output image containing more perceptive and salient visual clues. The image fusion process requires the use of appropriate feature extractors and fusion strategies to achieve the ideal fusion effect [1]. With effective capacity in processing multi-source visual inputs, image fusion has a wide range of applications and excellent development prospects in photo processing, visual tracking [2], [3], [4], [5] and video surveillance, etc.

Image fusion currently includes multi-sub-tasks, such as infrared and visible image fusion [6], multi-focus image fusion [7], multi-exposure image fusion [8], etc. Fusion models for different tasks have a certain versatility as they share the same underlying function of integrating multiple images. Therefore, image fusion algorithms can be simply divided into two types: traditional methods [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21] and deep learning-based methods [22], [23], [24], [25], [26].

The traditional methods are non-deep learning. According to the mathematic formulations, they are divided into the following categories:

(1) Multi-scale decomposition-based fusion: Multi-scale transformation is a widely used tool in image processing tasks. Infusion tasks, firstly, the source image is decomposed at multiple scales to obtain a multi-scale representation of the image. Then multi-scale representations from different images are fused according to specific fusion rules to obtain a fused multi-scale representation. Finally, the fused multi-scale representation is inversely transformed to obtain the fused image. Many algorithms that use multi-scale transforms have obtained promising fusion results, such as pyramid transform [27], wavelet transform [28] and contourlet transform [29], etc.

(2) Sparse representation-based fusion: Sparse representation (SR) is an image representation technique, which has been applied in many image-related tasks. It selects a sparse linear combination of atoms from the over-complete dictionary learned from high-quality images to represent the input. In the fusion task, the input image is divided into patches and then decomposed to obtain the corresponding sparse coefficients [30], [31], [32]. After decomposition, the fusion rules are applied to the coefficients of different source images to obtain the fused sparse coefficients. Finally, the fusion image is reconstructed through the fusion coefficient and dictionary.

(3) Fusion in other domains: In addition to methods based on multi-scale decomposition and sparse representation, there are other methods such as low-rank representation (LRR), subspace, etc. In particular, LRR can extract the features of the image in a specific domain (low-rank) for fusion to obtain a fused image [10]. The method based on subspace is to map the image from high-dimensional to low-dimensional, eliminating the interference of redundant information and facilitating the acquisition of the internal structure of the image [33], [34].

Although traditional methods can achieve acceptable fusion effects, the limitations of traditional methods for feature representation result in the focus on the field of deep learning [35], [36], [37]. In 2017, Liu [7] et al. proposed a
We propose a feature mutual mapping module to adaptively learn the fusion rules, formulating a unified framework for a variety of fusion tasks.

The main innovations of the proposed method include:

- We design an auto-encoder network that aggregates multi-scale representations to enhance the output features.
- We propose a feature mutual mapping module to adaptively learn the fusion rules, formulating a unified framework for a variety of fusion tasks.
- A single loss function is used during training, generating excellent fusion performance, and accelerating the training convergence.

The paper is organized as follows: In Section 2 we introduce the studies related to the proposed fusion method in this paper. In Section 3 we present the framework and training details of the proposed method. The experimental results and their analysis are reported in Section 4. Finally, the conclusions are drawn in Section 5.

2 RELATED WORKS

In recent years, many image fusion methods based on deep learning have been proposed \[44\]. Researchers have made different attempts to modify the network structure, fusion module, and loss function to obtain a fused image with better performance. In this section, firstly, we review several image fusion methods based on deep learning in recent years. Then we introduce the application of the self-attention mechanism in the computer vision field.

2.1 Deep Learning-based Image Fusion Methods

In 2017, Liu et al. \[45\] propose an image fusion method based on a convolutional neural network. In this paper, infrared and visible images are used as source images, generating the output containing visible details and infrared features. Firstly, a convolutional network is used to obtain a weight map that integrates information from two source images. According to the different characteristics of the infrared image and visible light image, multi-scale fusion is carried out through the image pyramid. Finally, the fusion image is obtained through Laplacian pyramid reconstruction. Experimental results show that the method achieves state-of-the-art results. However, this method only obtains the weight map through the convolutional neural network, without performing deep analysis.

In 2018, Ren et al. \[46\] use the VGG network to extract the features of infrared and visible source images, accompanied by a corresponding loss function to complete the fusion task. The fusion image is composed of general features extracted by the convolutional neural network in different proportions. This method has achieved good results in both subjective and objective evaluation indicators. However, due to the limitations of training methods and loss functions, this method is only suitable for infrared and visible image fusion tasks.

