IJLAI Transactions on Science and Engineering (ISSN: 3006-3418) 2024, vol. 2, no. 3, pp: 81–90

Key Technologies of Image Fusion Based on Deep Learning: A Survey

Lijuan Feng[1] and Jiangjiang Li[1]

School of Electronics and Electrical Engineering, Zhengzhou University of Science and Technology
Zhengzhou 450064 China
857003841@qq.com
Corresponding author: Jiangjiang Li

Abstract. Image fusion is an important branch of image processing. Because of the remarkable advantages of neural network in image feature extraction and classification, the application of neural network technology in the field of image fusion is also a research hotspot in recent years. Firstly, the infrared and visible light image fusion algorithms based on shallow and deep neural networks are summarized, and the research progress of image fusion technology is introduced in detail, and the research results of fusion algorithms are presented. Finally, the challenges faced by image fusion are discussed, and the future development direction of this field is forecasted.

Keywords: Image fusion, Deep learning, Deep neural networks.

1. Introduction

The purpose of image fusion is to obtain a fusion image containing rich details of the source image through computer processing of the source image captured by different types of sensors, which is easier for human visual sensory system to observe [1]. In addition, compared with a single source image, the fusion image can obtain the scene information of the target more clearly, which significantly improves the quality and clarity of the image. Therefore, the key to distinguish the advantages and disadvantages of a fusion algorithm lies in whether the algorithm can effectively extract the details and features of the source image, while avoiding the introduction of new noise2 in the fusion process. With the continuous progress of information technology, the role of sensors has become increasingly prominent, and the information obtained by the same type of sensors can no longer meet the needs of People's Daily life, so different types of sensors are needed to obtain more comprehensive information. Due to the huge differences in imaging principle, spatial resolution and texture features of images obtained by different types of sensors, image fusion technology is increasingly showing its importance in related fields [2-4].

Image fusion technology can be fused for a variety of different types of source images, such as infrared and visible images, medical images, remote sensing images, etc., their signals come from different ways, thus providing scene information from different aspects. In recent years, neural network technology has made good achievements in the field of computer vision and image processing, solving problems such as image classification [5,6], segmentation [7], super resolution [8] and so on. Due to the outstanding advantages of neural networks in extracting image features, different neural network algorithms are applied in the field of image fusion, and these algorithms can be divided into the following categories: Fusion methods based on pulse coupled neural networks, convolutional sparse representation, convolutional neural networks, generative adversarial networks and other framework fusion methods.

2. Image Fusion Classification

Image fusion technology was first proposed for image generation task [9,10], and infrared visible light fusion is an important branch of it. According to the fusion level, infrared visible fusion can be divided into pixel-level fusion, feature-level fusion and decision-level fusion.

Pixel level fusion. The most basic image fusion algorithm directly adopts a certain fusion strategy in the spatial domain to fuse pixels one by one and finally perform decision analysis. This hierarchical fusion algorithm is simple in design and can retain most of the spatial information of the image, but because it needs to be processed pixel by pixel, the fusion time is increased, and the advanced feature information of the image cannot be extracted. Because of the instability of the algorithm, the shortcomings of the source image are easily superimposed on each other, which is easy to be disturbed by image noise. The flow chart of the spatial domain is shown in Figure 1.

Feature level image fusion. The higher-level image fusion algorithm first extracts the features of the image into the feature space, and adopts a certain fusion strategy to fuse it, and finally reconstructs the image. Feature-level

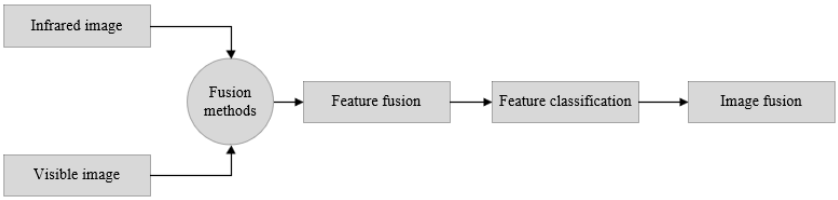

Fig. 1. Pixel-level image fusion

image fusion fuses images from the feature level to extract high-level semantic information of images. Because it extracts the image to the feature space for dimensionality reduction, it speeds up the processing speed of the fusion process and has high timeliness. However, it also has its limitations. Depending on the design of feature space, it can easily lead to the loss of image detail texture information.

Decision level fusion requires the decision classification of image features before fusion. The result of integrated decision of all source images is more accurate than that of a single image. However, since the decision of all source images will be transmitted to the decision level, the error will be correspondingly transmitted to the fused image, which increases the probability of error risk. The hierarchical fusion is usually targeted at specific tasks, and it is not generalized, so there are few researches on it.

Now it is generally believed that the fusion of visible and infrared images needs to meet the following conditions [11,12]:

1. Retain significant texture details in visible light images.
2. Highlight the intensity information of prominent targets in infrared images.
3. Reduce the generation of noise information, which is conducive to the application of advanced visual tasks.
4. In line with human visual perception, convenient for subjective analysis of images.

