Computers in Biology and Medicine 167 (2023) 107632

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

RIRGAN: An end-to-end lightweight multi-task learning method for brain MRI super-resolution and denoising

Miao Yu [a], Miaomiao Guo [a], Shuai Zhang [a], Yuefu Zhan [b], Mingkang Zhao [a], Thomas Lukasiewicz [c,d], Zhenghua Xu [a,*]

[a] State Key Laboratory of Reliability and Intelligence of Electrical Equipment, School of Health Sciences and Biomedical Engineering, Hebei University of Technology, Tianjin, China
[b] Department of Radiology, Hainan Women and Children's Medical Center, Haikou, China
[c] Institute of Logic and Computation, Vienna University of Technology, Vienna, Austria
[d] Department of Computer Science, University of Oxford, Oxford, United Kingdom

ARTICLE INFO

Keywords:
Multi-task learning
Super-resolution and denoising
Medical image analysis
Generative adversarial network

ABSTRACT

A common problem in the field of deep-learning-based low-level vision medical images is that most of the research is based on single task learning (STL), which is dedicated to solving one of the situations of low resolution or high noise. Our motivation is to design a model that can perform both SR and DN tasks simultaneously, in order to cope with the actual situation of low resolution and high noise in low-level vision medical images. By improving the existing single image super-resolution (SISR) network and introducing the idea of multi-task learning (MTL), we propose an end-to-end lightweight MTL generative adversarial network (GAN) based network using residual-in-residual-blocks (RIR-Blocks) for feature extraction, RIRGAN, which can concurrently accomplish super-resolution (SR) and denoising (DN) tasks. The generator in RIRGAN is composed of several residual groups with a long skip connection (LSC), which can help form a very deep network and enable the network to focus on learning high-frequency (HF) information. The introduction of a discriminator based on relativistic average discriminator (RaD) greatly improves the discriminator's ability and makes the generated image have more realistic details. Meanwhile, the use of hybrid loss function not only ensures that RIRGAN has the ability of MTL, but also enables RIRGAN to give a more balanced attention between quantitative evaluation of metrics and qualitative evaluation of human vision. The experimental results show that the quality of the restoration image of RIRGAN is superior to the SR and DN methods based on STL in both subjective perception and objective evaluation metrics when processing medical images with low-level vision. Our RIRGAN is more in line with the practical requirements of medical practice.

1. Introduction

Medical image, providing anatomical information to reveal structures of human body, is an indispensable component of clinical computer-aided diagnosis (CAD), which provide key points: the shape, size, kind of lesions and biomarkers to assist in lesion localization. This means that in clinical applications, high quality and high spatial resolution images are absolutely a must. While affected by many factors, such as hardware cost of image acquisition equipment, image mode, acquisition time, radiation dose, compression transmission, and so on, low-level vision medical images exist low resolution and high noise simultaneously, which directly affect the accuracy of disease diagnosis.

Single image super-resolution (SISR) is a low-level vision method aiming to get a high-resolution (HR) output from one low-resolution (LR) image to overcome hardware limitations and meet clinical requirements [1], has gained increasing research attention in the field of medical image for decades. As a notoriously challenging ill-posed problem, many different kinds of methods have been proposed to solve the SISR problem of appropriately filling in unknown extra pixels.

Traditionally, the mainstream algorithms of SISR can mainly be divided into three categories. Interpolation-based SISR methods, such as Bicubic interpolation and Bilinear interpolation, are very speedy but make the restored image blurred. Reconstruction-based SISR methods usually adopt sophisticated prior knowledge to restrict the possible solution space with the advantage of generating flexible and sharp details, but time-consuming, like the iterative back projection method [2]. Learning-based SISR methods, like manifold learning and sparse coding [3], uses a large number of training data to learn some corre-

* Corresponding author.
E-mail addresses: zs@hebut.edu.cn (S. Zhang), zhenghua.xu@hebut.edu.cn (Z. Xu).

https://doi.org/10.1016/j.compbiomed.2023.107632
Received 27 April 2023; Received in revised form 5 October 2023; Accepted 23 October 2023
Available online 27 October 2023

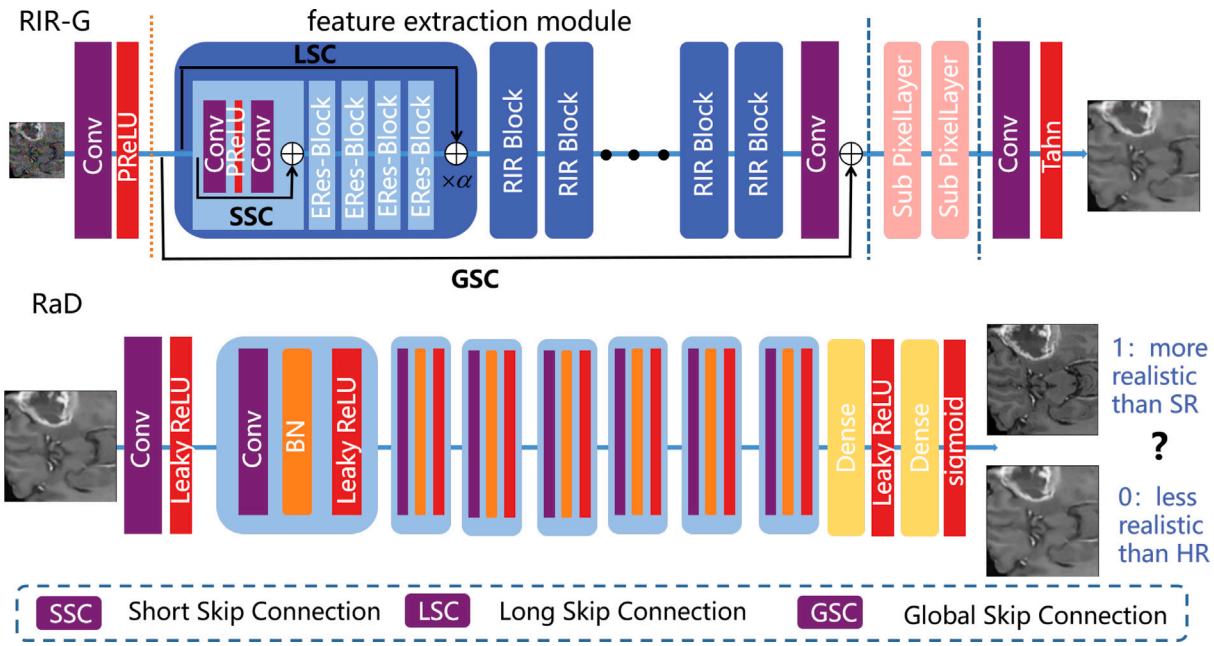


Fig. 1. Overall structure of *RIRGAN*, where the blue dotted line divides the RIR-G into three parts: feature extraction module, image amplification module and image restoration module. The combination of SSC, LSC and GSC realizes fully extraction of different hierarchical features in noisy LR inputs. And RaD tries to predict the probability that HR image (ground-truth) is relatively more realistic than output (SR image).

sponding relationship from LR-HR mapping. Since the down-sampling and degradation operations are coupled and ill-posed, traditional SISR methods cannot effectively restore some fine features and suffer from the risk of producing a blurry appearance and new artifacts. Currently, deep-learning-based SISR approaches [4,5] have been widely discussed and have led to dramatic improvements in medical image processing. With the proposal and rapid development of GAN, the GAN-based SISR methods [6,7] have made breakthroughs in human visual perception and achieved remarkable performance in various medical image modalities [8–10].

Unfortunately, there are three urgent issues in the research of low-level medical image SISR methods. The first issue is that the majority of approaches proposed to address high noise or low resolution problems in low-level medical image rely on single task learning (STL). However, in clinical settings, low-level medical images are commonly affected by both low resolution and image noise simultaneously, indicating that medical image SR and DN are not entirely independent tasks. The second issue is that the frequently used objective quantitative metrics peak signal-to-noise ratio (PSNR) and structural similarity index measure (SSIM) results are inconsistent with human subjective visual perception. Generated images with high PSNR and SSIM scores tend to be blurry, while generated images that appear to have clear details and textures typically poorly on PSNR and SSIM. The third issue is that high-performance SISR methods with complex features extraction blocks in deep or wide networks will lead to excessive network parameters and consume huge computational power [11], making it difficult to apply to some mobile devices with limited storage and computing resources [12], which is detrimental to the medical development in remote regions.

To overcome the aforementioned issues, we propose an end-to-end lightweight GAN-based method using RIR-Block for feature extraction, called RIRGAN shows in Fig. 1, which can recover potential high-frequency (HF) information and remove redundant noise information that affects diagnostic effect. Based on the characteristics of low-level vision medical images, RIRGAN combines the advantages of existing SISR models and introduces the idea of multi-task learning (MTL) [13], where super-resolution (SR) and denoising (DN) tasks share the bottom structure of the network with hard parameter sharing, and finally achieve the goal of executing both SR and DN tasks simultaneously through constraints of different loss function. The RIRGAN with perceptual realistic details pays more balanced attention to quantitative and qualitative metrics, and achieves satisfactory results in both of them by using a hybrid loss function. As a lightweight network, RIRGAN with the help of RIR-Blocks and RaD strikes a balance between effectiveness and efficiency.

Generally, the proposed RIRGAN has the following three main improvements. First, to fully extract the features of low-level images, we use RIR-Blocks to change the Res-Blocks in SRGAN and use short skip connection (SSC), long skip connection (LSC) and global skip connection (GSC) together to achieve local residual learning, regional residual learning, global residual learning and ensure the fully extraction of different hierarchical features of input image [14]. Specifically, with the assist of additional bypass connections, the model is more effective and robust for feature extraction. Thus, the model can use these features to generate more reliable details.