Li et al. \[22\] propose an autoencoder structure based on deep learning, applying different fusion strategies to the encoded image features. This method uses public data sets to train encoders and decoders, alleviating the limitation of insufficient training data for infrared and visible image fusion tasks. Since the fusion strategy does not participate in training, the autoencoder can also be used as a general framework for different tasks with different fusion strategies. However, the fusion strategy is not synchronized with the training, which means that the method requires manual selection for each fusion task.

In 2019, Ma et al. \[24\] introduce the method of the generative adversarial network into the image fusion task. This method uses a generator to generate a fusion image that retains the main infrared information and visible gradients. The discriminator is used to force the fused image to exhibit additional visible details. Specifically, the fusion image generated by GAN is clean and clear which is not sensitive to noise in the infrared image.

2.2 Self-attention Mechanism

In 2017, Lin et al. \[47\] propose the self-attention mechanism and apply it to the field of natural language processing. Subsequently, the work of Vaswani et al. \[48\] extends the
self-attention approach with widespread attention and application. Similar to the attention mechanism, the purpose of the self-attention mechanism is to integrate comprehensive information from a large number of connected nodes. The difference is that the self-attention mechanism reduces the dependence on external information, focusing on capturing the internal correlation of features. Its application in language processing is mainly to solve the problem of long-distance dependence by calculating the mutual influence between words.

In the field of computer vision, self-attention can effectively capture the long-range dependency between different positions [49]. Zhang et al. [50] use self-attention for image generation tasks and propose SAGAN. SAGAN adds a self-attention module on the basis of SNGAN [51], so that the generation of each position is no longer solely dependent on other positions adjacent to this position, while introducing a position that is farther and has a greater amount of information.

3 THE PROPOSED FUSION METHOD

In this section, we introduce our method in detail. Firstly, the overall framework of our method is presented in section 3.1. Then, we introduce the fusion method based on the image mapping relationship of different modalities, which is derived from the self-attention mechanism in section 3.2. Finally, the detail of the training phase is shown in section 3.3.

3.1 Framework of Network

As shown in Fig. 1, our network consists of three parts, including encoder, fusion module, and decoder. The red and blue rectangles on the left of Fig. 1 represent the encoders for different source images respectively. The green square represents the mapping fusion module. The other shapes denote the decoder.

In Fig. 1, “Iₓ” and “Iᵧ” represent different source images respectively. Unlike other methods [52], our network does not share the network weights of the two modalities, as images of different modalities present different information distribution. Therefore, from the perspective of feature mapping, two different feature spaces for each modality can better retain the unique information of a modal than sharing the same feature space. In order to characterize the corresponding modal information through the trained neural network, we set up two symmetrical encoder branches for infrared and visible images. Each encoder branch includes a 3 times 3 convolution and two res-blocks. In order to obtain a feature map of different scales, down-sampling is performed between each block to change the size of the feature.

As the Fig. 2 shows, the mapping fusion block is an adaptive fusion method to find the feature mapping between source images. The input of the mapping fusion module is the extracted feature of two modal images. This module achieves a better fusion effect by finding the global relationship of each spatial position between one modal image and another modal image. The specific implementation of the mapping fusion module will be described in section 3.2.

For the decoder, each block consists of a concatenation and convolution operation. In the decoding process, we aggregate features of different scales through feature aggregation [43], so that the fused image retains more multi-layer clues. Through a layer of convolution, the fused features are restored to the original shape and dimension, thus obtaining the final output.

3.2 The Mapping Fusion Module

In existing deep learning-based image fusion methods, commonly used fusion strategies include direct concatenation
and manual fusion rule design, etc. Direct concatenation of cross-modal features can obtain better fusion results, while delivering information redundancy. Manually designed fusion strategies are more flexible but difficult to find the optimal solution. Therefore, we try to find a strategy that can adaptively realize the multi-modal fusion according to the relationship between the cross-modal information in the image fusion task.

According to the analysis, we can conclude that performing direct concatenation fusion can not achieve fine-grained image information integration. Besides, the fusion process can only capture the local information constrained in the receptive field that is limited by the size of the multi-layer convolution kernels.

Therefore, we argue that finding the global relationship between different source image features is essential for generating adaptive fusion weights. The characteristics of the self-attention mechanism are employed to implement the global relationship construction. In the field of computer vision, self-attention usually determines the importance of the information in the image by calculating the relationship between each spatial point and other points on a single image [55]. In order to find the mapping relationship between the two modalities, we refer to the implementation of self-attention to involve the features of different modalities, performing a matrix multiplication operation to measure their degree of correlation. Then the result obtained by the dot product is input to a softmax operation and performs matrix multiplication with two modes separately. Through the above operations, we have obtained the global positional relationship on one modal with respect to another modal feature. We use the output map as the attention weight of the modal. As shown in Eq. 1, Map\(X\) and Map\(Y\) mean the map of different source images respectively. \(f_x\) and \(f_y\) represent the features of different source images respectively. \(fuse\) is the feature obtained by fusion.

\[
 fuse = Map_X \ast f_x + Map_Y \ast f_y.
\]  

(1)

The matrix multiplication operation requires a lot of computing resources, while the mapping relationship between the features can also well represent the mapping relationship between the two modal images, so we use the mapping fusion block for the image after two downsamplings. For the shallow features, the deep Map\(X\) and Map\(Y\) are upsampled, and then the shallow map is obtained through dense connection [54]. For the shallow fusion module, the fusion feature is obtained by implementing the map and the corresponding feature according to Eq. 1.

