frontiers | Frontiers in Neuroscience

TYPE Original Research
PUBLISHED 09 August 2022
DOI 10.3389/fnins.2022.986153

Check for updates

OPEN ACCESS

EDITED BY
Jiayi Ma,
Wuhan University, China

REVIEWED BY
Hancheng Zhu,
China University of Mining and Technology, China
Pengfei Chen,
Xidian University, China

*CORRESPONDENCE
Chuangeng Tian
tianchuangeng@163.com

SPECIALTY SECTION
This article was submitted to Brain Imaging Methods, a section of the journal Frontiers in Neuroscience

RECEIVED 04 July 2022
ACCEPTED 13 July 2022
PUBLISHED 09 August 2022

CITATION
Tang L, Hui Y, Yang H, Zhao Y and Tian C (2022) Medical image fusion quality assessment based on conditional generative adversarial network.
Front. Neurosci. 16:986153.
doi: 10.3389/fnins.2022.986153

Medical image fusion quality assessment based on conditional generative adversarial network

Lu Tang[1], Yu Hui[1], Hang Yang[1], Yinghong Zhao[1] and Chuangeng Tian[2]*

[1]School of Medical Imaging, Xuzhou Medical University, Xuzhou, China, [2]School of Information and Electrical Engineering, Xuzhou University of Technology, Xuzhou, China

Multimodal medical image fusion (MMIF) has been proven to effectively improve the efficiency of disease diagnosis and treatment. However, few works have explored dedicated evaluation methods for MMIF. This paper proposes a novel quality assessment method for MMIF based on the conditional generative adversarial networks. First, with the mean opinion scores (MOS) as the guiding condition, the feature information of the two source images is extracted separately through the dual channel encoder-decoder. The features of different levels in the encoder-decoder are hierarchically input into the self-attention feature block, which is a fusion strategy for self-identifying favorable features. Then, the discriminator is used to improve the fusion objective of the generator. Finally, we calculate the structural similarity index between the *fake* image and the *true* image, and the MOS corresponding to the maximum result will be used as the final assessment result of the fused image quality. Based on the established MMIF database, the proposed method achieves the state-of-the-art performance among the comparison methods, with excellent agreement with subjective evaluations, indicating that the method is effective in the quality assessment of medical fusion images.

KEYWORDS

attention mechanism, conditional, generative adversarial networks, image quality assessment, medical image fusion

Introduction

As the population aging becomes familiar, and the vulnerability of the human brain to physical, chemical, and viral attacks, the incidence of brain diseases such as intracranial tumors, intracranial infectious diseases, and cerebrovascular diseases is gradually increasing, which has seriously threatened human health and wellbeing (Chen et al., 2022; Gottesman and Seshadri, 2022). There are many medical imaging modalities for clinical diagnosis and treatment of brain diseases, including computed tomography

(CT), magnetic resonance imaging (MRI), positron emission tomography (PET), and so on. Different imaging methods always have their unique advantages in attracting clinicians to choose (Liu et al., 2019; Cauley et al., 2021; Preethi and Aishwarya, 2021). For example, CT could superbly display the histological structure of the skull and the density changes in the brain parenchyma, while MRI could faithfully restore the essential features of the nervous or soft tissue. Generally, it is difficult for medical experts to identify the necessary information from a single modality of brain images to ensure the reliability of clinical diagnosis (Townsend, 2008). Additionally, some early work found that radiologists could effectively improve the diagnostic accuracy if they can analyze imaging results of more than two modalities at the same time (Li and Zhu, 2020). From a technical point of view, multimodal medical image fusion (MMIF) just meets this clinical need. Therefore, recently, MMIF has received attention and extensive exploration by researchers (Li et al., 2020; Ma et al., 2020; Liu et al., 2021).

The purpose of MMIF is to complement the image in different modalities to obtain better image expression, quality, and information perception experience (Azam et al., 2022; Liu et al., 2022). The fused images may contain both anatomical structure and tissue metabolism information (e.g., image fusion of CT and MRI), which improves the applicability of image-based diagnosis or assessment of diseases, thereby simplifying diagnosis. At present, many high-quality MMIF methods have been proposed (Arif and Wang, 2020; Wang K. P. et al., 2020; Duan et al., 2021; Ma et al., 2022; Xu et al., 2022). Madanala and Rani (2016) proposed a two-stage fusion framework based on the cascade of discrete wavelet transform (DWT) and non-subsampled contour transform (NSCT) domains, realizing the combination of spatial domain and transform domain. Inspired by the Tchebichef moments' ability to effectively capture edge features, Tang et al. (2017) used the Tchebichef moments energy to characterize the image shape, and thus designed an MMIF method based on the pulse coupled neural network (PCNN). However, the performance evaluation of these MMIF models and fused images has not been fully explored.

Normally, the higher the image quality, the more features and information human observers can receive or perceive through the image. As the ultimate observers and beneficiaries of the fused images, medical experts, although they subjectively evaluate the fused images as the most direct and reliable solution, it will be a very time-consuming and labor-intensive task, and it is not very useful in practical applications. Hence, objective image quality assessment (IQA) is very necessary (Liu et al., 2018; Shen et al., 2020; Wang W. C. et al., 2020). Some existing objective quality assessment studies include deblocking images, screen content images, multiple distorted images, and noisy images, etc. (Gao et al., 2008; Min et al., 2019; Liu et al., 2020; Meng et al., 2020). For instance, in early work, Wang et al. (2004) developed structural similarity (SSIM) index based on the subjective perception of image structure information, which achieved a breakthrough in the objective evaluation of image quality. Kang et al. (2014) used deep learning techniques to accurately predict the quality of images without reference images, and their method greatly improved the performance and robustness of the algorithm. On the premise of highlighting the important detection objects, Lei et al. (2022) fuses multiple features of the images at the pixel level and designed an IQA method of main target region extraction and multi-feature fusion. However, among these IQA methods, they are proposed for general use in the field of image fusion, not specifically for MMIF. Note that the quality assessment of medical fusion images includes information fidelity, contrast, grayscale tolerance, and region of interest (ROI). In clinical practice, the ROI usually refers to the lesion area. And, the ROI has a great influence on the results of IQA, which is the most different from the natural image (Du et al., 2016a; Cai et al., 2020; Chabert et al., 2021). As a result, there is an urgent need for a dedicated objective IQA method for medical fusion images.