Furthermore, the second advancement of RIRGAN is to replace the original standard discriminator with the relativistic average discriminator (RaD) [15,16]. Specifically, the loss function of RaD estimates the probability that the HR image is relatively more realistic than the SR. This undoubtedly imposes higher requirements on the generator, which means that the generated SR image will have more realistic details.

Finally, to implement MTL for both SR and DN tasks, we train RIRGAN with a hybrid loss function which composed of pixel loss, perceptual loss, adversarial loss and total variation (TV) loss [17] through different weights. Pixel loss and perceptual loss are used to help improve the resolution of LR inputs, adversarial loss is used to guide the confrontation between generator and discriminator, TV loss and pixel loss are mainly responsible for removing noise.

Among all kinds of medical images, we choose brain MRI as our base data for three essential aspects: (i) *Software method are more economical than hardware*: The high-fidelity instruments have a great impact on MRI image quality, but powerful hardware is too expensive for radiological centers located in remote rural areas. (ii) *Image acquisition process is susceptible*: Clear MR images need long-term scanning, during which it is difficult for patients to ensure do not move slightly, but any small movement will produce artifacts that hinder doctor's

diagnosis [18]. (iii) *Published datasets are easy to link with other tasks*: As a dataset designed for segmentation task, the Multi-modal Brain Tumor Segmentation (BraTS) [19], after processed by SR model can be conveniently connected to CAD as a new foundational pre-/post-processing step such as image quality enhancement [20], segmentation [21], and lesion detect [18].

For this paper, the main contributions are as follows:

- We identify a common shortcoming of low-level vision methods: most of them are proposed for STL that can only solve one task in SR or DN at a time. This shortcoming severely limits their application in clinical practice. To alleviate this problem, in this work, we propose an end-to-end MTL method, in which SR and DN tasks share bottom structure with hard parameter sharing of the feature extraction network, and different tasks are given different attention through a hybrid loss. Ultimately, the network achieves the goal of performing both SR and DN tasks simultaneously.
- We proposed a lightweight RIR-generator (RIR-G) with SSC, LSC, and GSC, which combines local residual learning, regional residual learning, and global residual learning to better extract the different hierarchical features of low-level vision medical image, which is proved to achieve better performances in objective numerical metrics. We use RaD to further improve the perceptual quality of the output image of RIR-G. Stronger feature extraction ability and a more advanced confrontation training strategy ensure the clarity and definition of output.
- We considered both the reconstruction effect and algorithm efficiency when designing RIRGAN. RIRGAN pays more balanced attention to quantitative evaluation of metrics and qualitative evaluation of human vision. The relevant experimental results conducted on low-level vision medical images show that our RIRGAN has achieved satisfactory results.

2. Related work

Medical images, some providing anatomical information and revealing information about the structure of the human body, others providing functional information, locations of activity for specific activities and task-specific, are of great importance for medical diagnosis. The growing interest and development of single image super-resolution (SISR) algorithms dramatically influences the performance of medical image SISR tasks. Different from nature image SISR tasks, medical images, in general, have a lower signal-to-noise ratio, and the SR task on medical images usually needs to be pipelined by applications such as segmentation, classification and diagnosis, thus placing higher demands on how to retain sensitive information in image and enhance structures of interests (focusing lesions and their surrounding tissues).

The computer vision community has investigated many super-resolution (SR) and denoising (DN) methods to address low-level vision. Deep learning aims to extract high-level abstract features and learn potential distribution law of data through multi-layer nonlinear transformations. The studies of deep learning have led to dramatic improvements in SR and DN tasks.

In this paper, we propose an end-to-end lightweight multi-task learning (MTL) network for SR and DN, which is improved on the basis of SISR methods. Meanwhile, the SR task is the main task of our research. Therefore, the following content will be primarily discussed based on SISR.

As an important branch of low-level vision, SISR is widely used to improve medical image quality. We can view the process of SISR as a model for predicting one HR image from its LR counterpart. A key step in this process is to acquire the features of the LR images better and faster. SRCNN [22] does the pioneering work of deep-learning-based SISR methods, and Umehara et al. [23] use SRCNN scheme to enhance image resolution of CT images. However, due to the fact that the SRCNN framework only has three convolutional layers, it can only extract low-frequency (LF) features of inputs. The advent of SRGAN has allowed the SISR methods to focus on improving the quality of human vision rather than just pursuing high PSNR and SSIM scores [24]. Some medical image SISR methods have been improved based on SRGAN [25–27]. Despite achieving some successes, SRGAN-based medical image SISR methods still have some drawbacks, such as an insufficient feature utilization, numerous parameters. To further enhance the visual quality, we thoroughly studied three key components of SRGAN, i.e., network architecture, confrontation training, and loss function, and thus propose RIRGAN in this work.

Improving feature extraction module. The first improvement of RIRGAN is to change the feature extraction module by RIR-Block instead of Res-Block [29] in SRGAN. In recent years, there are many deep-learning-based SISR methods trying to improve the network performance by changing the feature extraction module. Chen et al. use Dense Block to further improve the feature extraction ability and realize the 3D multi-level super-resolution (SR) of MRI [30]. Wang et al. introduce Residual in Residual Dense Block (RRDB) in SRGAN to fully extract features of low-resolution (LR) image [9]. Zhu et al. [31] replace Res-Block in SRGAN with Enhanced Residual Block in EDSR [4], and the main difference is that removes the BN layer and increase residual scale. A fine inpainting method based on feature fusion and two-steps inpainting is proposed to overcome the problem of the existing image inpainting methods that cannot make full use of complete region to predict missing region features [32]. An image restoration method combining Semantic Priors and Deep Attention Residual Group is also proposed to solve the problems that the image inpainting methods lacking authenticity [33]. To some extent, these basic blocks are all based on the idea of residuals, but more complex bypass skip connections are added, as shown in Fig. 2. Complex feature extraction block means a large amount of parameters and too much computation. While compared to the common SOTA SISR networks, RIRGAN is a lightweight network with much smaller number of parameters as shown in Table 1.

Additional paths ensure that gradients can be transferred to each layer more effectively during back-propagation and prevent gradient vanishing. Thus, the model can easily achieve a very good performance, especially for minimizing pixel loss. However, the structure of dense connections also made the model liable to getting stuck at certain points, and insensitive to uncertain losses, such as GANs [31]. Second, compared with natural images, medical images with limited size and relatively lower contrast information, too many feature maps are often overqualified, so medical images do not need wide models. Therefore, although we change the feature extraction basic block in SRGAN, we did not choose the overly complex bypass connection module, but RIR-Block. Under the combined action of SSC, LSC and GSC, the RIR-generator (RIR-G) ensures the extraction of different hierarchical features of low-level vision medical image through local residual learning, regional residual learning, and global residual learning without imposing too much burden on the network.

Improving confrontation training of GAN. In addition to improving the structure of the generator, we also enhance the discriminator to change the confrontation training strategy. As we all know, GAN is good at generating images with rich details, but is hard to train. The standard discriminator in SRGAN estimates the probability of whether one input is real or fake, which makes SRAGN suffer from unstable and collapse mode that can affect the SISR results [24]. Many researches hope to find a way to improve the confrontation between the generator and the discriminator to get a higher quality generated image. Zhu et al. use WGAN with gradient penalty to achieve stabilized and efficient training and improved perceptual results [18]. Liang et al. add Gaussian noise to the input of discriminator to increase the difficulty of the discriminant task [34]. To generate more perceptually realistic images, Zhu et al. apply WGAN-based adversarial training [31]. Chen et al. propose an RNON compose of two independent GANs, and the two

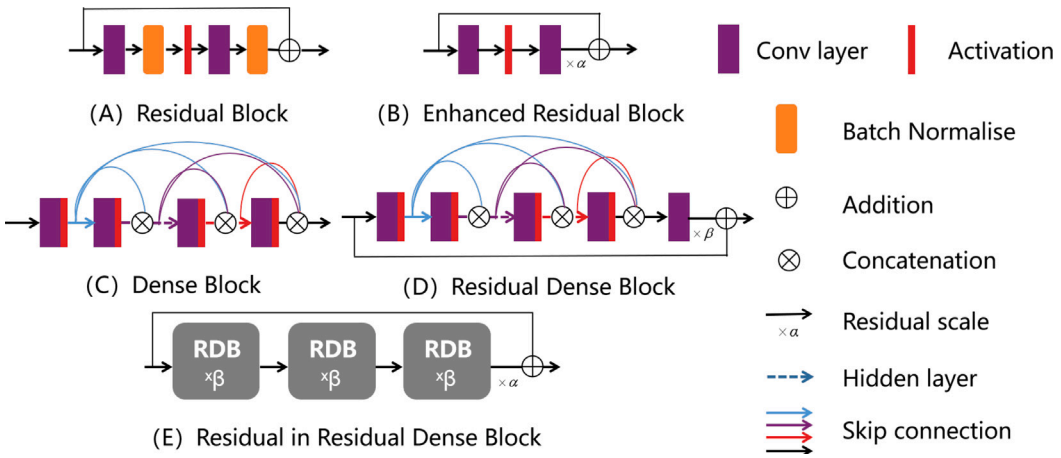


Fig. 2. Residual-based feature extraction basic block. These blocks are the basic unit of SOTA SISR networks: (A) Residual Block is used in SRGAN [24]; (B) Enhanced Residual Block is used in EDSR [4]; (C) Dense Block is used in SRDenseNet [28]; (D) Residual Dense Block is used in RDN [5]; (E) Residual in Residual Dense Block is used in ESRGAN [16].