### 3.3 Loss function

In the training phase, our method adopts an end-to-end training method. Through the learning of neural networks, we can get a multi-branch encoder for infrared visible images, a mapping fusion module, and the corresponding decoder. The structure of the network during training is shown in Fig. 1.

We propose an approach that relies on a single loss function to achieve the fusion task. Unlike the previous method [22], [55] that uses multiple loss functions to jointly supervise the training stage, we can achieve a better fusion effect using only a single loss function. In addition, the simplification of the loss function also speeds up the training process.

The SSIM [56] is a measure of the structural similarity between two images. The calculation process of structural similarity is shown in Eq. 2, \(X\) and \(Y\) represent two input images respectively. \(\mu\) and \(\sigma\) stand for mean and standard deviation respectively. \(\sigma_{XY}\) means the covariance between \(X\) and \(Y\). \(C_1\) and \(C_2\) are stability coefficients.

\[
SSIM(X, Y) = \frac{(2\mu_X\mu_Y + C_1)(2\sigma_{XY} + C_2)}{(\mu_X^2 + \mu_Y^2 + C_1)(\sigma_X^2 + \sigma_Y^2 + C_2)}
\]  

(2)

Due to the effectiveness of SSIM, it is used as the cost function in a variety of image fusion tasks [52], [57]. Since the structural similarity calculates the overall information of the image, neglecting the local visual connections. Therefore, simply using SSIM as the loss function cannot achieve a visually acceptable fusion effect. Taking infrared and visible image fusion as an example, since there are more noises in the infrared images, we usually hope to get more details from the visible image, such that other losses and parameters are introduced to enhance this information [52]. Inspired by [55], [58], we combined SSIM with the variance of the image to obtain a modified fusion loss function.

The variance of an image represents the distribution of data pixels in an image. As shown in Eq. 3, \(M\) and \(N\) represent the image size in the horizontal and vertical directions respectively. \(\mu\) is the mean of the image.

\[
\sigma^2(X) = \frac{\sum_{i=0}^{M-1} \sum_{j=0}^{N-1} [X(i,j) - \mu]^2}{MN}
\]  

(3)

Var\_SSIM calculates the structural similarity between fused image and source image patches. \(\sigma^2\) is the basis for choosing which source image to compare. As shown in Eq. 4, \(I_X\) and \(I_Y\) represent two input images respectively. \(I_F\) is a fused image. \(W\) is the sliding window, and its size is set to 11×11. Through the sliding window, the image is divided into multiple overlapping patches. We calculate the variance and the structural similarity of each patch. Finally,
4 EXPERIMENTS AND ANALYSES

In this section, we verify our approach in different image fusion tasks. First, we show the details of training and testing implementations in Sec 4.1. Then, different evaluation criteria for image fusion tasks are shown, including subjective and objective evaluations in Sec 4.2. In Sec 4.3, we show the performance of our method on the fusion of infrared and visible images. In Sec 4.4 and Sec 4.5, the performance of our method on the fusion of multi-focus images and medical images are reported respectively.

4.1 Training and test details

In the training stage, we select 40,000 pairs of corresponding infrared and visible images from the KAIST [59] data set as training data. The training image size is set to 256 times 256 pixels. During training, we use Adam as the optimizer and the learning rate is set to a constant 0.0001. The whole training process includes 4 epochs. Our model is implemented with NVIDIA TITAN Xp and Pytorch.

Through sufficient training on infrared and visible image data sets, we believe that the model has been sufficiently learning the feature extractor and fusion rules. Due to the similar characteristics of infrared and visible light images, multi-focus images, and medical image fusion tasks, we did not perform additional training on multi-focus image fusion and medical image fusion tasks. In the test, we used 10 pairs of images in [22] as the test set of infrared and visible image tasks. Multi-focus and medical images in [60] are used as test sets for multi-focus image fusion and medical image fusion tasks, respectively.

4.2 Subjective and Objective Evaluation

When evaluating the fusion results, we use two evaluation criteria: subjective evaluation and objective evaluation.