3. Traditional Image Fusion Method

In the early stage of the research, most of the methods to realize IVIF are mathematical transformation of the image, manual analysis in the spatial domain or transformation domain, and design the corresponding fusion rules to carry out image fusion. These methods include spatial domain based method, transform domain based method, sparse representation based method and multi-scale transform based method. The traditional method of visible and infrared image fusion usually consists of the following three steps: image transformation, image fusion and image inverse transformation.

3.1. Spatial Domain-based Approach

The image fusion method based on pixel domain fuses the source image in pixel domain by weighted fusion strategy to obtain the fused image [13,14]. The fusion strategies commonly used in this method include max-min fusion strategy, weighted average fusion strategy and other fusion strategies.

The max-min strategy obtains the maximum and minimum value of the pixel corresponding to the source image as the value of the corresponding pixel of the fused image. This method can obtain the fusion image quickly, and can retain more intensity information of the source image, so that the image has a high signal-to-noise ratio, but it is easy to lose a lot of detail texture information.

The weighted average strategy adds the pixel values corresponding to the source image and takes the average value as the corresponding pixel points. The fusion speed of this method is fast, but the fusion image contrast is not high, and the image edge information is easily lost.

Other fusion strategies need to be designed according to the needs of fusion images. For example, in order to reduce the contrast of images, the image is first transformed logarithmically and then weighted average, or the image is first transformed exponentially and then max-min strategy is adopted to highlight the significant information of images. These strategies are generally not generic and will not work for most images. The importance of image fusion algorithm based on spatial domain to pixels cannot be measured globally, and it is difficult to express the semantic information and context information of images. Therefore, this method has great limitations and can only be applied to some relatively simple visible and infrared image fusion.

3.2. Transform Domain-based Approach

In the image fusion problem, the method based on transformation domain is a commonly used method, which transforms the visible and infrared images into another space to obtain various components of the two images in the other space. These components can better distinguish the similarities and differences between the infrared and visible images, and adopt appropriate strategies to fuse each component. Then the fusion image is obtained by corresponding inverse transformation. The classical methods based on transform domain include Principal Component Analysis (PCA), Independent Component Analysis (ICA), IHS transform and so on [15,16].

Haribabu et al. [17] used PCA method to redistribute source image information, and used different weight factors to optimize model constraints on intensity and gradient information to reflect better visual effects and objective indicators. He et al. [18] believed that image decomposition features contained not only primary features, but also secondary features, and fusion of secondary features would lead to redundancy of fused image information and reduced visual effect. Therefore, they proposed an ICA-based method. The kurtosis information of ICA coefficient is used to distinguish the main and secondary features of the image, and the main features are fused and the secondary features are discarded to improve the visual perception effect of the fused image. In order to extract specific features from infrared images and visible images, Mishra et al. [19] proposed an algorithm based on IHS transform and regional variance matching, which generated high-frequency information fusion of 3×3 window pairs with a threshold of 0.5. Using the weighted average strategy as the low-frequency information fusion rule, the results show that it can enhance the edge contrast and produce more obvious texture resolution.

Although the method based on transform domain can make the feature difference of the image obvious by transforming space, the spatial feature, semantic feature and visual feature of the image are different, so it is difficult to distinguish all features by transforming a single domain, which easily leads to the loss of relevant information or information redundancy in the fused image.

3.3. Sparse Representation-based Approach

The sparse representation method simulates the sparse coding mechanism of human visual system, and uses the sparse representation of images on overcomplete dictionaries to fuse images. In this method, sparse representation coefficient is used to represent the image to be fused, and then fusion is performed according to a certain fusion strategy, and finally the fusion image is obtained by inverse transformation. Because of the sparsity of the weight coefficient, only a few coefficients can represent the significance information of the image.

Li et al. [20] firstly applied sparse representation to image fusion. After the image was represented on an overcomplete dictionary, they used the maximum strategy to fuse the sparse coefficients, which effectively solved the problems of image decomposition, fusion and restoration. In order to highlight more detail texture information as much as possible while retaining input image information, Li et al. [21] proposed a sparse representation based on visual salience and a detail injection model DIM. They designed a subfusion rule based on visual salience to represent images sparsely, and obtained high-quality fusion images by fusing the salient layer and the base layer respectively. Li et al. [22] proposed a fusion framework based on guided filtering and sparse representation. They proposed to calculate the weight graph by iterating the visual salience graph, and then optimize the weight graph by guided filtering. In order to ensure the salience of objects in fused images, Guo et al. [23] proposed a method based on sparse representation and prior texture saliency detection, which improved the potential low-rank representation to improve the clarity of texture details in fused images and obtain better visual quality.

Although the method based on sparse representation has achieved more research results and can achieve good results in supervised and infrared image fusion tasks, it still faces two most important problems, one is how to select sparse coding methods to obtain sparse coefficients, and the other is the construction of over-complete dictionaries, which limit the application scenarios of this method.