We discussed with radiologists and found that the quality of a medical fusion image mainly depends on its impact on disease diagnosis. That is, the medical fusion image retains disease-relevant information in the ROI, it will be acceptable and will be given a higher subjective evaluation score. To this end, we propose a novel medical fusion image quality assessment method that uses the radiologist's mean opinion scores (MOS) as the constraint on conditional generative adversarial networks (GANs). Concretely, the method firstly extracts the feature of different depths from MOS and two input source images with the aid of dual-channel encoder-decoder. Next, under the supervision of the attention mechanism, we fuse the feature information hierarchically, and generate the fused image through the up-sampling algorithm. Then, the discriminator (***D***) differentiates the source of the fused images to improve the generator (***G***) performance. Finally, we calculate the SSIM of the *fake* image and *true* image, and the constrain value corresponding to the maximum value of SSIM as the evaluation result. The experimental results show that the proposed method is superior to the previous IQA algorithms, and the objective results obtained are more consistent with the subjective evaluation of radiologists.

The content of this paper is arranged as follows. In see section "Methodology," the proposed method is mainly introduced from four aspects: Encoder-Decoder, ***G***, ***D***, and objective function. The details of the experiments are presented in see section "Experiments." See section "Discussion and conclusion" contains the discussion and conclusion of this paper.

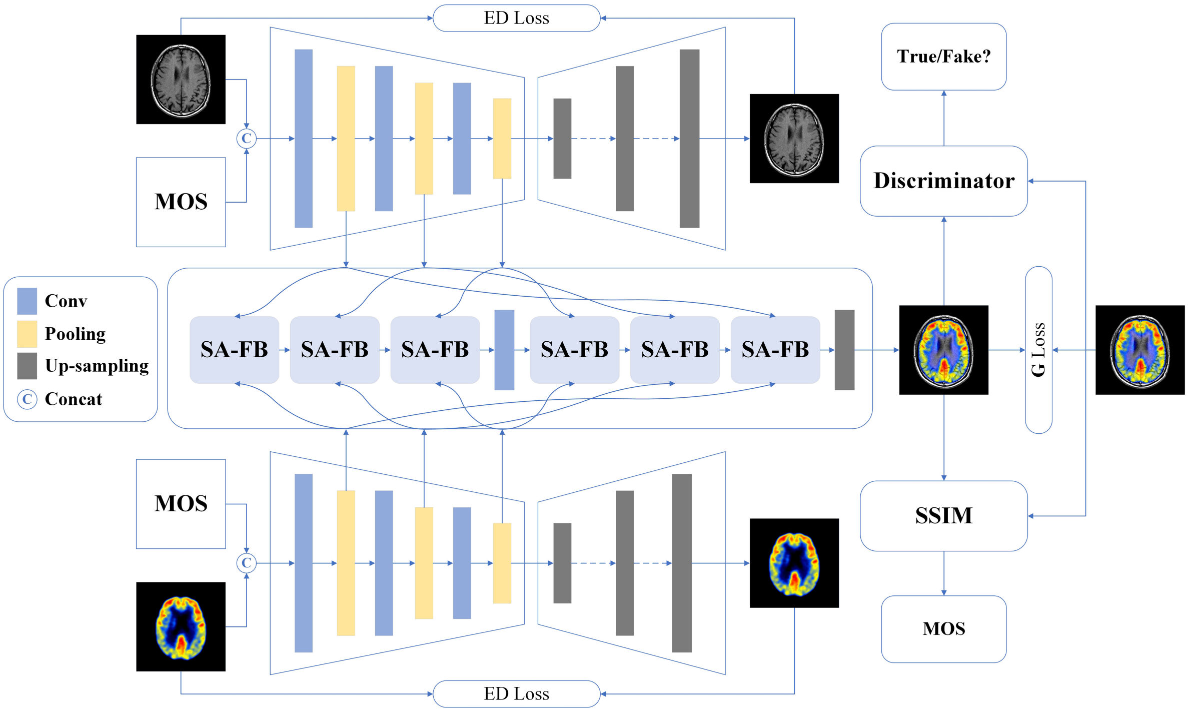

FIGURE 1
The overall architecture of our proposed method.

Methodology

The structure of our proposed model based on conditional generative adversarial network is shown in **Figure 1**, and the details are described below.

Dual-channel encoder-decoder

Among the existing multimodal medical images, each image has its unique imaging method and the advantage of displaying different human tissue. Therefore, accurately extracting the latent and deep key features of each modality image will be extremely conducive the image fusion (Ma et al., 2019). Besides, we also hope that MOS, the gold standard for image quality assessment, can participate in the feature extraction process of model learning images, in other words, learning the non-linear mapping relationship between MOS and fused images. To achieve this vision, we develop a dual-channel encoder-decoder structure.