Subjective evaluation starts from human vision to judge whether the fusion result conforms to visual habits, such
Fig. 5. Infrared and visible image fusion experiment on “human and tree” images.

TABLE 1
Quantitative evaluation results of infrared and visible image fusion tasks. The best two results are highlighted in **bold** and *italic* fonts.

| Methods    | EI      | CE     | SF     | EN     | Q_{\alpha/\beta} | MS. SSIM | SD     | VIF     |
|------------|---------|--------|--------|--------|------------------|----------|--------|---------|
| RP         | 44.9054 | 1.3420 | 12.7249| 6.5397 | 0.4341           | 0.8404   | 63.2427| 0.6420  |
| Wavelet    | 24.6654 | 1.4278 | 6.2567 | 6.2454 | 0.3214           | 0.8598   | 52.2292| 0.2921  |
| DTCWT      | 42.4889 | 1.6235 | 11.1296| 6.4791 | 0.5258           | 0.9053   | 60.1138| 0.5986  |
| CVT        | 42.9631 | 1.5894 | 11.1129| 6.4989 | 0.4936           | 0.8963   | 60.4005| 0.5930  |
| MSVD       | 27.6098 | 1.4202 | 8.5538 | 6.2807 | 0.3328           | 0.8652   | 52.9853| 0.3031  |
| GTF        | 35.0073 | 1.1440 | 9.5022 | 6.5781 | 0.4400           | 0.8169   | 66.0773| 0.4071  |
| DenseFuse  | 36.4838 | 1.4015 | 9.3238 | 6.8526 | 0.4735           | 0.8692   | 81.7283| 0.6875  |
| DeepFuse   | 33.8768 | 1.7779 | 8.3500 | 6.6102 | 0.3847           | 0.9138   | 66.8872| 0.5752  |
| IFCNN      | 44.9725 | 1.4413 | 11.8590| 6.6454 | 0.4962           | 0.9129   | 73.7083| 0.6090  |
| FusionGAN  | 32.5997 | 1.9353 | 8.0476 | 6.5409 | 0.2662           | 0.6135   | 61.6339| 0.4928  |
| NestFuse   | 37.5627 | 1.4463 | 9.5383 | 6.6492 | 0.4979           | 0.8816   | 80.2687| 0.7018  |
| U2Fusion   | 48.4915 | 1.3255 | 11.0368| 6.7227 | 0.3934           | 0.8774   | 66.5035| 0.7680  |
| Ours       | 45.5980 | 1.2941 | 11.5680| 6.8435 | 0.5581           | 0.9164   | 93.0976| 0.7897  |

as clarity, saliency information, etc. Therefore, the subjective evaluation method is to put the results of different methods together and directly display them.

4.3 Visible and Infrared Image Fusion

In Fig. 3 we show some images in the infrared and visible image fusion task. The upper row is the infrared image, the lower row is the corresponding visible image. The goal of infrared and visible image fusion is to combine the saliency information in the infrared image (the part of the highlighted object) with the detailed information of the visible image to obtain a fused image with better visual effects. In the testing phase, we selected 10 pairs of infrared and visible images to be fused and evaluated by objective methods.

In Fig. 3 we show a comparison of the visualization results of the involved 13 methods. In the red frame, the infrared source image has a bright human body, and the
Fig. 6. The multi-focus image fusion experiment results of "girl and toy" images.

Fig. 7. The multi-focus image fusion experiment results of "girl and toy" images.

TABLE 2
Quantitative evaluation results of multi-focus image fusion tasks. The best three results are highlighted in red, blue and brown fonts.

| Methods    | EI   | SF    | EN    | SCD  | MI   | SD    | VIF   |
|------------|------|-------|-------|------|------|-------|-------|
| ASR        | 69.0392 | 18.9739 | 7.5281 | 0.4839 | 15.0561 | 115.8590 | 1.0422 |
| CBF        | 68.5800 | 18,5826 | 7.5278 | 0.4771 | 15.0555 | 115.8917 | 1.0668 |
| CSR        | 69.8621 | 19.2245 | 7.5313 | 0.5292 | 15.0625 | 117.1057 | 1.1273 |
| DRPL       | 70.9735 | 19.4398 | 7.5334 | 0.5446 | 15.0667 | 117.3645 | 1.1389 |
| RP SP      | 70.7294 | 19.4944 | 7.5389 | 0.5645 | 15.0778 | 117.4334 | 1.1407 |
| MGFF       | 63.1033 | 17.3024 | 7.5685 | 0.7480 | 15.1370 | 120.3308 | 1.2356 |
| IFCNN(mean) | 65.5623 | 17.7762 | 7.5120 | 0.5253 | 15.0241 | 115.3850 | 1.0541 |
| IFCNN(max) | 70.7663 | 19.4004 | 7.5319 | 0.5817 | 15.0639 | 117.3573 | 1.1339 |
| IFCNN(sum) | 69.2496 | 18.9037 | 7.5334 | 0.5649 | 15.0667 | 116.6219 | 1.0979 |
| U2Fusion   | 69.2651 | 17.2710 | 7.5290 | 0.9874 | 15.0581 | 123.8306 | 1.2799 |
| Ours       | 77.1693 | 21.3001 | 7.5381 | 1.1038 | 15.0763 | 135.1799 | 1.4193 |