Table 1
Comparison of parameters between residual-based SOTA SISR method with complex bypass connections and RIRGAN.

Methods	MDSR [36]	SRGAN [24]	EDSR [4]	SRDensNet [28]
Params	6.7M	4.5M	43M	30.4M
Methods	RCAN [14]	ESRGAN [16]	RDN [5]	RIRGAN
Params	16M	30.2M	17.2M	3.3M

GANs are responsible for solving the problem of repairing irregular missing regions and the local color difference respectively [35].

We choose to replace standard discriminator in SRGAN with RaD using in Relativistic average GAN [15]. The RaD can judge whether one image is more realistic than the other image [16]. Our experiments show that the use of RaD can guide the generator to recover more realistic texture details with perceptual quality and boost the performance.

Improving learning methods of task. Third, for better application in the real world, we introduce the idea of MTL [13,37] so that RIRGAN can learn SR and DN tasks concurrently during training. Most medical images low-level vision methods based on deep learning are Single Task Learning (STL), for SR, or for DN. Yan et al. present a SR algorithm for SPECT reconstruction with compensation for non-uniform attention [38]. Mahapatral et al. introduce the triplet loss to SRGAN to realize cardiac MRI SISR, and achieve good results in big scaling factor [8]. Zhang et al. develop a CT SR method which can reconstruct LR sonograms into HR CT images [39]. In low-level vision of medical images, since people usually study SR and DN as two independent tasks, this leaves a lot of noise after SR processing. Although Zhang et al. mention that the image noise level and resolution of SPECT images are relatively poor, but they only use GAN in static SPECT image denoising [40]. We notice that the above works only focus on one problem in low-level vision medical image, which have poor clinical applicability, so there is still a lot of noise after SR processing.

Chen et al. use transformer architecture based image processing transformer (IPT) model to train on huge datasets (over 10 millions of images), then fineturned on specific small datasets, and finally achieved good results in SR, DN and deraining [41]. While our RIRGAN uses the share bottom structure, i.e., the SR and DN tasks share input and feature extraction layers, and hard parameter sharing, then use different losses to constrain the outputs of two tasks. The idea of MTL greatly improves the efficiency of the network, and ensures that RIRGAN can establish a nonlinear end-to-end mappings from noisy LR input to denoising and deblurring output, and has advantages in handling complex-problems.

3. Methods

Fig. 1 shows the overall structure of RIRGAN. In contrast to SRGAN, RIRGAN mainly consists of three advanced modules: RIR-G based on RIR-Block feature extraction module, RaD used to assist in adversarial training, and hybrid loss function for Multi-task learning (MTL). Specifically, several enhanced residual blocks composed RIR-Block with short skip connection (SSC) and long skip connection (LSC), which can help form a very deep network and enable the network to focus on learning of HF information. Global skip connection (GSC) connects low-frequency (LF) and high-frequency (HF) features together to prevent feature loss of input images. The RIR-G with SSC, LSC and GSC ensures that the network has the ability to extract multi-hierarchical features of noisy low-resolution (LR) input through local residual learning, regional residual learning, and global residual learning. RaD estimates the probability that high-resolution (HR) image is more realistic than super-resolution (SR) image. The use of new confrontation strategy can push the RIR-G to learn an inter-domain mapping and produce compelling rich detailed images and boost the performance in PSNR and SSIM. To achieve MTL of low-level brain MRI, we designed a hybrid loss function, which gives different constraints to the features extracted by RIR-G through different loss function. It also enables RIRGAN to give more balanced attention between quantitative and qualitative evaluations, and makes RIRGAN in line with the actual requirements of medicine.

3.1. Problem statement

Single image super-resolution. SISR aims to restore an HR image from one LR observation of the same object. The LR image y can be modeled as:

$$y = (x \otimes k) \downarrow_s + n, \tag{1}$$

Where x denotes the unknown HR image, k is the blurry kernel, and $x \otimes k$ is the convolution between both of them. $\downarrow$ is the down-sampling operator with scale factor s, and n represents the possible independent noise term. In the SR task, the size of the output image is enlarged relative to the input image. This is an extremely ill-posed problem for we do not know what exactly is the part of an image that needs to be completed.

Image denoising. The goal of image denoising (DN) task is to remove redundant noise information which pollutes input and restore the potentially clean image. Taking the most common additive noise as an example, the noise image can be simplified as:

$$y = x + n, \tag{2}$$

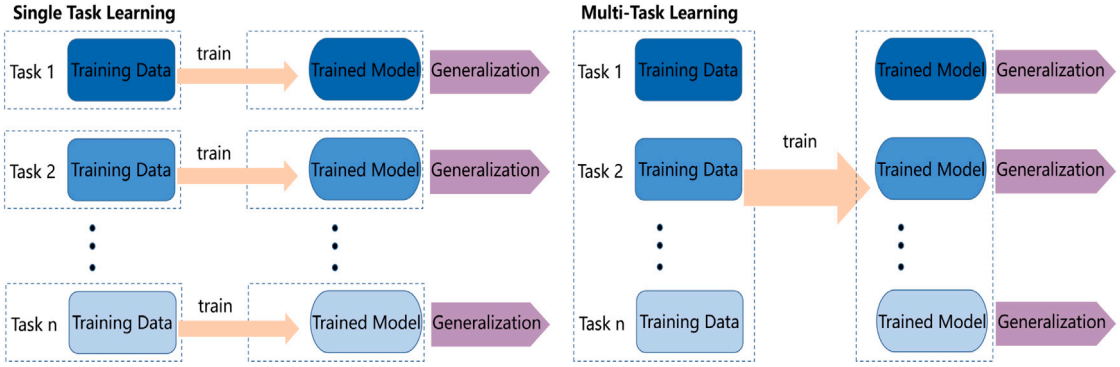


Fig. 3. Differences between STL and MTL when handling complex-task. STL decomposes complex-task into several sub-tasks and uses multiple networks to train each sub-task separately. In MTL, due to the inherent correlation between different sub-tasks, only one network needs to be trained to complete multiple tasks.

where x, y and n represent the clean image, noisy input and noise respectively. In the DN task, the size of the output image is consistent with the input image.

Our goal is to improve a noisy low-resolution (LR) brain MRI to a clean and high quality one. The main challenges are listed as follows. First, compare with natural images, noisy LR image in low-level medical vision contain more complex spatial variations and correlations, which limit the performance of traditional SISR methods. Second, the image with noise and artifact patterns creates difficulties for algorithms to produce a perfect output. Finally, up-sampling and degradation operations are coupled and ill-posed. Traditional SISR methods cannot be performed beyond a marginal degree, which cannot restore fine features effectively and suffer from the producing blurry and new artifacts [42]. To address these limitations, we introduced the idea of multi-task learning (MTL).

Multi-task learning. At present, most deep-learning-based models are single task learning (STL), i.e., training a model can only complete one task. When handling complex-tasks, the usual method is to decompose the complex-task into simple and independent sub-tasks to solve them separately, and then combine the results of the sub-tasks to get the results of the original complex-task [43,44]. In MTL, the main task uses domain specific information possessed by the training signals of related auxiliary tasks as an inductive bias. MTL model can learn multiple related tasks learn in parallel through the share bottom representation, and the gradients are simultaneously back-propagated to improve the generalization performance of main task. When handling complex tasks, the differences between STL and MTL are shown in Fig. 3.

Compared with STL, the advantages of MTL [13] are: (i) *High efficiency*: The basis that a network can learn multiple tasks is that all the sub-tasks have an inherent correlation. Multiple related tasks can be completed in one training, which hugely improves the efficiency of the network. (ii) *Strong model generalization ability*: In the learning process, a shared representation is used to share and supplement the domain information learned from each other, for the sake of promoting learning and improving the generalization effect. (iii) *Difficult to sink into local minima*: In STL, the backpropagation of gradient is easy to fall into local minima. While in MTL, local minima of different tasks are in different positions, which can help to implicitly escape local minima in the backpropagation of gradient. (iv) *Difficult to over-fitting*: Multiple tasks in shallow shared representation can weaken the ability of network and reduce the occurrence of network over-fitting.

Through the introduction above, we know that the focus of the SR task is to complete unknown pixel information, and the key point of the DN task is to removal redundant noise information, shown in Fig. 4. “complete” and “removal” seem to be two completely opposite words, but we have found the inherent correlation of two tasks in deep learning, i.e., feature extraction of high frequency details of input image, which is the basis for us to design an MTL network that can simultaneously complete SR and DN tasks.

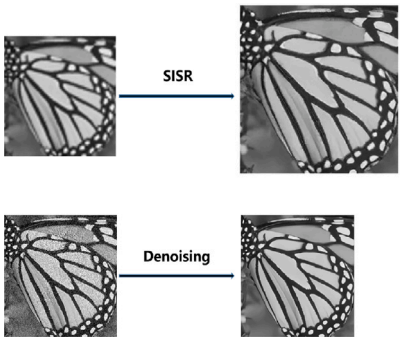


Fig. 4. Differences between SR (SISR) and DN tasks. The size of the output image in the SR task has enlarged compared to the input image, adding a lot of pixel information that was not present in the original input image. In the DN task, the output image has the same size as the input image, reducing redundant noise information in the input image.