First of all, we encapsulate three convolutional blocks, each of which contains two sets of convolutional layers, batch normalization (BN) layers, and activation layers. Specifically, the filter, stride, and padding of each convolutional layer are 3 × 3, 1, and 1, respectively. BN operation can effectively accelerate the network training as well as alleviate the problem of over-fitting. Thus, we append such operation after each convolutional layer. Considering that the image encoding process is important to learn image features and image fusion, we use a more comprehensive activation algorithm: Lleaky Rectified Line Unit (LeakyReLU). Then, we added max pooling operation instead of average pooling operation after each convolutional block. The reason is that the model should perform some specific feature selection under the constraints of MOS to learn more recognizable features. Each feature map output through the pooling operation is fed to the self-attention fusion block (SA-FB) separately, and more details will be explained in the next section. For the decoder, seven groups of deconvolution layer, BN layer, and Rectified Line Unit (ReLU) activation function layer complete the up-sampling operation of the feature maps. Finally, a reconstructed image of size 128 × 128 is obtained. It is worth noting that during the decoding operation, there is no feature map as output.

Perform the concatenating operation on the image of two different modalities (MI_i, $i = 1, 2$) and the corresponding MOS of their fused image, and the result is named MI_{imos}, and then input into two encoder-decoders, respectively. The feature map after the pooling layer is represented as F_{ij}, then the j-th feature map for the i-th modality can be marked as:

$$F_{ij} = ConvB(MI_{imos})_j \tag{1}$$

where *ConvB*(•) means the operation process of the *j*-th convolution block. The integer value range of *j* is one to three as only three convolution blocks are established in the encoding process. Here, sum of absolute difference is employed as the loss function for single modality image restoration, as defined by the following equation:

$$L_{ED} = \sum_{i} \sum \left| MI_i - \widehat{MI_i} \right|, i = 1, 2 \tag{2}$$

Where $\widehat{MI_i}$ refers to the original modal image restored by the decoder, and *i* represent the two modal images input to the dual-channel encoder-decoder, respectively.

Generator architecture

It is generally known that image fusion is the operation of synthesizing two or more images into one image, preserving the most representative features of each modality. To avoid the impact on image feature learning, independent of the dual-channel encoder-decoders, we design a feature fusion method based on the self-attention (SA) mechanism, as shown in **Figure 1**. Different levels of features contain different image information, for example, shallow features mean contour information while deep features represent texture information. For the three-level of feature F_{ij} yielded in the encoder, we develop the SA-FB to complete the fusion hierarchically. The structure diagram of SA-FB is shown in **Figure 2**.

In particular, the first SA-FB has only two inputs (i.e., F_{ij}), and the fusion feature F_{sa} is null. We do not carry out any feature selection operations (such as taking extreme values) during inputting, but directly feed the initial features F_{1j} and F_{2j} to SA after concatenating, and SA will sign weights to the features. Such setting can replace the manual feature selection algorithm, thus avoiding the loss of important information. SA is a variant of the attention mechanism from Sergey and Nikos (2017). It could coarsely estimate the foreground region to find prominent features that are in favor of later search. At the same time, it also reduces the dependence on external information, and is better at capturing the internal relevance of features. Immediately after, we adopt a convolution layer at the end of the SA. The convolution kernel size is set to 1 × 1 with stride 1 for adapt the output feature map weights. The output of this convolutional layer is concatenated with F_{sa}, and further input to a new convolution layer with a filter size of 3 × 3, and stride 1. In the end, a feature output F_{sa+1} that has undergone a complete SA-FB is obtained, and can be expressed as:

$$F_{sa+1} = safb(F_{ij}, F_{sa}), (i = 1, 2, j = 1, 2, 3) \tag{3}$$

where *safb*(•) is a series of operations of SA-FB. It should be mentioned that each convolution layer in the first three SA-FB is followed by BN layer and LeakyReLU as an activation function, which is similar to the encoder. The max pooling operation also appends after each SA-FB. The SA-FB in the up-sampling stage eliminates the pooling operation and changes the activation function to ReLU. On the basis of MOS as the condition to extract two modal image features, the ***G*** generates a fused image with 128 × 128. The parameters of the ***G*** are only renewed by the following loss function:

$$L_{fusion} = \frac{1}{N} \sum_{n=1}^{N} \left| y_{true} - \hat{y} \right|_1 \tag{4}$$

where y_{true} means the fused image with the corresponding MOS and the $\hat{y}$ represents the fused image produced by the ***G***. *N* is the total number of generations, and *n* represents the *n*-th generation. When training ***G***, minimize the following objective function:

$$L_G = V_G^{mos}(G, D) = E_{MI_1, MI_2 \sim P_{\text{data}M}}$$

$$[\log(1 - D(MI_1, MI_2, (G(MI_1, MI_2 \,|mos))))] + \alpha L_{fusion} \tag{5}$$

where P_{dataM} represents the distribution of MI_1 and MI_2, respectively, and $E_{MI_1, MI_2 \sim P_{dataM}}$ represents the expectation of $G(MI_1, MI_2 \,|mos)$. α is a weight hyperparameter and is set to 100 during training.

To sum up, we restrict the generator based on MOS conditional information, and achieve the goal of generating image content. This is similar to that the generator analyzes the fused image by simulating the human visual system (HVS) and learns the non-linear mapping relationship between MOS and image. That is, the generator simulates a radiologist to assess the quality of the fused image, there by producing a fused image that matches the quality of MOS (i.e., ***G*** has learned the evaluation experience of radiologist).