visible image has the details of the door frame. Other methods can only retain one modality of the contents, and the other modality is weakened by sufficient information fusion. The fused image obtained by our method can retain the highlights of the human body and the details of the door frame at the same time.

In Fig. 5, the bright human body in the infrared source image (red box) and the edge of the leaf in the visible image (blue box) are the fused objects. In other methods, the fusion of infrared and visible images by leaves is insufficient. At the same time, the saliency information in the fused image is not obvious, because the human body lacks visual clues in the corresponding area of the visible image. In our method, the feature map fusion module has learned the global relationship corresponding to the two source images. Accordingly, the red frame part retains the human body information highlighted in the infrared image, and the blue frame part blends the edges of the leaves and the subject in the two source images well.

In terms of objective evaluation, we compare with 12 methods, including classic and latest methods: Ratio of Low-pass Pyramid (RP) [67], Wavelet [68], Dual-Tree Complex Wavelet Transform (DTCWT) [69], Curvelet Transform (CVT) [70], Multi-resolution Singular Value Decomposition
Fig. 8. The medical image fusion experiment results of CT and MR images.

TABLE 3
Quantitative evaluation results of medical image fusion tasks. The best three results are highlighted in red, blue and brown fonts.

| Methods       | EI   | SF   | EN   | Q<sub>ab/f</sub> | SCD  | MS_SSIM | MI   | SD   | VIF  |
|---------------|------|------|------|-----------------|------|---------|------|------|------|
| CBF           | 82.1394 | 22.5670 | 5.7775 | 0.5500         | 0.8316 | 0.8520 | 11.5550 | 61.6619 | 0.8147 |
| CVT           | 79.0579 | 22.3262 | 6.0096 | 0.5229         | 0.7990 | 0.8859 | 12.0191 | 58.5920 | 0.8511 |
| DTCWT         | 77.6919 | 22.4307 | 5.8597 | 0.5559         | 0.8036 | 0.8994 | 11.7194 | 58.8025 | 0.8539 |
| GTF           | 55.8531 | 16.3053 | 5.2286 | 0.4476         | 0.4973 | 0.7481 | 10.4572 | 50.7192 | 0.4657 |
| MSVD          | 54.9092 | 19.9155 | 5.5083 | 0.4206         | 0.7683 | 0.8704 | 11.0166 | 53.8837 | 0.5566 |
| IFCNN(max)    | 83.1070 | 24.1949 | 5.6613 | 0.5737         | 1.1406 | 0.9329 | 11.3226 | 67.3829 | 0.9767 |
| IFCNN(mean)   | 76.0946 | 22.6930 | 5.5177 | 0.5466         | 0.9077 | 0.9197 | 11.0354 | 58.2606 | 0.8062 |
| IFCNN(sum)    | 76.8979 | 22.9619 | 5.5036 | 0.5416         | 0.8822 | 0.9165 | 11.0073 | 57.8363 | 0.8070 |
| U2Fusion      | 69.4348 | 20.1861 | 5.4758 | 0.4687         | 0.7430 | 0.8967 | 10.9516 | 54.0862 | 0.7678 |
| Ours          | 88.2840 | 25.7036 | 6.0503 | 0.6289         | 1.4161 | 0.9231 | 12.1005 | 77.8152 | 1.1499 |

Fig. 9. Four pairs of source images. The top row contains CT images, and the second row contains MR images.

(MSVD) [71], gradient transfer and total variation minimization (GTF) [72], DenseFuse [22], DeepFuse [73], a general end-to-end fusion network (IFCNN) [60], NestFuse [52], FusionGAN [24], U2Fusion [74]. In the quantitative comparison, we used the seven evaluation methods mentioned above. All quantitative comparison results are shown in Table 3. Our method achieved the best results on the four evaluation indicators (Q<sub>ab/f</sub>, MS_SSIM, SD, VIF), and achieved the second and third results on the remaining indicators (second: EI, CE; third: SF, EN).