Our RIRGAN adopts share bottom structure with hard parameter sharing in MTL, two tasks, SR and DN share the bottom network structure of input layers and feature extraction layers, and network parameters, then use different loss functions to constrain the output of two tasks, which ensures our RIRGAN is an end-to-end model. Put it differently, our RIRGAN implements the sharing of network parameters in the form of shared features. In our model, the dominant main task is the SR task, and the DN task is the related auxiliary task as an inductive bias for SR task. We use $I_{l_{r+n}}$ to represent the input LR image with random Gaussian noise to simulate low-level visual medical images, and I_{mtl} represents the output of RIRGAN.

$$I_{mtl} = G(I_{l_{r+n}}, s; \theta_G), \tag{3}$$

Where G is the RIR-G and θ_G denotes its trainable weights. And then we use L_{mtl} represents the loss function between the image generated by G and the ground-truth I_{gt}. Through backpropagation we can calculate gradients and update weights θ_G:

$$\hat{\theta_G} = \arg\min_{\theta_G} L_{mtl}(G(I_{l_{r+n}}), I_{gt}). \tag{4}$$

3.2. Lightweight RIR-G

To fully extract the multi-hierarchical features of noisy LR input, the first improvement of RIRGAN is to change the basic block of the feature extraction module in SRGAN. Specifically, RIR-G is composed of three parts: feature extraction module, image amplification module,

and image restoration module. The feature extraction module first uses a Conv layer and a PReLU layer to extract low-frequency (LF) information of input, which is shown on the left side of the red dotted line in Fig. 1. While extracting high-frequency (HF) information, different from SRGAN, we change the Res-Block by RIR-Block [14]. In RIR-Block, five Enhanced Residual Blocks (which shows in Fig. 2, and short for ERes-Block in Fig. 1) with LSC and residual scale α form the RIR-Block. Several RIR-Block groups form a very deep network, which enables the network to focus on learning HF information. Then we use GSC fuse LF and HF information, which ensures blending different hierarchical features together. To realize the amplification with scale factor 4, we connect two sub-pixel layers in series. And the image restoration module includes a convolution and a tanh layer and finally outputs the SR image.

The advantages of using RIR-Block based RIR-G to extract features of noisy LR input are threefold:

(i) RIR-G removes the BN layer. BN layer will normalize the feature and this normalization can accelerate the speed and reduce the difficulty of network training, which is conducive to detection [45] and segmentation [46,47] in medical imaging. However, the normalization also changes original image information to a certain extent and makes the network lose flexibility. As classic low-level vision regression tasks, SR and DN require various features to provide quality support for output images. While the authenticity and accuracy of extracted features are the key point to ensure the reliability of output. The feature normalization caused by the BN layer is disadvantageous for SR, DN, and image generation tasks (it may bring extra artifacts).

(ii) RIR-G uses more bypass connection path. The skip connection used in SRGAN ensures local feature fusion to achieve local residual learning. While in RIRGAN, we use RIR-Block based RIR-G instead the Res-Block based generator by increasing SSC, LSC and GSC. The SSC has the same function as the skip connection in SRGAN to realize the local residual learning. The newly added bypass connections LSC enables the network to achieve regional feature fusion by regional residual learning and helps to stabilize the training of the network. After obtaining multi-level region features, we use GSC realize global residual learning to achieve global feature fusion, which adaptively preserves the hierarchical features globally. In other words, our RIR-G realized the extraction of different hierarchical features of noisy LR input and RIR-Block optimizes feature transmission between neural network layers and enhances the diversity and quality of the generated image. From the perspective of forward propagation, SSC, LSC and GSC together accumulate the extracted hierarchical features, increase the stability, and help prevent training saturation and over-fitting.

(iii) RIR-G is more inline with medical image. We do not choose the other extraction blocks in Fig. 2 with too complex and that much bypass connections [5,16,28] for: (1) complex feature extraction block means a large amount of parameters and too much computation, (2) training a wider network with dense blocks would become complicated, (3) the structure of dense connections also made the model liable to getting stuck at certain points, and made the model insensitive to uncertain losses such as GANs [31]. In practical medical applications, too deep and so wide models are unnecessary. RIRGAN is a lightweight network only have 3.3M parameters, which is much smaller than other SOTA SISR method with complex bypass connections shown in Table 1. The lightweight model has obvious advantages in the research of some mobile devices with limited storage and computing resources [11,12], which is beneficial to the medical development of underdeveloped remote areas. And a number of experiments on some benchmarks have proved that our proposed RIRGAN can achieve comparable performance against the SOTA with a more lightweight model.

The formal process of RIR-G can be written as follows. First, by defining the input image as $I_{l_{r+n}}$, the output feature of the LF feature extraction module is:

$$F_0 = H_{LF}(I_{l_{r+n}}), \tag{5}$$

Where $H_{LF}(\bullet)$ denotes convolutional operation. And F_0 is the extracted LF feature. It is not only the output of the LF feature extraction network but also the input of the HF feature extraction network composed of N RIR-Blocks. The output of the HF feature extraction network we call it as:

$$F_{HF} = H_{RIR}(F_0), \tag{6}$$

Where $H_{RIR}(\bullet)$ denotes RIR-Blocks. To prevent the sharp increase in network parameters and limit the application in practical, we set the number of RIR-Blocks in generator to $N = 8$.

We made a global feature fusion of LF feature F_0 and HF feature F_{HF} to adaptively preserve the hierarchical features globally through GSC, then get

$$F_{GF} = F_0 + F'_{HF}. \tag{7}$$

Finally, the extracted feature maps of the input through image amplification module A and image restoration module can get

$$I_{mtl} = G(I_{l_{r+n}}) = A(F_{GF}(I_{l_{r+n}}), s). \tag{8}$$

Algorithm 1 The algorithm of RIR-G for noisy LR MR image

Input: noisy LR MR image: $I_{l_{r+n}}$
Output: restored MR image: I_{mtl}
1: $I_{l_{l+n}}$ enters into RIR-G;
2: // Stage 1. feature extraction
3: Extract LF features based on Eq. (5);
4: Extract HF features based on Eq. (6);
5: Fusion of hierarchical features based on Eq. (7);
6: // Stage2. amplification
7: Use 2 sub-pixel layers to amplify the F_{GF} by $s = 4$;
8: // Stage 3. restoration
9: Obtain I_{mtl} through image restoration module.

3.3. RaD

In RIRGAN, we use RaD to push the generator to learn an inter-domain mapping and produce compelling targets. Generally, the discriminator predicts the probability that the input image is real or fake [24], which makes GAN very difficult to train. RaD tries to estimate the probability that a real image (I_{gt} or HR image) is relatively more realistic than a fake one (I_{mtl} or SR image), as shown in Fig. 1.

Specifically, our discrimination loss function becomes

$$\begin{aligned} L_D^{Ra} = &- E_{I_{gt}}[\log(D_{Ra}(I_{gt}, I_{mtl}))] \\ &- E_{I_{mtl}}[\log(1 - D_{Ra}(I_{mtl}, I_{gt}))], \end{aligned} \tag{9}$$

Where I_{gt} represents the ground-truth, $E_{I_{mtl}}[\bullet]$ represents the average of all fake data in the mini-batch.

The corresponding loss function of our generator should also be changed to:

$$\begin{aligned} L_G^{Ra} = &- E_{I_{gt}}[\log(1 - D_{Ra}(I_{gt}, I_{mtl}))] \\ &- E_{I_{mtl}}[\log(D_{Ra}(I_{mtl}, I_{gt}))]. \end{aligned} \tag{10}$$

Benefits from the gradients from both generated SR image and real HR image in adversarial training, our RIRGAN can learn more realistic details with high quality.

3.4. Hybrid loss function

Although the features of noisy LR brain MRI have been fully extracted, and high-quality output can be generated thanks to RaD, now RIRGAN can only achieve STL. To realize MTL, the features extracted by the feature extraction module need to be constrained by different loss functions, that is why we proposed a hybrid loss function. SR and DN tasks have the same loss functions, but also have different loss functions. We will introduce them in detail in the following content.

Pixel loss. Pixel loss is the most common loss function in low-level vision, which can be used for both SR and DN tasks. It directly compares each pixel value of the output image get by network processing with the corresponding pixel in ground-truth. Pixel loss prefers to encourage blurry and encourage network to find an average of many plausible solutions and lead to over-smooth results. We choose L1 loss as our pixel loss, which is also known as the mean absolute error (MAE). The goal of L1 loss is to minimize the absolute sum of square of the difference between ground-truth and output,

$$\mathcal{L}_{pix} = \mathcal{L}1 = \sum_{i=1}^{n} \left| I_{gt} - I_{mtl} \right|. \tag{11}$$

Compared with MSE loss, L1 loss is more stable for outliers, which can add constraints to network and help the network converge quickly. The image trained by L1 loss has good numerical performance on the metrics of PSNR.

Perceptual loss. Owing to the signal aliasing in LR inputs, HR images are difficult to be faithfully restored. Perceptual loss is proposed to enhance visual quality by minimizing error in a feature space instead of pixel space and making the resulting image more semantically similar to target image. The formula for perceptual loss is as follows:

$$\mathcal{L}_{per}(I_{gt}, I_{mtl}) = E(|\mathcal{V}_l(I_{gt}) - \mathcal{V}_l(I_{mtl})|^2), \tag{12}$$

where $\mathcal{V}$ is a pre-trained VGG19 model and l denotes the feature maps of the specific layer of $\mathcal{V}$.

Perceptual loss is applied in both SR and DN tasks to enhance the texture of a restored image. Applying perceptual loss in GAN can generate images with more natural details, partly reducing visually unpleasant artifacts, especially when handling fine-scale details images. Although the results after perceptual loss training have improved perceptually realistic and were more in line with human visual requirements, while perceptual loss cause the pixel-wise differences to the ground-truth, so it cannot achieve optimal results on metrics of RSNR and SSIM [48].