To evaluate the quality of the fused image FI_{12}, first of all its original two modal images FI_1 and FI_2 should be input and then generate the fusion image FI_{fake} by ***G***. Where *1* and *2* represent two modal images, respectively. We have created five *fake* MOS ($MOS_k = 0.2k$, $k \in [1, 5]$, $k \in \mathbb{Z}$) as the conditional constraints ***G***, so the FI_{fake} can be renewed to $FI_{fake,k}$, which represents the fused image generated under the five constraints. Finally, the SSIM between FI_{12} and $FI_{fake,k}$ is calculated, and the MOS corresponding to the optimal value is taken as the assessment result, as follows:

$$Q = \max SSIM(FI_{12}, G(FI_1, FI_2 \,|MOS_k)) \tag{6}$$

Discriminator architecture

The discriminator needs to determine whether the generated image conforms to the real data distribution, so its structure is much simpler than the generator. In the proposed method, the input of the ***D*** is the generated fusion image or the original fusion image, all of which are 128 × 128 in size, and down-sampling is implemented using the discriminator

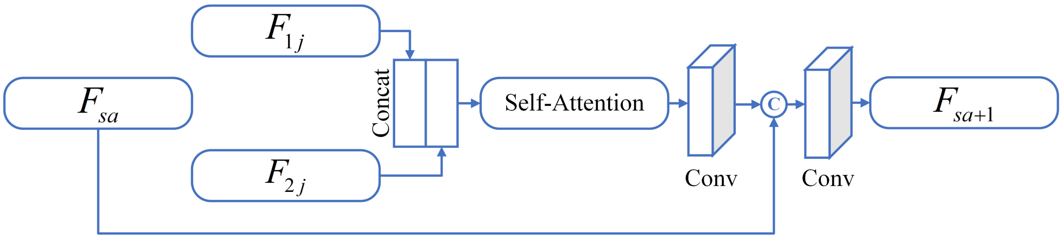

FIGURE 2
The diagrammatic sketch of SA-FB.

block (DB). Each DB consists of a convolution layer with a filter size of 3 × 3, stride of 2 and padding of 1, and followed by BN processing. The LeakyReLU is used as the activation function for each block. The image passes through four DB in sequence, and after each DB, the size of the feature map becomes a quarter of that before input. An independent convolutional layer with convolution kernel 3 × 3 and stride 1 is appended to the last DB, and the final obtained feature map is 6 × 6. At last, the discriminator will judge the authenticity of the result. We apply mean square error (MSE) as the loss function to optimizing the parameters of the ***D***. Further, the objective function of ***D*** can be reformulated as:

$$L_D = V_D^{mos}(G, D) = E_{y_{\mathrm{true}} \sim P_{data}}[\log D(y_{true} \\ |mos)]E_{MI_1, MI_2 \sim P_{dataM}}[\log(1 - D(G(MI_1, MI_2\,|mos)))] \quad (7)$$

where P_{data} represents the distribution of y_{true} and $E_{y_{true} \sim P_{data}}$ represents the expectation of y_{true} .

Total objective loss function

As shown in **Figure 1**, we use MOS as a condition to limit the content of the image generated by ***G***, and ***D*** determines whether the distribution of the generated fused image is true or false. ***G*** and ***D*** are trained against each other, and finally achieve the goal of Nash Equilibrium. Therefore, the optimization process of the whole network can be expressed by Eq. 8:

$$L_{all} = \min_G \max_D V(G, D) + \beta L_{ED} \quad (8)$$

where $V(G, D)$ can be obtained by Eqs. 5 and 7, respectively. β is a weight hyperparameter and is set to 20 in this experiment.

Experiments

Dataset

Image quality assessment has been developed in full swing in many fields and has made substantial progress. But, in the past period, the short-lived time of the MMIF algorithm has resulted in few research dedicated to the quality assessment of medical fusion images. In order to enable the medical image fusion algorithm to restore the brain structure more accurately and reflect tissue metabolic information more objectively, meeting the needs of clinical diagnosis, based on our previous work (Tang et al., 2020), we construct a special multimodal medical image fusion image database (MMIFID) with subjective evaluation of radiologists. Particularly, this work uses brain images from the AANLIB dataset, provided by Harvard Medical School and accessible online. The image size is 256 × 256, which can be browsed directly on the online web page. Most importantly, since image registration is completed for each combination of different modal images, it is one of the most widely used datasets. We selected 120 pairs of images in the AANLIB dataset and fused the images through ten image fusion algorithms. **Figure 3** shows examples of results generated by ten fusion algorithms. Consistent with our previous work (Tang et al., 2020), radiologists subjectively evaluated the quality of the fused image and gave a score (1 is the lowest and 5 is the highest), and finally obtained the MOS.

Evaluation metrics

To comprehensively evaluate the performance of the proposed method, that is, the consistency of the model's assessment of the fused image quality with the MOS score, we adopted four commonly used performance metrics: Spearman Rank-order Correlation Coefficient (SRCC), Kendall Rank-order Correlation Coefficient (KRCC), Pearson Linear Correlation Coefficient (PLCC), and Root Mean Square Error (RMSE). To sum up, the higher SRCC, KRCC and PLCC value and lower RMES value mean better model performance. Note, the model is evaluated at the end of each training epoch, and the final model is the checkpoint model with the best evaluation performance within 200 epochs.