4.4 Multi-focus Image Fusion

In Fig. 7 we present some examples of multi-focus images. The first row of image content is focused on the foreground, the second row of image content is focused on the background. Therefore, the purpose of multi-focus image fusion is to fuse two images focused on different positions to obtain a fused image with clear content. In the test, we used twenty pairs of multi-focus images to fuse.

We show the comparison of the visual fusion results of ten methods, including classic and latest methods: ASR [75], CBF [76], CSR [77], DRPL [78], RP_SP [79], MGFF [80], IFCNN [60] and U2Fusion [74]. In terms of visual effects, various methods can achieve better fusion results. But our method is not trained on the multi-focus image data set, and it can achieve comparable results with other methods.
### Table 4
The objective evaluation with different networks. The best results are highlighted in **bold** fonts.

|      | EI   | CE   | SF   | EN   | Q_{ab/f} | MS_SSIM | SD   | VIF   |
|------|------|------|------|------|----------|---------|------|-------|
| VIF-Net | 34.7517 | 1.3231 | 8.2169 | 6.7120 | 0.4725 | 0.8568 | 85.6736 | 0.5236 |
| Ours-Net | **38.2261** | **1.3859** | **9.9340** | **6.6467** | **0.4939** | 0.8042 | **86.1420** | **0.6493** |

### Table 5
The objective evaluation with different loss functions. The best results are highlighted in **bold** fonts.

|      | EI   | CE   | SF   | EN   | Q_{ab/f} | MS_SSIM | SD   | VIF   |
|------|------|------|------|------|----------|---------|------|-------|
| Ours(Mean) | 38.2261 | 1.3859 | 9.9340 | 6.6467 | 0.4939 | 0.8042 | **86.1420** | 0.6493 |
| Ours(Var) | **44.5137** | 1.1959 | 11.3889 | 6.8980 | 0.5436 | 0.9235 | **89.8534** | 0.7556 |

This shows that our fusion module and network structure have been fully trained in a single data set to learn quite transferable feature extraction capabilities as well as the fusion rules.

In the quantitative comparison, we use the seven evaluation methods mentioned above. All quantitative comparison results are shown in Table 2. Our method achieved the best results on the five evaluation indicators (EI, SF, SCD, SD, VIF), and achieved the second results on the remaining indicators (EN, MI).

### 4.5 Medical Image Fusion

In Fig. 9, we visualize some medical images that need to be fused, including CT images and MR images. The purpose of medical image fusion is similar to the task of infrared and visible image fusion. It combines the salient information in the two source images to generate a fused image to assist visual judgment. This has a significant effect on improving the results of medical diagnosis. In the test, we use eight pairs of medical images (CT and MR) to fuse.

In Fig. 8, we show the fusion results of CT and MR images obtained by ten methods, including traditional methods and deep learning-based methods: CBF [76], CVT [70], DTCWT [69], GTF [72], IFCNN [60], MSVD [71], U2Fusion [74] and ours. Different features are shown in the red boxes of the two source images. We hope to obtain these two different kinds of information at the same time in the fused image. Among the fusion results of ten methods, some methods did not preserve both information well (f, j). Others use fusion to reduce the saliency of the information in the original image. Our method guarantees the salient information of the two source images and achieves a better fusion effect.

In Table 3, we use nine objective evaluation methods to compare the fusion results. The red font ratio indicates the best result. Our method performs the best in eight indicators (EI, SF, EN, Q_{ab/f}, SCD, MI, SD, VIF) and ranks second in the remaining one metric (MS_SSIM).

### 4.6 Ablation Study

#### 4.6.1 Network structure

In order to verify the effectiveness of our network structure, the VIF-Net [55] is used to compare to our method, which uses a similar loss function and training method. The framework of VIF-Net is shown in Fig. 10. In order to compare with VIF-Net, we keep the loss function consistent with it, so that our method is only different from VIF-Net in the network structure. At the same time, we chose the same cascade fusion for the fusion module. As shown in Table 4, our method has achieved advantages in multiple objective evaluations, which also shows the effectiveness of the network structure. The network structure we proposed obtains the multi-scale information of the image through scale transformation. Besides, multi-scale information is enriched through feature aggregation, so better performance is obtained.

#### 4.6.2 Loss function

In terms of verifying the effect of the network structure, we use the same loss function as VIF-Net which uses the average of the pixel values of each sliding window as the criterion for selecting the patch of source image. Then we improve the loss function, using variance as the selection criterion for each sliding window. While keeping the network structure and fusion module unchanged, we compare the impact of the results produced by different loss functions. As Table 5 shows, the result of our proposed loss function is ahead of the original loss in all aspects. Compared with the mean, variance can characterize the distribution of information, which makes the model have better performance and a wider application range.