Adversarial loss. Recently, GAN become popular to hallucinate details. GAN consists of a generator and a discriminator. The discriminator aims to distinguish generated fake images from real ground-truth images, while the generator aims to fool the discriminator. Adversarial loss is the loss function used to balance generator and discriminator in GAN. The discriminator and generator are constantly confronting each other under the effect of adversarial loss, and finally guide the generator, generating images with natural details. Put it another way, an adversarial loss is applied to distinguish ground-truth or generated one. Thus, the basic adversarial loss function is defined as:

$$\mathcal{L}_{adv} = -E_{I_{gt}}[\log D(I_{gt})] - E_{I_{lr+n}}[\log(1 - D(G(I_{lr+n})))], \tag{13}$$

where I_{lr+n} denotes the low-level vision input with low-resolution and noise in our task.

GAN is good at generation task of "creating something out of nothing", which makes it very suitable for ill-posed low-level vision task like SR and DN. Why does GAN works in our task? In fact, GAN is equivalent to letting the model learn what is noise information and what is useful information for SR, i.e., dividing the original visual feature space into noise space and useful space for SR, so that the model not fit the noise space when doing SR task [49]. To generate more perceptually realistic images we choose to use RaD, and the detailed formula can be found in Section 3.3.

Total variation loss. The total variation (TV) of an image contaminated by noise is significantly larger than that of an image without noise, so limiting the TV will limit the noise of image. The basic idea of TV denoising is that if the details of an image have a lot of HF information (such as spikes, noise, etc.), the sum of gradient amplitudes (TV) of the entire image can be reduced, the difference between adjacent pixel values in the image can be reduced by minimizing TV, thus achieving the goal of DN [50]. By defining an SR image as x, TV is the sum of gradients in the pixel domain:

$$\mathcal{L}_{tv}(x) = \sum_{i,j} |x_{i+1,j} - x_{i,j}| + |x_{i,j+1} - x_{i,j}|, \tag{14}$$

where $x_{i+1,j}$ and $x_{i,j+1}$ is the adjacent pixels of the pixel $x_{i,j}$ in the given SR image. The advantage of TV denoising is that it can remove noise while preserving information such as boundaries in the image.

In the process of low-level vision, a little noise on input may have a great impact on output, because usually many low-level vision algorithms will amplify noise inevitably. At this time, it is necessary to add some regularization items to maintain image smoothness. TV loss is such a good regularization loss, which is capable in promoting the image's spatial smoothness and reducing noise [17].

Hybrid loss function. In MTL, the role of different losses in different tasks is usually different. Among them, pixel loss and perceptual loss are good at improving the resolution of output. Adversarial loss is used to guide the confrontation between generator and discriminator, and make sure the output with more natural details. TV loss and pixel loss play a major role in noise removing. The existence of TV loss guarantees that RIRGAN can remove noise when doing SR task. This makes it unnecessary to train SR and DN separately as two independent tasks.

If we simply add up the losses of different tasks, it may lead to MTL being dominated by a certain task or deviation of the overall learning task. The model may tend to fit the main task it thinks, and the effect of other tasks will be negatively affected, resulting in poor final results. To train the model in the desired direction, we configure a fixed weight parameter for each loss and combine multiple loss functions into a hybrid loss function.

In the process of training, we find that the TV loss converged very quickly. That is to say, in RIRGAN, DN is a relatively simpler auxiliary task, so the weight of TV loss should be much smaller than the weight for SR tasks according to Liu et al. [37]. At the same time, the existence of TV loss will make the output of the network over smooth, which is inconsistent with the purpose of our main task, SR task. Therefore, we hope to reduce the impact of TV loss on SR task as much as possible on the premise that it can complete the DN task, which makes RIRGAN achieve the effect of balancing performance of the overall network.

Our hybrid loss function is:

$$\mathcal{L}_{mtl} = \lambda \times \mathcal{L}_{pix} + \gamma \times \mathcal{L}_{per} + \beta \times \mathcal{L}_{tv} + \eta \times \mathcal{L}_{adv}, \tag{15}$$

where the hybrid loss function is a weighted sum of these loss terms, with the weights λ, γ, β, and η controlling the relative importance of each term.

4. Experiments

4.1. Dataset and preprocessing

We pick the Multi-modal Brain Tumor Segmentation (BraTS) dataset [19] as our experimental dataset, which provides MRI scans of 210 patients with glioblastoma (HGG) and 75 patients with lower grade glioma (LGG). BraTS is a multi-modal dataset, which constrains 4 versions of brain MRI scans, including native (T1), contrasted enhanced T1-weighted (T1ce), T2-weighted (T2), and T2 Fluid Attenuated Inversion Recovery (T2-FLAIR) volumes, and serves as a good proxy for medical images. We picked the T1ce version as our training images, because the images of this version are the most complete for the brain, which is suitable for our experimental requirements compared with the other three versions.

BraTS is specially designed for segmentation task, where the shape of each 3D CT volume is 240 × 240 × 155, i.e., containing 155 2D image slices with the size of 240 × 240. However, only the middle parts of these 2D image slices contain useful information (i.e. the brain), while the rest is purely black and useless. Therefore, in order to help

the model better learn the image features, "center-crop" operation is conducted in pre-processing to remove black areas at the edge of the image and retain only useful information in the middle of the image. Consequently, in training stage, HR ground-truth slices are cropped to the size of 96 × 96 to maximize the elimination of useless black areas (and the size of SR output images in training are also set as 96 × 96). Differently, in testing stage, the HR medical images are cropped to a larger size, 128 × 128, which is because of the following reasons: (i) Most of the existing single image super resolution (SISR) works [8,25,31] adopt the image size of 128 × 128 in their experiments, so we follow the same setting to keep fair comparison. (ii) More importantly, in real-world clinical use, medical images are of diverse sizes, so we cannot guarantee that the important information of each image always appears in a central area with a fixed size, e.g., 96 × 96; so, to retain all useful information, we tend to be more conservative with image cropping in practical usage, which inevitably make the images still contain some black edge areas; therefore, we believe using a larger center-crop size of 128 × 128 to include a certain extent of black areas in the testing data will make our experiments more practical.

In addition, in the SISR task, low-level vision medical images are also needed, which should have exactly the same content as the high-level vision medical images but with lower resolution. However, in clinical practices, it is almost impossible for any medical imaging device to obtain paired real high-resolution and low-resolution images simultaneously at the same time; and if the high-resolution and low-resolution images are obtained from different devices (or) at different time, due to the change of the patients' postures, the resulting medical images will have different content; and even small differences in the paired high-resolution and low-resolution images will bring significant bias to the evaluation of model performances. Consequently, a widely adopted approach in the field of SISR is to degrade high quality medical to simulate low-level medical images [8,25,31]. Similarly, as for the denoising tasks, it is also impossible to obtain a pair of real high-noise and low-noise medical images with exactly the same content simultaneously using any medical imaging device, so the existing denoising works [8,51,52] usually use additional artificial noises to simulate the high noise images. Therefore, in this work, we follow the same operations as in [8,25,31] to use a down-sampling factor of $s = 4$ to obtain low-resolution medical images (with the sizes of 24 × 24 in training and 32 × 32 in testing) from HR medical images, and then Gaussian noises are further added as in [8,51,52] to generate low-level vision medical images with not only low-resolution but also high-noise.

Please note that although due to the above fact that it is impossible to obtain paired real high-level and low-level medical images with the same content, we were certainly unable to quantitatively compare the similarity between simulated low-level medical images and real low-level medical images, we still invited four radiologists with more than five years' clinical experience from the Department of Radiology, Hainan Women and Children's Medical Center, China, to subjectively evaluate the rationality of using the simulated low-level medical images. Based on their clinical experience and expertise, all four doctors believe that these simulated low-level CT images show characteristics that are relatively similar to the real low-level CT images in their clinical practices, so it is reasonable to use them to simulate the real ones in the experimental studies. The workflows of training and testing are shown in Fig. 5.

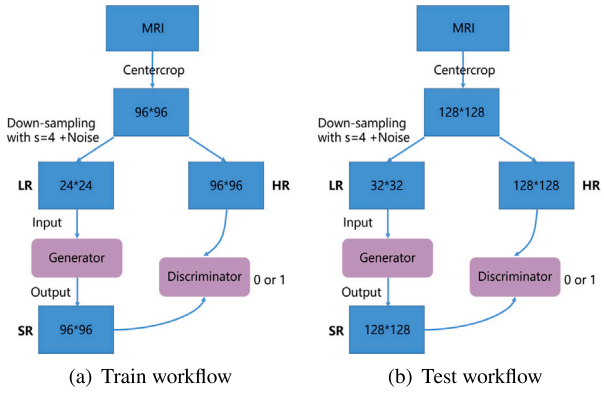


Fig. 5. Workflows during training and testing process when RIRGAN handles BraTS, for a ×4 up-sampling task.

4.2. Baselines

Our RIRGAN is proposed through modification based on the SISR method, and SR task is the main task in our MTL model. In order to evaluate the performances of the proposed RIRGAN, we chose four classical and representative SISR methods: Bicubic, SRResNet, SRGAN [24] and RDN [5]. The reasons for selecting these methods as the baselines are as follows. (i) Bicubic is the most common interpolation amplification method in practical applications, and we also choose it as our image degradation method. (ii) SRResNet is a residual block based perception-driven method without using GAN. (iii) SRGAN is arguably the most widely adopted GAN-based SISR model, and it is also used as the backbone of the proposed RIRGAN. (iv) RDN is selected as the representative of PSNR-oriented method, which is the state-of-the-art deep learning based end-to-end SISR model with Residual Dense Blocks (8 blocks are used here). (v) Our RIRGAN is a lightweight residual learning based network, and for the sake of fairness, we did not choose the SR methods with too large parameters as our baseline.