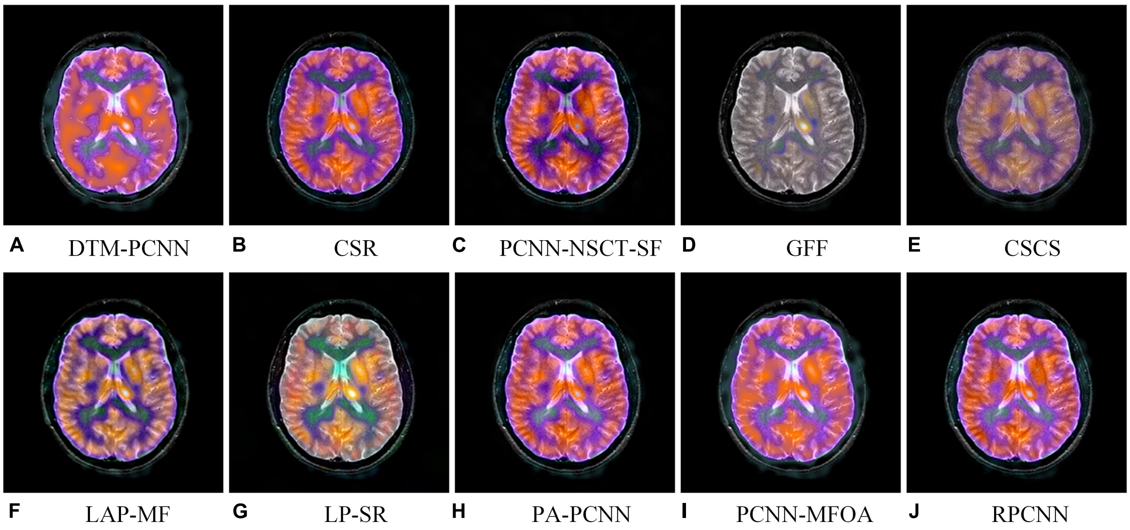

FIGURE 3
An example of fused images generated by ten different MMIF algorithms. Algorithms include **(A)** discrete Tchebichef moments and pulse coupled neural network (DTM-PCNN) (Min et al., 2019), **(B)** convolutional sparse representation (CSR) (Liu et al., 2016), **(C)** pulse-coupled neural network with modified spatial frequency based on non-subsampled contourlet transform (PCNN-NSCT-SF) (Das and Kundu, 2012), **(D)** guided filtering (GFF) (Li et al., 2013), **(E)** cross-scale coefficient selection (CSCS) (Shen et al., 2013), **(F)** union Laplacian pyramid with multiple features (LAP-MF) (Du et al., 2016b), **(G)** Laplacian pyramid and sparse representation (LP-SR) (Liu et al., 2015), **(H)** parameter-adaptive pulse-coupled neural network (PA-PCNN) (Yin et al., 2019), **(I)** pulse coupled neural network using the multi-swarm fruit fly optimization algorithm (PCNN-MFOA) (Tang et al., 2019), and **(J)** reduced pulse-coupled neural network (RPCNN) (Das and Kundu, 2013).

Comparison methods

The results are compared with those of the state-of-the-art (SOTA) image fusion quality metrics, which are listed as follows:

Mutual Information (Q_{MI}) (Hossny et al., 2008): As an objective method for evaluating image fusion performance, this method can measure the features and visual information from the input initial image and the fused image. The MI method we adopted is optimized by Hossny et al. (2008).

Non-linear Correlation Information Entropy (Q_{NCIE}) (Wang and Shen, 2004): Wang et al. propose a method based on non-linear correlation measures. This method evaluates the performance of image fusion algorithms by analyzing the general relationship between the source image and the fused image.

Gradient based fusion metric (Q_G) (Xydeas and Petrovic, 2000): This performance metric measures the amount of visual information transmitted from the source image to the fused image.

Ratio of spatial frequency error (Q_{rSFe}) (Zheng et al., 2007): This is a new metric based on extended spatial frequencies, and its original intention is to guide the algorithm to obtain a better fusion image.

The metric proposed by Yang et al. (2008) (Q_Y): According to the structural similarity between the source image and the fused image, this method treats redundant regions and complementary / conflicting regions, respectively.

A metrics based on edge preservation (Q_{EP}) (Wang and Liu, 2008): An image fusion metric method is proposed based on the perspective of edge information preservation.

A metric based on an absolute image feature measurement (Q_P) (Zhao et al., 2007): Based on phase congruency and its moments, a pixel-level image fusion performance metric is defined, which provides an absolute measure of image features.

Table 1 shows the performance of the above methods on our MMIFID, and the last row is the performance of the method proposed in this paper. Generally, SRCC, KRCC, and PLCC

TABLE 1 Comparison of quality assessment performance of different models.

Methods	SRCC	KRCC	PLCC	RMSE
Q_{MI}	0.2545	0.3604	0.2772	0.3804
Q_{NCIE}	0.2647	0.3608	0.2920	0.4093
Q_G	0.2488	0.3322	0.2444	0.2791
Q_{rSFe}	0.1801	0.2076	0.3126	0.2872
Q_Y	0.1884	0.2400	0.2503	0.4002
Q_{EP}	0.0960	0.1275	0.2235	0.2970
Q_P	0.1093	0.1216	0.0803	0.3007
Proposed	**0.8259**	**0.7426**	**0.8197**	**0.1709**

The bold values are the results of our proposed method, which achieves the best performance.

can measure the agreement between MOS and the objective scores, while RMSE can calculate its absolute error. Thus, the higher the SRCC, KRCC, and PLCC values, the better the quality evaluation metrics. The smaller the RMSE, the higher accuracy of the assessment. From **Table 1**, we can observe that the proposed method outperforms all SOAT methods. Furthermore, it can also be noticed that our proposed metrics are obviously better than these methods, which especially highlights that the quality assessment methods for medical images differ from natural images. Therefore, it is necessary to explore the special indicators for the quality evaluation of medical fusion images.

Ablation experiment

As we know, image fusion can be divided into two categories: early fusion and late fusion. The early fusion fuses the image directly together and then carries on the process of feature extraction and selection, while the late fusion allows the images to go through the process of feature extraction and selection, respectively, and then perform image feature fusion. Therefore, our two ablation experiments are to downgrade the proposed method to the early fusion and late fusion model, named Early-FM and Late-FM, respectively. Specifically, Early-FM first concatenates F_1, F_2 and MOS, and then completes feature learning through the single-channel encoder-decoder structure (e.g., we use the single-channel encoder-decoder to replace dual-channel encoder-decoder). The features output by the third convolutional block will be used to generate the fused image. Different from Early-FM, the Late-FM first concatenates the images of the two modalities and their respective MOS, and then inputs them to the dual-channel encoder-decoder, respectively, to complete feature learning. The third convolution block of the two channels outputs features, and the fused features are obtained by fusion operation. Finally, $\boldsymbol{G}$ generates the fused image. For the third ablation experiment, we eliminated the SA mechanism in SA-FB, and the rest of the structure is consistent with the proposed method, which is marked as proposed w/o SA. We train the Early-FM, Late-FM and the proposed w/o SA based on the same method applied in the proposed method and tabulate their test performances in **Table 2**.