#### 4.6.3 Fusion method

In addition to the network structure and loss function, the design of the fusion method is another core of the image fusion algorithm. After adopting our network structure...
and loss function, we compared three methods, including direct addition, direct cascade, and dense feature mutual mapping. The results are shown in Table 6. The best fusion image performance can be obtained by using the feature mutual mapping module we proposed for fusion. As the simplest method, direct addition has achieved good results in some algorithms. However, it may cause the problem of neutralization of the pixel value of the corresponding position, which makes the poor quality of the fused image. The concatenation fusion method allows the model to actively learn the fusion weights during the training process, but the concatenation is too simple to learn the fusion method we need well. The feature mutual mapping module we proposed learns a satisfactory fusion method in training through more complex structures, and can be directly transferred to more fusion tasks.

4.6.4 Network depth

After determining that our network structure and loss function are effective, we continue to explore the impact of the depth of the network on the fusion performance. As shown in Fig.4, our encoder network is composed of a Conv layer and two res-blocks which we define its depth as three layers. The initial network depth is determined by our fusion method, so we try to prove whether a deeper network can deliver better results.

Every time we deepen the network, we add downsampling and res at the encoder, and the decoder part will also increase the corresponding aggregation structure. In Table 7, we show the results of two experiments after deepening the network, which are 4-layers and 5-layers. The results show that deepening the network does not significantly boost fusion performance. This also shows that after the network depth reaches a certain level, continuing to deepen the network will not have a positive effect on the performance improvement, and may even cause a negative effect.

5 Conclusion

We proposed a novel image fusion method based on feature mutual mapping and multi-scale autoencoder. In the feature mutual mapping and fusion module, the network can adaptively learn the fusion strategy during the training process by finding the mapping relationship between the features of the two source images. The multi-scale autoencoder and the aggregation operation of different scale features in the decoding process make the fusion fully performed at different granularities. Experiments on a variety of tasks and their public data sets prove the effectiveness and versatility of our method.

References

[1] S. Li, X. Kang, L. Fang, J. Hu, and H. Yin, “Pixel-level image fusion: A survey of the state of the art,” Information Fusion, vol. 33, pp. 100–112, 2017.
[2] H. Li, X.-J. Wu, and J. Kittler, “Mdlatlrr: A novel decomposition method for infrared and visible image fusion,” IEEE Transactions on Image Processing, vol. 29, pp. 4733–4746, 2020.
[3] T. Xu, Z.-H. Feng, X.-J. Wu, and J. Kittler, “Learning low-rank and sparse discriminative correlation filters for coarse-to-fine visual object tracking,” IEEE Transactions on Circuits and Systems for Video Technology, vol. 30, no. 10, pp. 3727–3739, 2019.
[4] T. Xu, Z. Feng, X.-J. Wu, and J. Kittler, “Adaptive channel selection for robust visual object tracking with discriminative correlation filters,” International Journal of Computer Vision, vol. 129, no. 5, pp. 1359–1375, 2021.
[5] T. Xu, Z.-H. Feng, X.-J. Wu, and J. Kittler, “An accelerated correlation filter tracker,” Pattern Recognition, vol. 102, p. 107172, 2020.
[6] X. Jin, Q. Jiang, S. Yao, D. Zhou, R. Nie, J. Hai, and K. He, “A survey of infrared and visual image fusion methods,” Infrared Physics & Technology, vol. 85, pp. 478–501, 2017.
[7] Y. Liu, X. Chen, H. Peng, and Z. Wang, “Multi-focus image fusion with a deep convolutional neural network,” Information Fusion, vol. 36, pp. 191–207, 2017.
[8] H. Xu, J. Ma, and X.-P. Zhang, “Mef-gan: multi-exposure image fusion via generative adversarial networks,” IEEE Transactions on Image Processing, vol. 29, pp. 7203–7216, 2020.
[9] H. Li, Y. Wang, Z. Yang, R. Wang, X. Li, and D. Tao, “Discriminative dictionary learning-based multiple component decomposition for detail-preserving noisy image fusion,” IEEE Transactions on Instrumentation and Measurement, vol. 69, no. 4, pp. 1082–1102, 2019.
[10] H. Li and X.-J. Wu, “Multi-focus image fusion using dictionary learning and low-rank representation,” in International Conference on Image and Graphics. Springer, 2017, pp. 675–686.
[11] G. Pajares and J. M. De La Cruz, “A wavelet-based image fusion tutorial,” Pattern recognition, vol. 37, no. 9, pp. 1855–1872, 2004.
[12] X. Li, X. Guo, P. Han, X. Wang, H. Li, and T. Luo, “Laplacian redecomposition for multimodal medical image fusion,” IEEE Transactions on Instrumentation and Measurement, vol. 69, no. 9, pp. 6880–6890, 2020.
[13] Z. Zhou, B. Wang, S. Li, and M. Dong, “Perceptual fusion of infrared and visible images through a hybrid multi-scale decomposition with gaussian and bilateral filters,” Information Fusion, vol. 30, pp. 15–26, 2016.
[60] Y. Zhang, Y. Liu, P. Sun, H. Yan, X. Zhao, and L. Zhang, “Ifcnn: A general image fusion framework based on convolutional neural network,” Information Fusion, vol. 54, pp. 99–118, 2020.