4.3. Implementation settings

Our experiments are implemented using the PyTorch framework[1] and run on two NVIDIA GeForce GTX 2080Ti GPUs. The implementation details of the proposed RIRGAN are shown as follow. Each RIR-Block used in RIR-G contains 5 Enhanced Residual Blocks with SSC, which form a residual group through LSC and s residual scale $a = 0.2$. And the number of RIR-Blocks in RIR-G is $N = 8$. Layers in LF feature extraction and HF feature extraction module have $w_0 = 64$ filters. The Sub-Pixel layers in image amplification module have $w_1 = 256$ filters. And the last convolutional layer in image restoration module has $w_2 = 3$ filters. In RIR-G, expect for the convolutional kernel size of first and last convolutional layers are 9 × 9, other convolutional layers are all with kernel size 3 × 3. Our RIRGAN follows the structure of the discriminator in SRGAN [24], where the numbers of feature maps are 64, 64, 128, 128, 256, 256, 512, 512.

We trained the RIRGAN on the pre-processed BraTS slices for 200k update iterations, with a mini-batch size of $n = 16$ using Adam optimizer with parameters $\beta_1 = 0.5$, $\beta_2 = 0.999$, a learning rate of 10^{-4} of the generator and a learning rate of 10^{-6} of the discriminator. The specific weights in the hybrid loss function specifically designed for MTL in Eq. (15) are $\lambda = 0.006$, $\gamma = 1$, $\beta = 2 \times 10^{-9}$, and $\eta = 5 \times 10^{-3}$ respectively. From the composition of the hybrid loss function, we can also see that the SR task is the dominant main task and the DN task is the related auxiliary task.

4.4. Evaluation metrics

To evaluate the performances of our proposed RIRGAN and other SOTA methods, we selected two kinds of widely used objective evaluation metrics.

Image quality evaluation. PSNR is an error sensitive image quality evaluation metric based on the statistics of image pixel information, and is the most widely used evaluation criteria for SR and DN tasks.

[1] https://pytorch.org/

Table 2
Quantitative multi-task learning results of the proposed RIRGAN and the baselines in terms of numerical objective metrics. ↑ represents the larger the better, ↓ represents the smaller the better, – means that we are not concerned about the size of this value, but we still marked the maximum one. The best results of each row are marked bold.

Metric		Bicubic	SRResNet	SRGAN	RDN	RIRGAN
PSNR	Mean↑	25.66	24.91	25.05	28.24	**28.26**
	Var↓	5.691	10.30	14.26	17.34	**3.413**
	Std↓	2.386	3.210	3.776	4.165	**1.847**
	Min–	19.10	17.52	16.98	15.07	**21.12**
	Max–	30.18	30.92	33.21	**34.14**	31.62
	$\Delta\downarrow$	11.08	13.40	16.23	19.07	**10.50**
SSIM	Mean↑	0.7893	0.7301	0.8294	0.8661	**0.8833**
	Var↓	0.0007	0.0007	0.0010	0.0031	**0.0005**
	Std↓	0.0268	0.0262	0.0313	0.0552	**0.0213**
	Min–	0.6977	0.6455	0.7004	0.6062	**0.8129**
	Max–	0.8318	0.8405	0.8815	**0.9356**	0.9304
	$\Delta\downarrow$	0.1341	0.1950	0.1811	0.3294	**0.1175**

The larger the value of PSNR, the smaller the distortion of the generated image relative to the ground-truth. The formal definition of PSNR is as:

$$\mathrm{PSNR}(I_{mtl}, I_{gt}) = 10 \times \log_{10}\left(\frac{(L)^2}{\frac{1}{N}\sum_{i=1}^{N}(I_{mtl}(i) - I_{gt}(i))^2}\right), \tag{16}$$

where L denotes the maximum pixel ($L = 1.0$ in our case), and N is the number of all pixels in I_{mtl} and I_{gt}. The unit of PSNR is dB.

SSIM is an image quality evaluation metric based on image structure information statistic, which measures image similarity from brightness, contrast, and structure. SSIM value range 0 to 1, and the larger the value, the smaller the image distortion. The formal definition of SSIM is as:

$$\mathrm{SSIM}(x, y) = \frac{2\mu_x\mu_y + \kappa_1}{\mu_x^2 + \mu_y^2 + \kappa_1} \cdot \frac{\sigma_{xy} + \kappa_2}{\sigma_x^2 + \sigma_y^2 + \kappa_2}, \tag{17}$$

where x, y denote two images to be compared, μ and σ^2 are the mean and variance, σ_{xy} is the covariance between x and y, and κ_1, κ_2 are constant relaxation terms.

Unfortunately, both PSNR and SSIM are limited to measure the fidelity quality, but they do not fully consider the visual characteristics of human (human is highly sensitive to the contrast difference with low spatial frequency, human sensitivity to brightness contrast differences is higher than chromaticity, and human perception of an area is affected by its surrounding areas, etc.). So the evaluation results of PSNR and SSIM cannot be consistent with human subjective perception sometimes. Put it differently, the over-smoothed images are reported to achieve higher PSNR and SSIM scores than texture rich images [18,24].

Network output stability evaluation. The Mean value refers to the sum of all data in a group of data divided by the number of this group of data.

$$\bar{X} = \frac{X_1 + X_2 + X_3 + \cdots + X_n}{n}, \tag{18}$$

where X_i represents the specific value of Metrics for each image, and n represents the number of images.

Variance (Var) is used to calculate the difference between each observation value and the Mean value, expressed as:

$$\sigma^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n}, \tag{19}$$

where σ^2 represents the variance.

Standard deviation (Std) can reflect the dispersion degree of a data set and is the arithmetic square root of variance,

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n}}. \tag{20}$$

The smaller the Std is, the more stable the output of the network.

When considering the minimum (Min) and maximum (Max), we not only hope that they are as large as possible, but also rather more hope that the gap between them, Δ, is not too large.

$$\Delta = Max - Min. \tag{21}$$

4.5. Main results

It is not practical to physically collect a large amount of LR-HR pairs low-level brain MRI training data. We use Bicubic interpolation to degrade BraTS slices by ×4 down-sampling, and add random Gaussian noise to simulate low-level MRI in reality.

The experimental results of RIRGAN and four selected SR method baselines on brain MRI dataset BraTS with noisy LR inputs are shown in Table 2 and Fig. 6. We compared the proposed method RIRGAN with SR methods: Bicubic interpolation, SRResNet (perception-driven method), SRGAN (GAN-based method), and RDN (PSNR-oriented method), where we use the same inputs as our RIRGAN, LR image with scale factor 4 and Gaussian noise.

4.5.1. Comparison in numerical objective metrics with SR methods

In Table 2, we randomly selected 175 brain MRI slice images and calculated the mean PSNR and SSIM values of the output images obtained by each selected baseline method relative to the ground-truth. From Table 2, it can be seen that our proposed RIRGAN achieves the best results in both PSNR and SSIM, i.e., it not only outperforms the classic Bicubic, SRResNet, and SRGAN, but also achieves much better results than the SOTA SISR baseline, RDN.

The reasons for RIRGAN's superior performances of RIRGAN are as follows: (i) the architecture of RIR-G realize the fully extraction of different hierarchical features, through local residual learning, regional residual learning, and global residual learning especially focusing on the extraction of HF features; (ii) we replace the standard discriminator with RaD, which can guide RIR-G to generate more details and textures; (iii) the hybrid loss function in RIRGAN makes it more suitable for low-level vision in medical images.

4.5.2. Comparison in visual perception with SR methods

Fig. 6 shows the visualized results of RIRGAN and the other four baselines on BraTS together with ground-truth. Specifically, we first find that the images processed by Bicubic are both blur and noisy, this is because the interpolation method cannot remove the noise in low-level medical image but simply enlarge the image to the corresponding multiple; however, please also note that not achieving good super-resolution results does not mean Bicubic is bad, actually, Bicubic and other interpolation solutions also have many advantages and have thus been widely used in many other research fields, e.g., domain adaptation.

With the help of perceptual loss, the images obtained by SRResNet and SRGAN add various texture details, but we can see from Table 2 that the brightness and structure of images are changed (reflected by SSIM values), and in Fig. 6 the output image processed by SRResNet and SRGAN contains grid artifacts, especially at the edges of the image. This proves that perception loss based SR method and GAN based SR method can improve the visual similarity, but they also lead to artifacts that affect image quality [48]. Although RDN with complex feature extraction module has already become the most advanced model in the field of SISR in recent years, the output images of RDN that only use pixel loss are blurry and losing some textures. This is because the pixel loss in PSNR-oriented RDN contributes to both SR task and DN task, but it will obviously lead to over-smooth, which is unacceptable in low-level medical image processing. In Table 2, we can also clearly see that the output image quality of the RDN is less stable than the proposed RIRGAN, with more significant differences between images, which makes it less applicable than RIRGAN in medical practices.