Two main conclusions can be drawn from the experimental results. First, the performance results of both Early-FM and Late-FM are worse than those of the hierarchical fusion strategy we designed (i.e., the proposed method without or with SA). More concretely, the results comparison between Early-FM and proposed method are notably improved by 11.82% for SRCC, 12.18% for KRCC, and 14.18% for PLCC, while the RMSE decreased by 7.16%. For Late-FM, the proposed method also improves SRCC, KRCC, and PLCC by 9.71, 9.99, and 13.64%, respectively, while reducing RMSE by 7.08%. It is conceivable that the unnecessary noise in the early fusion will affect the quality of the fused image, and the late fusion may lose important details of the image. Thus, the obtained results are not pleasing. Second, the performance of the proposed method with SA as guidance is better than that without SA, which means that with the assistance of the SA mechanism, the process of model learning features is superior.

TABLE 2 Comparative results of ablation experiments.

Methods	SRCC	KRCC	PLCC	RMSE
Early-FM	0.7077	0.6208	0.6779	0.2425
Late-FM	0.7288	0.6427	0.6833	0.2417
Proposed w/o SA	0.7825	0.7113	0.7867	0.2020
Proposed w SA	**0.8259**	**0.7426**	**0.8197**	**0.1709**

The bold values are the results of our proposed method, which achieves the best performance.

Discussion and conclusion

Multimodal medical image fusion, as a way to express multimodal diagnostic information at the same time, has gradually gained attention in the field of medical imaging. However, the diagnostic information that a radiologist can perceive is *not only* related to the amount of initial image information contained in the fused image, *but also* to the quality of the fused image. Therefore, the quality assessment of MMIF plays an increasingly important role in the field of image processing and medical imaging diagnosis. At the same time, it has also aroused the interest of many scholars in the industry.

As MMIF is gradually gaining recognition in the medical field, quality assessment of fused images has also developed vigorously as an emerging field. An excellent objective assessment method can *not only* achieve the purpose of image quality control, *but also* guide the optimization of image fusion algorithms, so as to find the best algorithm for image fusion of different modalities. For instance, a certain algorithm can achieve very good results for image of MRI and CT, but it is not suitable for image fusion of MRI and SPECT, and maybe another algorithm should be more suitable. Unfortunately, most of existing IQA research methods are based on natural images, and it is difficult to achieve satisfactory performance for medical fusion images (see section "Comparison methods"). On the basis of previous work, we augmented the medical image database, MMIFID, which takes the doctor's MOS as the gold standard for subjective evaluation. The image content generated by $\boldsymbol{G}$ is constrained by MOS as a condition, and the non-linear mapping relationship between subjective evaluation and fused image is learned. The experimental results show that the objective evaluation results obtained from the model can match the subjective evaluation values well. In addition, compared with other IQA algorithms, we found that the proposed method outperforms the SOTA methods. Finally, we enumerate the potential limitations of this work as follows: (1) Although the database we built, as far as we know, is the largest multimodal medical image fusion database with MOS. However, it may still be a challenge for training GANs. In the future, we will continue to work on expanding the database. (2) Currently, the images contained in MMIFID are brain data, and we hope to add other body parts to the database in the future. (3) This work uses SSIM to calculate and obtain the final fusion image quality evaluation results, which may affect the accuracy of assessment to a certain extent. It would be better if the final evaluation result could also be directly assigned by GANs. Future, we will continue to explore the impact of fusing two modalities image through different methods, and design another novel IQA algorithm based on the idea of no reference.

Data availability statement

The original contributions presented in this study are included in the article; further inquiries can be directed to the corresponding author. The brain images are accessible online: https://www.med.harvard.edu/aanlib/home.html.

Author contributions

LT and HY wrote the main manuscript and contributed to the final version of the manuscript. CT and YH implemented the algorithm and conducted the experiments. YZ supervised the project and collected the data. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by the National Natural Science Foundation of China (82001912), Project Funded by China Postdoctoral Science Foundation (2018M642325), and QingLan Project of Jiangsu Province (2021).

Conflict of interest

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

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Arif, M., and Wang, G. J. (2020). Fast curvelet transform through genetic algorithm for multimodal medical image fusion. *Soft Comput.* 24, 1815–1836. doi: 10.1007/s00500-019-04011-5

Azam, M. A., Khan, K. B., Salahuddin, S., Rehman, E., Ali Khan, S., Attique Khan, M., et al. (2022). A review on multimodal medical image fusion: compendious analysis of medical modalities, multimodal databases, fusion techniques and quality metrics. *Comput. Biol. Med.* 144:105253. doi: 10.1016/j.compbiomed.2022.105253

Cai, C. L., Chen, L., Zhang, X. Y., and Gao, Z. (2020). End-to-End optimized ROI image compression. *IEEE Trans. Image Process.* 29, 3442–3457. doi: 10.1109/TIP.2019.2960869

Cauley, K. A., Hu, Y., and Fielden, S. W. (2021). Head CT: toward making full use of the information the X-rays give. *Am. J. Neuroradiol.* 42, 1362–1369. doi: 10.3174/ajnr.a7153

Chabert, S., Castro, J. S., Munoz, L., Cox, P., Riveros, R., Vielma, J., et al. (2021). Image quality assessment to emulate experts' perception in lumbar MRI using machine learning. *Appl. Sci. Basel* 11:6616. doi: 10.3390/app11146616