3.4. Multi-scale Transformation-based Approach

In image fusion, the method based on multi-scale transformation usually performs multi-scale decomposition of two source images, and then fuses the high frequency and low frequency sub-bands on each scale decomposition layer according to different fusion rules, and finally performs multi-scale inverse transformation to generate the fusion image. Common multi-scale transformation methods include fusion based on pyramid transform, fusion based on wavelet transform and multi-scale geometric decomposition fusion.

(1) Image fusion based on pyramid transformation.

Luo et al. [24] first proposed pyramid transformation to achieve human stereoscopic vision fusion. They used Laplacian pyramid for multi-scale decomposition and adopted the maximum fusion strategy to fuse sub-bands at different scales, but the image contrast generated by this algorithm was not high and it was easy to generate noise.

Fang et al. [25] proposed a contrast pyramid, which took into account the difference between successive layers of the Laplacian pyramid and obtained a ratio factor for each layer while carrying out a positive transformation. Compared with the Laplacian pyramid, this algorithm improved the contrast of the fused image and increased the noise resistance of the algorithm, but the algorithm complexity and running time also increased correspondingly. Yao et al. [26] added the median filter to the fusion algorithm based on pyramid transformation, which made the algorithm more robust and stable, and had stronger anti-noise ability. However, the contrast of the algorithm was not high, and human visual perception was not high.

Some researchers improve fusion performance by focusing on the salience information of images. Sun et al. [27] proposed an image fusion network PCANet based on principal component analysis and pyramid transformation. They used principal component analysis to transform images into low-dimensional space, and then fused them by weighted average fusion rule after pyramid transformation in this space. The network aggregates the features of the salient targets of infrared images and the detailed texture features of visible images, which improves the quality of image fusion, but also increases the time cost. Wang et al. [28] proposed a contrast pyramid algorithm based on regional energy, which firstly decomposed source data through contrast pyramid transformation. Then the energy, standard deviation and similarity of each region are calculated. Then the regional fusion operator is determined by threshold and standard deviation. Finally, the fusion image is reconstructed by contrast pyramid inversion, which preserves the intensity information of the source image well, but easily loses the edge information.

(2) Image fusion based on wavelet transform.

Wavelet transform represents the image by extracting different frequency components of the image, and these frequency components have orthogonality. The fusion image can be obtained by fusing these different frequency components and carrying out inverse wavelet transform. Because of the orthogonality of each frequency component, the fused image has less redundant information and retains more detail.

Wavelet transform was first proposed by Aghamaleki et al. [29] to be used in the field of image processing. Due to its advantages of orthogonality and direction selectivity, wavelet transform has received wide attention, and has been applied in various fields such as image decomposition and image fusion. According to different fusion rules, Dogra et al. [30] used multiple operators to construct the wavelet coefficients of the fusion images, which can highlight important frequency components. Compared with other fusion rules, it has better fusion performance, which not only avoids information loss, but also improves the resolution and quality of the images. Nagaraja et al. [31] proposed a wavelet transform based on spectral images. In this method, a weighted average method based on bilateral filtering was selected, and the spatial consistency of natural images was utilized to merge high-frequency and low-frequency subbands, further preserving the details of images from different sources. In order to improve the visual perception of fused images, Avci et al. [32] proposed an image fusion method based on wavelet transform and directional contrast in their work. In this method, the source image is first transformed by wavelet to obtain a multi-resolution architecture, and then the corresponding sub-band signal of each input image is selected for fusion according to the directional contrast. Finally, the fusion image is obtained by inverse wavelet transformation. This method better preserves the details of the original image and is more in line with human visual perception.

Although wavelet transform has the advantages of fast computation speed and less redundant information, it is more sensitive to noise. For images with more noise, noise suppression should be carried out in advance.

(3) Multi-scale geometric decomposition fusion.

Multi-scale geometric decomposition is widely used because of its directivity and anisotropy. Multi-scale geometric decomposition can be divided into Ridgelet, Curvelet, Bandelet, Contourlet, etc.

Gong et al. [33] proposed a scene enhancement method to solve the problem that image fusion effect is affected by visible light illuminance and weather. In this method, the source image is decomposed by curvilinear wave transform, the low frequency part is fused by improved sparse representation, and the high frequency part is fused by parametric adaptive pulse coupling. Finally, the fused image is obtained by curvilinear wave inversion. This method has good performance in detail processing, edge protection and so on.

In order to extract important information from medical images, Hao et al. [34] proposed a directional contrast based multimodal medical image fusion in the Nonsubsampled Contourlet (NSCT) domain. Firstly, the source image is converted by NSCT. Then a fusion rule based on phase consistency and directional contrast is used to fuse the low frequency coefficients and high frequency coefficients. Finally, the fusion image is constructed by using inverse NSCT transformation of all components. The fusion framework proposed by them provides an effective method for accurate analysis of multimodal images.