The output brain MRI of RIRGAN is the closest to ground-truth visually. The aforementioned content proves that even if the loss function

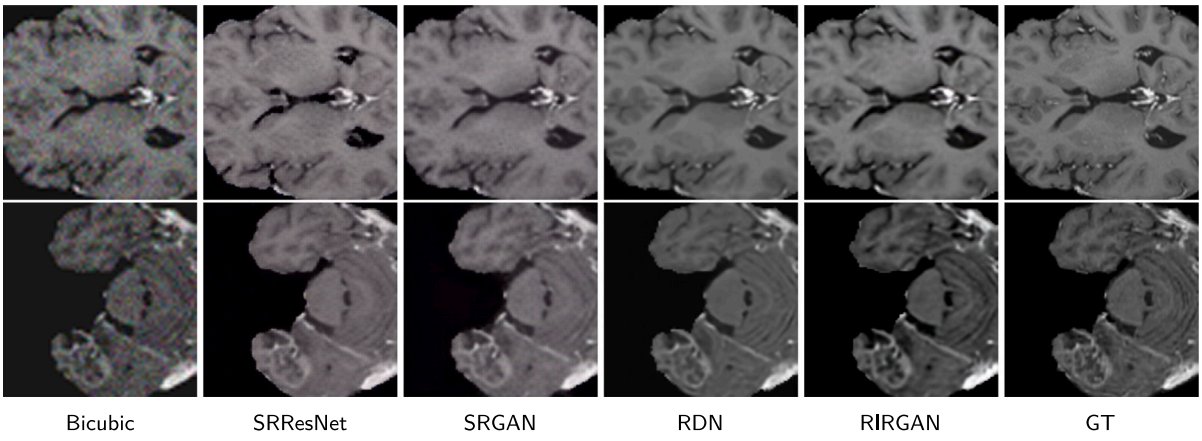
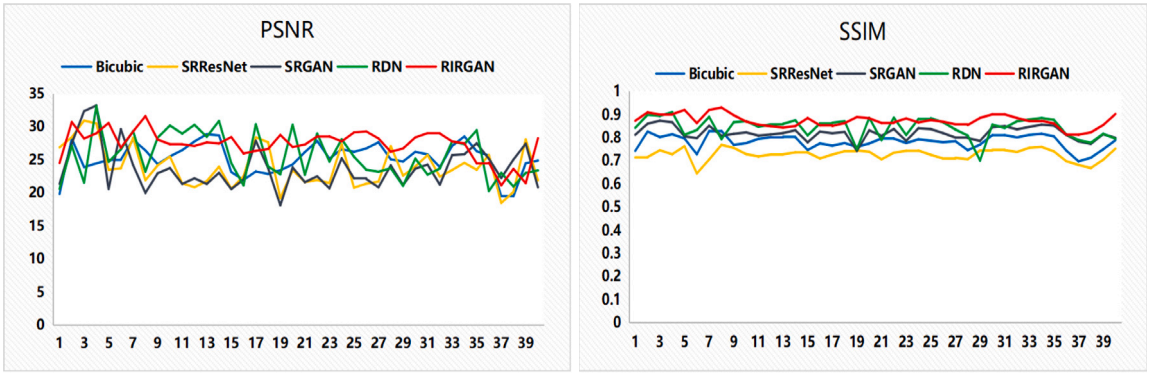


Fig. 6. Visualized multi-task learning results for RIRGAN and the baselines using inputs of ×4 LR image and random Gaussian noise. With the help of pixel loss and perceptual loss, SRResNet, SRGAN and RDN eliminate the noise in input images to some extent while completing the SR task, but with very poor effects; on the other hand, MTL-based RIRGAN has much better visual effect.



(a) PSNR (b) SSIM

Fig. 7. The multi-task learning results of RIRGAN and the SR baselines, which aim to reflect the models' stability by measuring the quality of output images. We randomly selected 40 slices for testing and plotted the PSNR and SSIM values of the output images of each method. The higher the curve in the figure, the higher the Mean value, and thus the better the output quality of the network. The closer the curve in the figure is to the horizontal, the better the stability of the network's output. The results indicate that our RIRGAN performs in both PSNR (a) and SSIM results (b), so the image quality generated by RIRGAN is both high and stable.

with the DN attribute (perceptual loss in SRResNet and SRGAN and pixel loss in RDN) is included, the STL-based SR methods still cannot complete MTL task effectively and commendably.

In fact, we know that the ground-truth image in BraTS contains Speckle noise due to the acquisition instrument and other reasons. Thanks to the existence of TV loss, our RIRGAN removes the random Gaussian noise added to the input image, and at the same time, it also removes the Speckle noise of the image itself to a certain extent. This is because Gaussian noise and Speckle noise are both additive noise. According to the additional experiments in Section 4.8, our RIRGAN is sensitive to a certain kind of noise after training.

Compare with other SR methods, RIRGAN restores the potential information contained in noisy LR inputs. While compare with ground-truth, RIRGAN also plays a role in image enhancement for denoise of Speckle noise.

4.5.3. Comparison of output stability with SR method

Similarly, to study whether the outputs of our RIRGAN are more stable than other SR models when processing low-level brain MRI, we further made a simple comparison to illustrate. In practical applications, we hope that the outputs of the model are not only effective but also stable. In other words, the clinical application requires that the model should ensure that the quality of each output image should be generally maintained at a certain level, rather than some of the quality is particularly good, while some are very poor.

Fig. 7 shows the PSNR and SSIM values of 40 slices randomly selected in noisy LR testing inputs after SR methods and RIRGAN processing. We use the 40 images as a sample results to reflect the overall effect. It can be seen from the PSNR results in Fig. 7(a) and SSIM results in Fig. 7(b) that the red line representing RIRGAN is not only at the top of the figure but also has less fluctuation compared with other lines. To further and more accurately evaluate the stability of the output of each model, we choose six indicators, namely, Mean, Var, Std, Min, Max and Δ, shown in Table 2. They are the most important measures that represent the trend of a group of data sets and are used to describe the trend and dispersion of data sets.

Combining the results in both Fig. 7 and Table 2, we can see clearly that, compared with the classic SR methods, Bicubic, SRResNet, and SRGAN, and the SOTA SR baseline, RDN, in the multi-task learning of low-level brain MRI, RIRGAN not only has good results on the mean value of PSNR and SSIM, but also the quality of output image remains at a relatively stable level, and the difference between image quality is moderate, proving that the performances of RIRGAN is stable and high-quality. In summary, these curves and values reveal the effectiveness and stability of the proposed RIRGAN when handling low-level brain MRI than other STL-based SR methods.

4.6. Ablation study

To indicate the effectiveness and necessity of the proposed three improvement in RIRGAN (RIR-G, RaD and hybrid loss function for

Table 3
Ablation studies on BraTS using scaling factor ×4, random Gaussian noise, and 200K updates, where the best results are bold.

Different combination of RIR-G, RaD and Hybrid Loss								
RIR-G		×	✓	×	×	✓	×	✓
RaD		×	×	✓	×	✓	✓	✓
Hybrid Loss		×	×	×	✓	×	✓	✓
PSNR	Mean↑	26.04	25.84	25.26	27.14	27.35	23.50	**28.17**
	Var↓	10.05	11.78	9.963	17.52	10.60	16.84	**3.452**
	Std↓	3.171	3.432	3.156	4.186	3.256	4.103	**1.858**
	Min–	14.80	16.51	17.09	14.30	18.36	13.53	**21.12**
	Max–	32.07	32.25	31.44	**33.74**	32.47	32.30	31.62
	$\Delta\downarrow$	17.27	15.74	14.35	19.44	14.11	18.77	**10.50**
SSIM	Mean↑	0.8502	0.8131	0.8058	0.8538	0.8541	0.8175	**0.8799**
	Var↓	0.0025	0.0020	0.0026	0.0011	0.0008	0.0030	**0.0006**
	Std↓	0.0503	0.0449	0.0507	0.0327	0.0291	0.0548	**0.0235**
	Min–	0.5367	0.3935	0.6798	0.7513	0.6406	0.5760	**0.8129**
	Max–	**0.9829**	0.8605	0.9072	0.9172	0.8979	0.8895	0.9304
	$\Delta\downarrow$	0.4462	0.4670	0.2274	0.1659	0.2573	0.3135	**0.1175**

Table 4
Different combinations of training input and testing input in different groups and purposes.

Group	Training input	Testing input	Purpose
SR-only	LR	LR	Whether RIRGAN trained by LR can complete the SR task
SR-test1	LR	LR+noise	Whether RIRGAN trained by LR can complete the DN task
SR-test2	LR+noise	LR	Whether RIRGAN trained by MTL can complete the STL task
MTL	LR+noise	LR+noise	Whether RIRGAN trained by MTL can complete the MTL task

MTL) when processing low-level vision of brain MRI, ablation studies are further conducted, where several intermediate models that only use one or two improvements. All the networks have the same feature extraction basic block number ($N = 8$) and input images are (I_{lr+n}). The baseline is the SRGAN use 8 residual-blocks. We should be aware that SRGAN is a proven training balanced GAN model, which means that the generator and discriminator in SRGAN are matched.

We then add one of RIR-G, RaD or hybrid loss to the baseline, resulting in the results from 2nd to 4th combination in Table 3, respectively. The experimental results indicate that the use of RIR-G or RaD alone cannot improve the numerical metrics. We believe that there are several reasons for this: firstly, the improvement of GAN-based model improves the perception quality of images, while PSNR and SSIM values contradict with perception; secondly, after changing the generator or discriminator alone separately will break the balance already formed in SRGAN, that is, a powerful generator requires a mighty discriminator to work with it. The hybrid loss, which is specially designed for low-level vision in medical image, enables the network achieve MTL. Therefore, when processing complex inputs, even using hybrid loss alone can significantly improve the performance of the network. This thus proves that it is reasonable to introduce MTL in low-level medical image vision.