Chen, S., Zhao, S., and Lan, Q. (2022). Residual block based nested U-type architecture for multi-modal brain tumor image segmentation. *Front. Neurosci.* 16:832824. doi: 10.3389/fnins.2022.832824

Das, S., and Kundu, M. K. (2012). NSCT-based multimodal medical image fusion using pulse-coupled neural network and modified spatial frequency. *Med. Biol. Eng. Comput.* 50, 1105–1114. doi: 10.1007/s11517-012-0943-3

Das, S., and Kundu, M. K. (2013). A neuro-fuzzy approach for medical image fusion. *IEEE Trans. Biomed. Eng.* 60, 3347–3353. doi: 10.1109/TBME.2013.2282461

Du, J., Li, W. S., Lu, K., and Xino, B. (2016a). An overview of multi-modal medical image fusion. *Neurocomputing* 215, 3–20. doi: 10.1016/j.neucom.2015.07.160

Du, J., Li, W., Xiao, B., and Nawaz, Q. (2016b). Union Laplacian pyramid with multiple features for medical image fusion. *Neurocomputing* 194, 326–339. doi: 10.1016/j.neucom.2016.02.047

Duan, J. W., Mao, S. Q., Jin, J. W., Zhou, Z., Chen, L., and Chen, C. L. P. (2021). A novel GA-based optimized approach for regional multimodal medical image fusion with Superixel segmentation. *IEEE Access.* 9, 96353–96366. doi: 10.1109/ACCESS.2021.3094972

Gao, X. B., Lu, W., Li, X. L., and Tao, G. (2008). Wavelet-based contourlet in quality evaluation of digital images. *Neurocomputing* 72, 378–385. doi: 10.1016/j.neucom.2007.12.031

Gottesman, R. F., and Seshadri, S. (2022). Risk factors, lifestyle behaviors, and vascular brain health. *Stroke* 53, 394–403. doi: 10.1161/strokeaha.121.032610

Hossny, M., Nahavandi, S., and Creighton, D. (2008). Comments on 'Information measure for performance of image fusion. *Electron. Lett.* 44, 1066–1067. doi: 10.1049/el:20081754

Kang, L., Ye, P., Li, Y., and Doermann, D. (2014). "Convolutional neural networks for no-reference image quality assessment," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (Columbus, OH), doi: 10.1109/CVPR.2014.224

Lei, F., Li, S., Xie, S., and Liu, J. (2022). Subjective and objective quality assessment of swimming pool images. *Front. Neurosci.* 15:766762. doi: 10.3389/fnins.2021.766762

Li, C. X., and Zhu, A. (2020). Application of image fusion in diagnosis and treatment of liver cancer. *Appl. Sci.* 10:1171. doi: 10.3390/app10031171

Li, S., Kang, X., and Hu, J. (2013). Image fusion with guided filtering. *IEEE Trans. Image Process.* 22, 2864–2875. doi: 10.1109/TIP.2013.2244222

Li, X. X., Guo, X. P., Han, P. F., Wang, X., Li, H., and Luo, T. (2020). Laplacian rede composition for multimodal medical image fusion. *IEEE Trans. Instr. Meas.* 69, 6880–6890. doi: 10.1109/TIM.2020.2975405

Liu, R. S., Liu, J. Y., Jiang, Z. Y., Fan, X., and Luo, Z. (2021). A bilevel integrated model with data-driven layer ensemble for multi-modality image fusion. *IEEE Trans. Image Process.* 30, 1261–1274. doi: 10.1109/TIP.2020.3043125

Liu, Y., Chen, X., Wang, Z., Wang, Z. J., Ward, R. K., and Wang, X. (2018). Deep learning for pixel-level image fusion: recent advances and future prospects. *Inform. Fusion* 42, 158–173. doi: 10.1016/j.inffus.2017.10.007

Liu, Y., Chen, X., Ward, R. K., and Wang, Z. J. (2016). Image fusion with convolutional sparse representation. *IEEE Signal Process. Lett.* 23, 1882–1886. doi: 10.1109/LSP.2016.2618776

Liu, Y., Chen, X., Ward, R. K., and Wang, Z. J. (2019). Medical image fusion via convolutional sparsity based morphological component analysis. *IEEE Signal Process. Lett.* 26, 485–489. doi: 10.1109/LSP.2019.2895749

Liu, Y., Liu, S., and Wang, Z. (2015). A general framework for image fusion based on multi-scale transform and sparse representation. *Inform. Fusion* 24, 147–164. doi: 10.1016/j.inffus.2014.09.004

Liu, Y., Shi, Y., Mu, F., Cheng, J., Li, C., and Chen, X. (2022). Multimodal MRI volumetric data fusion with convolutional neural networks. *IEEE Trans. Inst. Meas.* 71, 1–15. doi: 10.1109/TIM.2022.3184360

Liu, Y., Wang, L., Cheng, J., Li, C., and Chen, X. (2020). Multi-focus image fusion: a Survey of the state of the art. *Inform. Fusion* 64, 71–91. doi: 10.1016/j.inffus.2020.06.013

Ma, J. Y., Xu, H., Jiang, J. J., Mei, X., and Zhan, X.-P. (2020). DDcGAN: a dual-discriminator conditional generative adversarial network for multi-resolution image fusion. *IEEE Trans. Image Process.* 29, 4980–4995. doi: 10.1109/TIP.2020.2977573

Ma, J., Tang, L., Fan, F., Huang, J., Mei, X., and Ma, Y. (2022). Swinfusion: cross-domain long-range learning for general image fusion via Swin transformer. *IEEE CAA J. Autom. Sin.* 9, 1200–1217. doi: 10.1109/JAS.2022.105686