Zhou et al. [35] proposed an image fusion technology based on NSST (non-subsampled shearlet transform), which had flexible orientation features and optimal shift invariance, and had better fusion performance and lower computing cost compared with NSCT. They proposed a new rule for the fusion of low frequency and high frequency subband coefficients of source images. The fusion method firstly used NSST to decompose the source image

into different frequency components, then used the regional average energy model to fuse the low frequency subband coefficients of visible and infrared images, used the local direction contrast model to fuse the corresponding high frequency subband coefficients, and finally obtained the final fusion image by NSST inverse transformation. The experimental results showed that it achieved better subjective effect and higher objective evaluation index.

Compared with the fusion methods based on pyramid transform and wavelet transform, multi-scale geometric decomposition fusion has better source image decomposition ability and better analysis ability for high-dimensional data. However, due to the increase of algorithm complexity, time cost will also increase correspondingly.

4. Deep Learning-based Image Fusion Methods

4.1. Convolutional Neural Network Method

Convolutional Neural Network (CNN) is a typical neural network model, which has hierarchical feature representation mechanisms of image data at different levels of abstraction, and each stage is composed of many feature maps. The coefficients in the feature map are called neurons. Different from the traditional multi-layer perceptual neural network, the neurons of two adjacent stages in CNN are locally connected through convolution operation and weight sharing strategy, which can greatly reduce the parameters of network learning [36]. Because CNN is widely used in image recognition, object detection and other fields and has achieved good results, so try to apply CNN in the field of image fusion. Li Yu et al. [37] applied convolutional neural network to multi-focus image fusion for the first time, solving the problem that traditional image fusion methods needed to manually design complex active level measurement and fusion rules, but this method only achieved good results in multi-focus image fusion. Literature [38,39] combined CNN model with image pyramid to apply infrared and visible image fusion and medical image fusion. This method solved the one-sidedness that the CNN model could only be used for multi-focus images, and the generated fusion images performed well in both subjective and objective evaluation. However, training a good CNN model requires a large amount of label data, which will reduce the efficiency of the algorithm.

In short, CNN has a strong ability in feature extraction and data representation. On the one hand, it can effectively learn features from training data without human intervention; on the other hand, complex relationships between different signals can be modeled by deep convolutional networks, which are suitable for multi-source image fusion, especially the fusion of image data obtained from sensors with large class differences.

4.2. Generate Adversarial Network Methods

The concept of Generative Adversarial Network (GAN) was first proposed by Goodfellow et al. [40] and had attracted wide attention in the field of deep learning. The generative adversarial network consists of two parts, namely generator and discriminator. The generator generates a new sample from the input data, which the discriminator uses to determine whether the sample is from the real data or generated by the generator, iterating until the discriminator can no longer distinguish the generated sample data from the real data. Because of the powerful generative ability of Gans, researchers have applied them in the field of image fusion to generate new fused images. Rao et al. [41] applied GAN to the fusion of infrared and visible images for the first time. By this method, infrared images and light images can be input into the generator to obtain the fusion image, and then the fusion image and visible image are input into the discriminator together.

With the application of GAN in the field of image fusion, some improved GAN models have also been applied to other multimodal image fusion such as medicine. As shown in references [42,43], residual networks are added to the network structure of GAN and applied to the field of image fusion. However, GAN-based image fusion methods only rely on adversarial training to retain more detail information, which is unstable and will lose a lot of detail information. In order to solve this problem, researchers improved the network structure and loss function of GAN [44], and proposed a new GAN model to be applied in infrared and visible image fusion.

4.3. Autoencoder-based Image Fusion

Autoencoder generally includes encoder, fusion layer and decoder. The general procedure of Ae-based fusion method consists of two parts [45]. Firstly, the autoencoder is pre-trained on a public data set to ensure that its encoder and decoder have excellent feature extraction and feature reconstruction capabilities. Then the feature fusion strategy is designed according to the features extracted by the encoder, and the fusion image is reconstructed by the decoder.

Wu et al. [46] proposed the residual structure network DenseNet, which used various feature layers to connect with each other and effectively saved the feature information of source images. Wang et al. [47] proposed Densefuse on the basis of the DenseNet structure. This fusion network densely connected the convolutional layers in the encoder, in which the output of each layer was connected to each other. This network obtained more features from the source image, and they also designed two fusion layers to fuse these features efficiently. Liu et al. [48] proposed an end-to-end fusion network with information retention and feature transmission capabilities, which introduced residual-dense blocks to fully extract depth features of source images, and divided feature information into background layer and detail layer. In addition, they also introduced a weight module to generate adaptive weights to retain similar information in two source images. Dhiravidachelvi et al. [49] proposed an image fusion network, DIVFusion, which was conducive to dark conditions. The network used the light enhancement network to increase the intensity information of visible light images at night, and at the same time saved the texture details through the texture enhancement network, effectively solving the problem of serious degradation of visible light images under dark conditions.