[61] C. Xydeas, and V. Petrovic, “Objective image fusion performance measure,” Electronics letters, vol. 36, no. 4, pp. 308–309, 2000.

[62] B. S. Kumar, “Multifocus and multispectral image fusion based on pixel significance using discrete cosine harmonic wavelet transform,” Signal, Image and Video Processing, vol. 7, no. 6, pp. 1125–1143, 2013.

[63] A. M. Eskicioglu and P. S. Fisher, “Image quality measures and their performance,” IEEE Transactions on communications, vol. 43, no. 12, pp. 2959–2965, 1995.

[64] J. W. Roberts, J. A. Van Aardt, and F. B. Ahmed, “Assessment of image fusion procedures using entropy, image quality, and multispectral classification,” Journal of Applied Remote Sensing, vol. 2, no. 1, p. 023522, 2008.

[65] Y.-J. Rao, “In-fibre bragg grating sensors,” Measurement science and technology, vol. 8, no. 4, p. 355, 1997.

[66] H. R. Sheikh and A. C. Bovik, “Image information and visual quality,” IEEE Transactions on image processing, vol. 15, no. 2, pp. 430–444, 2006.

[67] A. Toet, “Image fusion by a ratio of low-pass pyramid,” Pattern Recognition Letters, vol. 9, no. 4, pp. 245–253, 1989.

[68] J. J. Chipman, T. M. Orr, and L. N. Graham, “Wavelets and image fusion,” in Proceedings. International Conference on Image Processing, vol. 3. IEEE, 1995, pp. 248–251.

[69] J. J. Lewis, R. J. O’Callaghan, S. G. Nikolov, D. R. Bull, and N. Canagarajah, “Pixel-and region-based image fusion with complex wavelets,” Information fusion, vol. 8, no. 2, pp. 119–130, 2007.

[70] F. Nencini, A. Garzelli, S. Baronti, and L. Alparone, “Remote sensing image fusion using the curvelet transform,” Information fusion, vol. 8, no. 2, pp. 143–156, 2007.

[71] V. Naidu, “Image fusion technique using multi-resolution singular value decomposition,” Defence Science Journal, vol. 61, no. 5, p. 479, 2011.

[72] J. Ma, C. Chen, C. Li, and J. Huang, “Infrared and visible image fusion via gradient transfer and total variation minimization,” Information Fusion, vol. 31, pp. 100–109, 2016.

[73] K. R. Prabhatkar, V. S. Srikan, and R. V. Babu, “Deepfuse: A deep unsupervised approach for exposure fusion with extreme exposure image pairs,” in ICCV, vol. 1, no. 2, 2017, p. 3.

[74] H. Xu, J. Ma, J. Jiang, X. Guo, and H. Ling, “L2fusion: A unified unsupervised image fusion network,” IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020.

[75] Y. Liu and Z. Wang, “Simultaneous image fusion and denoising with adaptive sparse representation,” IET Image Processing, vol. 9, no. 5, pp. 347–357, 2014.

[76] B. S. Kumar, “Image fusion based on pixel significance using cross bilateral filter,” Signal, image and video processing, vol. 9, no. 5, pp. 1193–1204, 2015.

[77] Y. Liu, X. Chen, R. K. Ward, and Z. J. Wang, “Image fusion with convolutional sparse representation,” IEEE signal processing letters, vol. 23, no. 12, pp. 1882–1886, 2016.

[78] J. Li, X. Guo, G. Lu, B. Zhang, Y. Xu, F. Wu, and D. Zhang, “Drpl: Deep regression pair learning for multi-focus image fusion,” IEEE Transactions on Image Processing, vol. 29, pp. 4816–4831, 2020.

[79] Y. Liu, S. Liu, and Z. Wang, “A general framework for image fusion based on multi-scale transform and sparse representation,” Information fusion, vol. 24, pp. 147–164, 2015.

[80] D. P. Bavirisetti, G. Xiao, J. Zhao, R. Dhuli, and G. Liu, “Multi-scale guided image and video fusion: A fast and efficient approach,” Circuits, Systems, and Signal Processing, vol. 38, no. 12, pp. 5576–5605, 2019.