Ma, J., Yu, W., Liang, P., Li, C., and Jiang, J. (2019). FusionGAN: a generative adversarial network for infrared and visible image fusion. *Inform. Fusion* 48, 11–26. doi: 10.1016/j.inffus.2018.09.004

Madanala, S., and Rani, K. J. (2016). "PCA-DWT based medical image fusion using non sub-sampled contourlet transform," in *Proceedings of the 2016 International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES)*, (Paralakhemundi), doi: 10.1109/SCOPES.2016.7955608

Meng, C. L., An, P., Huang, X. P., Yang, C., and Liu, D. (2020). Full reference light field image quality evaluation based on angular-spatial characteristic. *IEEE Signal Process. Lett.* 27, 525–529. doi: 10.1109/LSP.2020.2982060

Min, X. K., Zhai, G. T., Gu, K., Yang, X., and Guan, X. (2019). Objective quality evaluation of dehazed images. *IEEE Trans. Intell. Transp. Syst.* 20, 2879–2892. doi: 10.1109/TITS.2018.2868771

Preethi, S., and Aishwarya, P. (2021). An efficient wavelet-based image fusion for brain tumor detection and segmentation over PET and MRI image. *Multimedia Tools Appl.* 80, 14789–14806. doi: 10.1007/s11042-021-10538-3

Sergey, Z., and Nikos, K. (2017). "Paying more attention to attention: improving the performance of convolutional neural networks via attention transfer," in *Proceeding of the International Conference on Learning Representations (ICLR)*, (Paris).

Shen, R., Cheng, I., and Basu, A. (2013). Cross-scale coefficient selection for volumetric medical image fusion. *IEEE Trans. Biomed. Eng.* 60, 1069–1079. doi: 10.1109/TBME.2012.2211017

Shen, Y. X., Sheng, B., Fang, R. G., Li, G., Dai, L., Stolte, S., et al. (2020). Domain-invariant interpretable fundus image quality assessment. *Med. Image Anal.* 61:101654. doi: 10.1016/j.media.2020.101654

Tang, L., Qian, J., Li, L., Hu, J., and Wu, X. (2017). Multimodal medical image fusion based on discrete Tchebichef moments and pulse coupled neural network. *Int. J. Imaging Syst. Technol.* 27, 57–65. doi: 10.1002/ima.22210

Tang, L., Tian, C., and Xu, K. (2019). Exploiting quality-guided adaptive optimization for fusing multimodal medical images. *IEEE Access* 7, 96048–96059. doi: 10.1109/ACCESS.2019.2926833

Tang, L., Tian, C., Li, L., Hu, B., Yu, W., and Xu, K. (2020). Perceptual quality assessment for multimodal medical image fusion. *Signal Process.* 85:115852. doi: 10.1016/j.image.2020.115852

Townsend, D. W. (2008). Dual-modality imaging: combining anatomy and function. *J. Nuclear Med.* 49, 938–955. doi: 10.2967/jnumed.108.051276

Wang, K. P., Zheng, M. Y., Wei, H. Y., Qi, G., and Li, Y. (2020). Multi-modality medical image fusion using convolutional neural network and contrast pyramid. *Sensors* 20:2169. doi: 10.3390/s20082169

Wang, P., and Liu, B. (2008). "A novel image fusion metric based on multi-scale analysis," in *Processing of the International Conference on Signal Processing (ICSP)*, (Beijing). doi: 10.1109/TIP.2017.2745202

Wang, Q., and Shen, Y. (2004). "Performances evaluation of image fusion techniques based on nonlinear correlation measurement," in *Proceedings of the IEEE Instrumentation and Measurement Technology Conference*, (Como), doi: 10.1109/IMTC.2004.1351091

Wang, W. C., Wu, X. J., Yuan, X. H., and Gao, Z. (2020). An experiment-based review of low-light image enhancement methods. *IEEE Access* 8, 87884–87917. doi: 10.1109/ACCESS.2020.2992749

Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.* 13, 600–612. doi: 10.1109/TIP.2003.819861

Xu, H., Ma, J., Jiang, J., Guo, X., and Ling, H. (2022). U2Fusion: a unified unsupervised image fusion network. *IEEE Trans. Pattern Anal. Mach. Intell.* 44, 502–518. doi: 10.1109/TPAMI.2020.3012548

Xydeas, C. S., and Petrovic, V. S. (2000). "Objective pixel-level image fusion performance measure," in *Proceedings of the Society of Photo-Optical Instrumentation Engineers (SPIE) 4051, Sensor Fusion: Architectures, Algorithms, and Applications IV*, (Orlando, FL), doi: 10.1117/12.381668

Yang, C., Zhang, J. Q., Wang, X. R., and Liu, X. (2008). A novel similarity based quality metric for image fusion. *Inform. Fusion* 9, 156–160. doi: 10.1016/j.inffus.2006.09.001

Yin, M., Liu, X., Liu, Y., and Chen, X. (2019). Medical image fusion with parameter-adaptive pulse coupled neural network in nonsubsampled Shearlet transform domain. *IEEE Trans. Instr. Meas.* 68, 49–64. doi: 10.1109/TIM.2018.2838778

Zhao, J. Y., Laganiere, R., and Liu, Z. (2007). "Performance assessment of combinative pixel-level image fusion based on an absolute feature measurement," in *Processing of the International Conference on Innovative Computing, Information and Control (ICICIC)*, (Ottawa, ON).

Zheng, Y., Essock, E. A., Hansen, B. C., and Haun, A. M. (2007). A new metric based on extended spatial frequency and its application to DWT based fusion algorithms. *Inform. Fusion* 8, 177–192. doi: 10.1016/j.inffus.2005.04.003