Some researchers improve the performance by changing the network structure. Cao et al. [50] proposed a dual encoder-decoder network named CUFD for image feature extraction and decomposition, effectively extracting deep features and shallow features, and finally using weighted average rule and max hybrid weighted rule for fusion. Cao et al. [51] proposed an image fusion network based on variational autoencoder, which included an image fusion network and an infrared feature compensation network. They used Gaussian probability density product to fuse the mean and variance of each infrared and visible feature map, and used residual block and symmetric jump connection methods in the network to improve the efficiency of network training. Devi et al. [52] proposed UNFusion, which used an encoder-decoder architecture with downsampling operation to learn the contextual features of images. They fused the multi-resolution scale after downsampling with the full resolution scale that retained local detail information, and passed the multi-scale context features at different stages through a cross-stage fusion module. In addition, they used an upsampling module to solve artifacts and ambiguity problems.

Some researchers have introduced attention mechanisms in the study of image fusion in order to retain important information that needs attention. By giving more weight to the features that need to be focused on, the fusion image can effectively retain the significance information of the source image. Cheng et al. [53] proposed a normalized attention model called MSFNet, which provided normalization mechanisms for three different specifications. Through these normalized attention mechanisms, MSFNet was able to highlight and combine depth features in spatial and channel dimensions, and then reconstructed the final fused image using the combined spatial and channel attention graphs. This method extracted and reconstructs multi-scale depth features efficiently, and solved the problem of loss of target region and texture details caused by lack of multi-scale features and global dependence in convolution operation.

Different from the above design schemes, in order to make use of the differences of multi-level features to improve the use of global information, some researchers have proposed a multi-stage multi-level feature fusion scheme based on attention mechanism. Wang et al. [54] proposed an end-to-end infrared and visible image fusion network called SEDRFuse, which had the capability of information retention and feature transmission. They introduced Residual Dense Block (RDB) module in the network to ensure adequate extraction of depth features from source images. In addition, they introduced a weight module for generating adaptive weights to retain similar information in two images, thereby reducing the loss of intermediate information during transmission. Thus, by combining the RDB module and the weight module, the SEDRFuse network enabled efficient feature extraction and reduced information loss to produce high-quality infrared and visible image fusion results. In order to solve the problem of distribution alignment of different image domains, Liu et al. [55] proposed an unsupervised network based on regional feature loss function. The network was trained using regional feature loss function constraints to divide weights according to the importance of different regions. In their model, an attention-based network connection was also introduced for cross-scale information transmission, making full use of information at different scales.

Some researchers have divided the source image into a base layer and a detail layer for fusion. Munasinghe et al. [56] proposed a new fusion network DIDFuse based on autoencoder. They first used encoder to divide the image into background and detail feature maps with low frequency and high frequency information respectively, and used loss function to constrain the background feature maps of source images to be similar while the detail feature maps were different. Finally, the decoder recovered the fused image. In order to solve the problem that the effectiveness of existing methods was affected by a single fusion strategy, Wang et al. [57] adopted the guided filtering method to decompose the source image into the base layer and the detail layer, and designed different fusion strategies for the two layers to adapt to different image fusion tasks.

Some methods use self-supervised learning, which is realized by fine-tuning the autoencoder in the fusion phase. Feng et al. [58] proposed a SSL-WAEIE framework based on self-supervised learning. The network included WAE module for extracting multi-level features and CIEN module for information exchange, which was

the first attempt to conduct image fusion in a self-supervised learning way through network information exchange. In addition, Sauvalle et al. [59] proposed a self-supervised feature adaptive fusion framework, which improved the robustness of the fusion method and made the fusion result contain more complementary information between source images. Fu et al. [60] proposed NestFuse, a network and space-channel attention model based on nested connections. They designed a new fusion strategy, developed a spatial attention model and a channel attention model, described the importance of each spatial location and each channel with depth features, and had better feature fusion performance than other algorithms.

5. Conclusion

Although the first traditional visible and infrared image fusion methods can fuse images well, such methods rely on artificially designed fusion strategies, and most of the methods are not universal, which makes them have certain limitations. With the wide application of deep learning in the field of image processing, many methods based on deep learning have emerged in the field of visible and infrared image fusion. Deep learning can train the network adaptive fusion model through a large amount of data. Visible light and infrared image fusion methods based on deep learning generally have higher fitting degree and can fuse images of higher quality compared with traditional methods, and have certain universality. However, such algorithms also have shortcomings. It requires a large number of data sets for training, and fusion efficiency is low, requiring more time.

In order to facilitate the research of relevant fields, this paper introduces the commonly used data sets and evaluation indicators in visible and infrared image fusion tasks, mainly reviews and analyzes the feature-level visible and infrared image fusion methods, and points out its shortcomings and areas for improvement. Although major breakthroughs have been made in the field of visible and infrared image fusion, there are still many serious challenges and problems to be faced:

1. Real-time performance of high-quality image fusion. Most of the current visible and infrared image fusion algorithms are aimed at static images. However, in some practical applications, dynamic video fusion of infrared and visible light is required, such as auto autonomous driving fusion detection, which requires real-time high-quality fusion. The traditional fusion methods of visible and infrared images can fuse at a faster speed, but the fusion quality is low, and it is not universal. Although the visible and infrared image fusion methods based on deep learning can fuse high-quality images, most fusion models are relatively complex, making the fusion timeliness low, and lightweight models will reduce the quality of fusion images. Therefore, the fusion method of real-time high-quality image fusion is an important direction to be studied.
2. Unregistered image fusion. The existing public visible and infrared image fusion data sets are strictly aligned after registration, but there are some problems in the actual scene such as lens shooting deviation, scene incomplete coincidence, parallax and so on. At present, the registration is basically carried out by manual or mature image registration algorithms. However, due to the need of real-time fusion, the fusion algorithm needs to have the ability of image registration and fusion, which also requires in-depth research by researchers in related fields.
3. Super resolution image fusion. The image taken by the existing visible light sensor has a higher resolution than that taken by the infrared sensor, so to fuse them, the infrared image needs to be superdivided to improve its resolution. Although there are many hypersegmentation algorithms that can achieve better hypersegmentation images, it will increase the time cost of fusion, so how to achieve hypersegmentation image fusion is a difficult problem that needs to be challenged.
4. Image fusion under harsh conditions. In practice, there are often many images under extreme conditions, such as multiple exposures of visible images and severe noise, at which time the texture details of visible images are not reflected. In view of this situation, it is also an important direction for future research to design a suitable visible and infrared image fusion task algorithm to mine the texture details of visible images under extreme conditions and reflect them in the fused images.
5. Advanced visual task-driven image fusion. At present, most visible and infrared image fusion methods are designed to improve the image fusion index, ignoring the promotion of advanced visual tasks such as target detection and semantic segmentation. Although there are also methods by adding semantic loss function, fusion network and advanced visual task network generation confrontation into fusion network, etc. However, how to realize the combination of fusion network and high vision task network to achieve end-to-end promotion generation still needs further research.
6. Assessment reliability of the criteria. Appropriate evaluation criteria can objectively reflect the quality of image fusion. At present, there are many evaluation indicators for visible light and infrared image fusion, but there is no authoritative evaluation standard, and the evaluation criteria selected by different fusion algorithms are also different. The development of a reliable and authoritative evaluation standard can promote the further development of the IVIF mission.

6. Conflict of Interest

The authors declare that there are no conflict of interests, we do not have any possible conflicts of interest.

Acknowledgments. This work was supported by the Science and technology research project, name "Research on key technologies of image fusion based on deep learning" (Project number: 242102210187).

References

1. Karim S, Tong G, Li J, et al. Current advances and future perspectives of image fusion: A comprehensive review[J]. Information Fusion, 2023, 90: 185-217.
2. Zhang X, Demiris Y. Visible and infrared image fusion using deep learning[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(8): 10535-10554.
3. Yin S, Wang L, Teng L. Threshold segmentation based on information fusion for object shadow detection in remote sensing images[J]. Computer Science and Information Systems, 2024. doi: 10.2298/CSIS231230023Y.
4. Yin S, Li H, Sun Y, et al. Data Visualization Analysis Based on Explainable Artificial Intelligence: A Survey[J]. IJLAI Transactions on Science and Engineering, 2024, 2(2): 13-20.
5. Roy S K, Deria A, Hong D, et al. Multimodal fusion transformer for remote sensing image classification[J]. IEEE Transactions on Geoscience and Remote Sensing, 2023, 61: 1-20.
6. Ma W, Wang K, Li J, et al. Infrared and visible image fusion technology and application: A review[J]. Sensors, 2023, 23(2): 599.
7. Teng L, Qiao Y, Shafiq M, et al. FLPK-BiSeNet: Federated learning based on priori knowledge and bilateral segmentation network for image edge extraction[J]. IEEE Transactions on Network and Service Management, 2023, 20(2): 1529-1542.
8. He J, Yuan Q, Li J, et al. A self-supervised remote sensing image fusion framework with dual-stage self-learning and spectral super-resolution injection[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2023, 204: 131-144.
9. Zhao Z, Bai H, Zhu Y, et al. DDFM: denoising diffusion model for multi-modality image fusion[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 8082-8093.
10. Tang L, Zhang H, Xu H, et al. Rethinking the necessity of image fusion in high-level vision tasks: A practical infrared and visible image fusion network based on progressive semantic injection and scene fidelity[J]. Information Fusion, 2023, 99: 101870.
11. Liu J, Dian R, Li S, et al. SGFusion: A saliency guided deep-learning framework for pixel-level image fusion[J]. Information Fusion, 2023, 91: 205-214.
12. Liu J, Lin R, Wu G, et al. Coconet: Coupled contrastive learning network with multi-level feature ensemble for multi-modality image fusion[J]. International Journal of Computer Vision, 2024, 132(5): 1748-1775.
13. Yin S. Object Detection Based on Deep Learning: A Brief Review[J]. IJLAI Transactions on Science and Engineering, 2023, 1(02): 1-6.
14. Jiang, Y., Yin, S. Heterogenous-view Occluded Expression Data Recognition Based on Cycle-Consistent Adversarial Network and K-SVD Dictionary Learning Under Intelligent Cooperative Robot Environment[J]. Computer Science and Information Systems, vol. 20, no. 4, 2023. https://doi.org/10.2298/CSIS221228034J.
15. Senkerik R. Remote Sensing Image Fusion Based on PCA and Wavelets[C]//Intelligent Data Engineering and Analytics: Proceedings of the 10th International Conference on Frontiers in Intelligent Computing: Theory and Applications (FICTA 2022). Springer Nature, 2023, 327: 25.
16. Kalamkar S. Multimodal image fusion: A systematic review[J]. Decision Analytics Journal, 9, 2023: 100327.
17. Haribabu M, Guruviah V, Yogarajah P. Recent advancements in multimodal medical image fusion techniques for better diagnosis: an overview[J]. Current Medical Imaging, 2023, 19(7): 673-694.
18. He C, Li K, Zhang Y, et al. Camouflaged object detection with feature decomposition and edge reconstruction[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023: 22046-22055.
19. Mishra V K, Kumar R, Nareti U, et al. Pansharpening Using IHS Method on Multi-sensor Data and Multiple Feature Extraction Using Modified Otsu Thresholding[J]. Journal of the Indian Society of Remote Sensing, 2024, 52(1): 113-126.
20. Li L, Lv M, Jia Z, et al. Sparse representation-based multi-focus image fusion method via local energy in shearlet domain[J]. Sensors, 2023, 23(6): 2888.
21. Li L, Lv M, Jia Z, et al. Sparse representation-based multi-focus image fusion method via local energy in shearlet domain[J]. Sensors, 2023, 23(6): 2888.
22. Li X, Wan W, Zhou F, et al. Medical image fusion based on sparse representation and neighbor energy activity[J]. Biomedical Signal Processing and Control, 2023, 80: 104353.
23. Guo P, Xie G, Li R, et al. Multimodal medical image fusion with convolution sparse representation and mutual information correlation in NSST domain[J]. Complex & Intelligent Systems, 2023, 9(1): 317-328.
24. Luo X, Fu G, Yang J, et al. Multi-modal image fusion via deep laplacian pyramid hybrid network[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 33(12): 7354-7369.
25. Fang A, Zhao X, Yang J, et al. A light-weight, efficient, and general cross-modal image fusion network[J]. Neurocomputing, 2021, 463: 198-211.
26. Yao J, Zhao Y, Bu Y, et al. Laplacian pyramid fusion network with hierarchical guidance for infrared and visible image fusion[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 33(9): 4630-4644.

27. Sun Y, Xu H, Ma Y, et al. Dual spatial-spectral pyramid network with transformer for hyperspectral image fusion[J]. IEEE Transactions on Geoscience and Remote Sensing, vol. 61, pp. 1-16, 2023.
28. Wang J, Xie X, Li D, et al. GRPAFusion: A gradient residual and pyramid attention-based multiscale network for multimodal image fusion[J]. Entropy, 2023, 25(1): 169.
29. Aghamaleki J A, Ghorbani A. Image fusion using dual tree discrete wavelet transform and weights optimization[J]. The Visual Computer, 2023, 39(3): 1181-1191.
30. Dogra A, Kumar S. Multi-modality medical image fusion based on guided filter and image statistics in multidirectional shearlet transform domain[J]. Journal of Ambient Intelligence and Humanized Computing, 2023, 14(9): 12191-12205.
31. Nagaraja Kumar N, Jayachandra Prasad T, Prasad K S. An intelligent multimodal medical image fusion model based on improved fast discrete curvelet transform and type-2 fuzzy entropy[J]. International Journal of Fuzzy Systems, 2023, 25(1): 96-117.
32. Avci D, Sert E, zyurt F, et al. MFIF-DWT-CNN: Multi-focus image fusion based on discrete wavelet transform with deep convolutional neural network[J]. Multimedia Tools and Applications, 2024, 83(4): 10951-10968.
33. Gong X, Hou Z, Wan Y, et al. Multispectral and SAR image fusion for multi-scale decomposition based on least squares optimization rolling guidance filtering[J]. IEEE Transactions on Geoscience and Remote Sensing, vol. 62, pp. 1-20, 2024.
34. Hao H, Zhang B, Wang K. MGFuse: An infrared and visible image fusion algorithm based on multiscale decomposition optimization and gradient-weighted local energy[J]. IEEE Access, 2023, 11: 33248-33260.
35. Zhou Z, Fei E, Miao L, et al. A perceptual framework for infraredCvisible image fusion based on multiscale structure decomposition and biological vision[J]. Information Fusion, 2023, 93: 174-191.
36. Ravi J, Subbayamma B V, Kumar P V, et al. Multi-image fusion: optimal decomposition strategy with heuristic-assisted non-subsampled shearlet transform for multimodal image fusion[J]. Signal, Image and Video Processing, 2024, 18(3): 2297-2307.
37. Li L, Li C, Lu X, et al. Multi-focus image fusion with convolutional neural network based on Dempster-Shafer theory[J]. Optik, 2023, 272: 170223.
38. Shao X, Xie X, Jiang Q, et al. Multi-focus image fusion based on transformer and depth information learning[J]. Computers and Electrical Engineering, 2024, 119: 109629.
39. Yin S, Wang L, Wang Q, et al. M2F2-RCNN: Multi-functional faster RCNN based on multi-scale feature fusion for region search in remote sensing images[J]. Computer Science and Information Systems, vol. 20, no. 4, pp. 1289-1310, 2023. https://doi.org/10.2298/CSIS230315054Y.
40. Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial networks[J]. Communications of the ACM, 2020, 63(11): 139-144.
41. Rao Y, Wu D, Han M, et al. AT-GAN: A generative adversarial network with attention and transition for infrared and visible image fusion[J]. Information Fusion, 2023, 92: 336-349.
42. Li K, Liu G, Gu X, et al. DANT-GAN: A dual attention-based of nested training network for infrared and visible image fusion[J]. Digital Signal Processing, 2024, 145: 104316.
43. Chang L, Huang Y, Li Q, et al. DUGAN: Infrared and visible image fusion based on dual fusion paths and a U-type discriminator[J]. Neurocomputing, 2024, 578: 127391.
44. Fan Y, Li H, Sun B. Cycle GAN-MF: A Cycle-consistent Generative Adversarial Network Based on Multifeature Fusion for Pedestrian Re-recognition[J]. IJLAI Transactions on Science and Engineering, 2024, 2(1): 37-44.
45. Meng X, Wang X, Yin S, et al. Few-shot image classification algorithm based on attention mechanism and weight fusion[J]. Journal of Engineering and Applied Science, 2023, 70(1): 14.
46. Wu C M, Ren M L, Lei J, et al. CAEFusion: A New Convolutional Autoencoder-Based Infrared and Visible Light Image Fusion Algorithm[J]. Computers, Materials & Continua, 2024, 80(2).
47. Wang H, Li L, Li C, et al. Infrared and Visible Image Fusion Based on Autoencoder Composed of CNN-Transformer[J]. IEEE Access, vol. 11, pp. 78956-78969, 2023.
48. Liu H, Yan H. An end-to-end multi-scale network based on autoencoder for infrared and visible image fusion[J]. Multimedia Tools and Applications, 2023, 82(13): 20139-20156.
49. Dhiravidachelvi E, Devadas T J, Kumar P J S, et al. Enhancing image classification using adaptive convolutional autoencoder-based snow avalanches algorithm[J]. Signal, Image and Video Processing, 2024, 18(10): 6867-6879.
50. Cao B, Cao H, Liu J, et al. Autoencoder-Based Collaborative Attention GAN for Multi-Modal Image Synthesis[J]. IEEE Transactions on Multimedia, 2023, 26: 995-1010.
51. Cao B, Bi Z, Hu Q, et al. Autoencoder-driven multimodal collaborative learning for medical image synthesis[J]. International Journal of Computer Vision, 2023, 131(8): 1995-2014.
52. Devi M G, Akila I S. Image fusion: A deep Y shapedCresidual convolution auto-encoder with MS-SSIM loss function[J]. Journal of Radiation Research and Applied Sciences, 2024, 17(4): 101089.
53. Cheng C, Xu T, Wu X J. MUFusion: A general unsupervised image fusion network based on memory unit[J]. Information Fusion, 2023, 92: 80-92.
54. Wang C, Gao G, Wang J, et al. GCT-VAE-GAN: An Image Enhancement Network for Low-Light Cattle Farm Scenes by Integrating Fusion Gate Transformation Mechanism and Variational Autoencoder GAN[J]. IEEE Access, vol. 11, pp. 126650-126660, 2023.
55. Liu T, Yan S, Wang G. Remove and recover: Two stage convolutional autoencoder based sonar image enhancement algorithm[J]. Multimedia Tools and Applications, 2024, 83(18): 55963-55979.
56. Munasinghe A P, Chathumini K G L. Motion Deblurring Through Autoencoder-Based Image Restoration[C]//2024 4th International Conference on Advanced Research in Computing (ICARC). IEEE, 2024: 137-142.

57. Wang E, Zhou H, Wen G, et al. Bearing performance degradation assessment using adversarial fusion convolutional autoencoder based on multi-source information[J]. Transactions of the Institute of Measurement and Control, 2024, 46(5): 992-1011.
58. Feng Y, Zhang Y, Zhou Z, et al. Memristor-based storage system with convolutional autoencoder-based image compression network[J]. Nature Communications, 2024, 15(1): 1132.
59. Sauvalle B, de La Fortelle A. Autoencoder-based background reconstruction and foreground segmentation with background noise estimation[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2023: 3244-3255.
60. Fu Q, Fu H, Wu Y. Infrared and Visible Image Fusion Based on Mask and Cross-Dynamic Fusion[J]. Electronics, 2023, 12(20): 4342.

Biography

Lijuan Feng is with the School of Electronics and Electrical Engineering, Zhengzhou University of Science and Technology. Research direction is computer application and AI.

Jiangjiang Li is with the School of Electronics and Electrical Engineering, Zhengzhou University of Science and Technology. Research direction is computer application and AI.