We further add two components to the baseline, resulting in the results from the 5th and 6th combination in Table 3 respectively. It can be seen that use RIR-G + RaD would perform better than only one of them. This is because after using a complex block, a more powerful discriminator is needed to guide the generator to make better use of the extracted features, and RaD has done this well, which proves the RIR-G and RaD are matched and effective in low-level medical image vision. When RIR-G + hybrid loss components added to baseline, under the same training settings, the results show a great decline. We inspected each output image, and found that the network output is very unstable, some image quality is particularly good, while some are very poor. This proves that our generation task is too difficult compared to the SR or DN task of STL. Without the guidance of a powerful discriminator, the generator falls into a local optimum. This result is consistent with the conclusion in 2nd and 5th columns in Table 3.

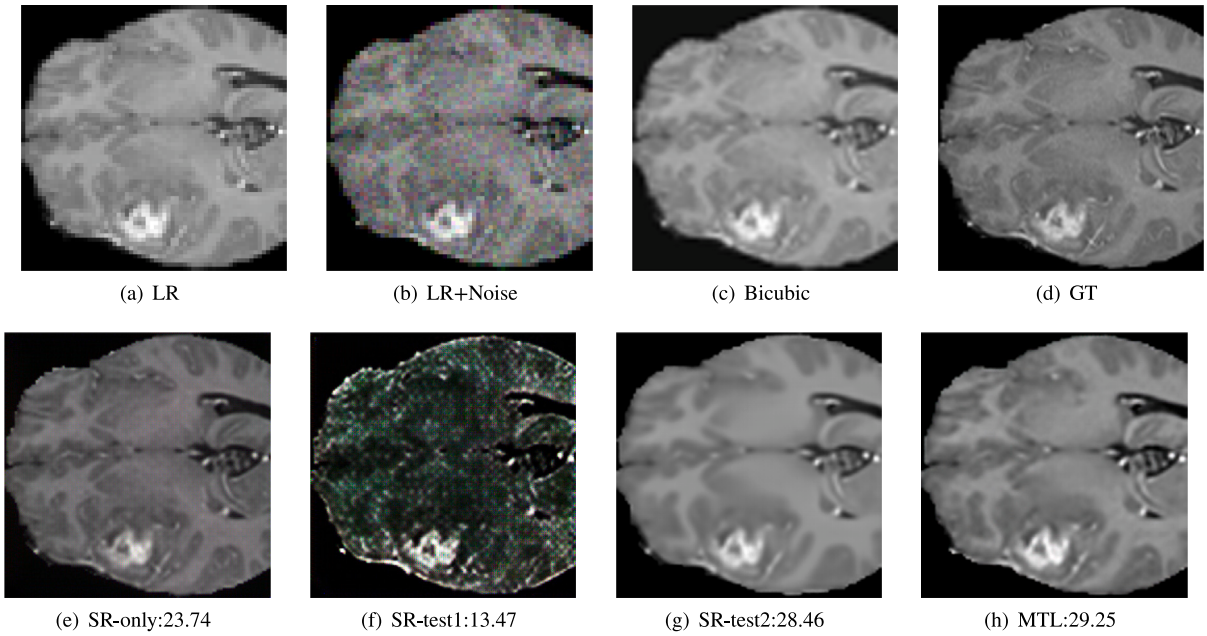
Careful people will find that there is a case missing from the two components in Table 3, RaD + hybrid loss. When RaD + hybrid loss components added to baseline, under the same training settings, discrimination loss of the network will quickly drop to 0. This is because RaD is too powerful compared with the generator at this time, leading to the network mode collapse. This results can also be mutually confirmed with the results in 3rd and 5th columns, that is, the abilities of the generator and discriminator need to match with each other. It also proves that GAN is difficult to train.

When we use these three components simultaneously forms our RIRGAN, and it performs the best in Table 3. The experiment of ablation study proves that the three improvements of RIRGAN are all effective when processing low-level medical images, and the effects are the best when the three improvements cooperate with each other. This is because the additional bypath connection in RIR-G can enhance the deep model's feature learning capability by obtaining different hierarchical features. RaD can guide RIR-G use the multi-hierarchical features to generate higher quality images. And the hybrid loss function enables the network to achieve MTL, which improves the performance of the network by tackling SR and DN problems, and gets more effective and robust images. Therefore, the above observations demonstrate that the proposed three improvements are all effective and essential to achieve the superior quality medical image.

4.7. Effectiveness of RIRGAN for different tasks

Further experiments are conducted to verify the effectiveness of MTL. Let us take a look at the performance of RIRGAN in different combinations of training and testing sets through this part, shows in Table 4 and Fig. 8. We divided the input into two types: one is the degradation images with ×4 down-sampling by Bicubic called **LR** (shows in Fig. 8(a)), the other type is to add random Gaussian noise to the LR, which we called it **LR+noise** (shows in Fig. 8(b)). According to the different combination of training input and testing input, we divide the experiments into four groups, and set Table 4 for details.

The first group, we call it **SR-only**, uses LR as training input, we find that the network converges very fast, which only needs about 100k iterations. The testing input in SR-only is also LR, and the SR-only visual result is shown in Fig. 8(e). This kind of input was same with the ordinary STL SR method. Put it another way, the experiment of the SR-only group is used to test whether RIRGAN could complete STL. The output of the SR-only group has a good visual effect compared to the baseline Bicubic with LR input, but a low PSNR value and the brightness is also significantly different from GT, which proves that



(a) LR (b) LR+Noise (c) Bicubic (d) GT

(e) SR-only:23.74 (f) SR-test1:13.47 (g) SR-test2:28.46 (h) MTL:29.25

Fig. 8. Visualized results of the effectiveness of multi-task learning in PSNR. Bicubic is selected as the baseline.

Table 5
The robustness of RIRGAN to additive noise and salt & pepper noise with different iterations in multi-task learning. RIRGAN converged by training with additive noise for 200k iterations, but salt & pepper noise needs more training iterations, which indicates that salt & pepper noise is more difficult to handle than additive noise.

Kinds	Additive			Salt & Pepper			Salt & Pepper		
Noise	Gaussian	Speckle	Localvar	S&P	Salt	Pepper	S&P	Salt	Pepper
Iteration	200k	200k	200k	200k	200k	200k	350k	350k	350k
PSNR	28.23	28.67	28.19	24.69	24.70	24.67	27.07	27.16	27.04
SSIM	0.8809	0.8965	0.8808	0.8837	0.8841	0.8838	0.8962	0.8957	0.8956

our RIRGAN can complete the STL task as a perception-driven, but the effect is not PSNR-superior.

The second group we named it as **SR-test1**, which uses LR as training input and use LR+noise as testing input, and the output result shows in Fig. 8(f), which is not a meaningful output. It proves that a model without noise training could not complete the task of DN.

The third group is called **SR-test2**, which we use the training input of LR+noise and testing input of LR. The purpose of this experiment is to verify whether RIRGAN trained by MTL can complete the STL task. The results in Fig. 8(g) achieves a high PSNR value, but a blurred image. This result is similar to the original PSNR-oriented SR method, which obviously does not meet the requirements of medical practical applications.

The fourth group is **MTL** group, using LR+noise as both training input and testing input, which is the training method we proposed in Section 4.5. The image in the MTL group shows in Fig. 8(h) which balanced the visual effect with numerical values of metrics.

From the four groups of comparative results, we can see that RIRGAN can effectively complete the SR task (both PSNR-oriented and perception-driven) through different combinations of training input and testing input. Comparing the results of SR-test2 and MTL group, we find that RIRGAN after trained by LR+noise input, when processing LR testing input, model tends to produce overly smooth output (shows in Fig. 8(g)). Although the PSNR value of this blurred image may be high, compared with ground-truth (GT, shows in Fig. 8(d)), it obviously loses various details and causes distortion. When the testing input is LR+noise, the output image (shows in Fig. 8(h)) contains realistic textures and details, which prove the effectiveness of MTL. Furthermore, the SR-test2 and MTL group experiments prove that the pre-trained RIRGAN model obtained through noisy LR inputs can handle multiple tasks: SR (SR-test2 group) and SR+DN (MTL group). These two groups of experiments also demonstrate that the MTL based model has high generalization ability.

4.8. Robustness to different kinds of noise

Similarly, to investigate the influence of different kinds of common noise, additive noise and salt&pepper noise, on the performance of RIRGAN, experiments are further conducted, and the results are shown in Table 5.

Generally, when processing additive noise, we use LR image (downsampling to factor 4) with random Gaussian noise to train RIRGAN, then use the trained model to process different types of additive noise: Gaussian, Speckle, Localvar. In Table 5, we can see that RIRGAN works well on different kinds of additive noise. When handling salt&pepper noise, the model which pre-trained by Gaussian noise cannot be directly used to remove salt&pepper noise, so it is necessary to re-train the network with salt&pepper noise. When trained 200k iterations, the model has not converged and the image denoising effect is not ideal. Therefore, we continue to train the model to 350k iterations and achieved relatively satisfactory results.

After analyzing the reason, we find that additive noise did not change the pixel value of the original image, only superimposed different noise on it. But salt&pepper noise is a kind of impulse noise. Salt noise will randomly change the original pixel value to a white dot, while pepper noise will randomly change the original pixel to a black dot. Because salt&pepper noise changes the pixel value of the original image, it is more difficult for the network to process salt&pepper noise than additive noise. Meanwhile, the images in BraTS itself contain a small amount of additive noise, and then add salt&pepper noise,

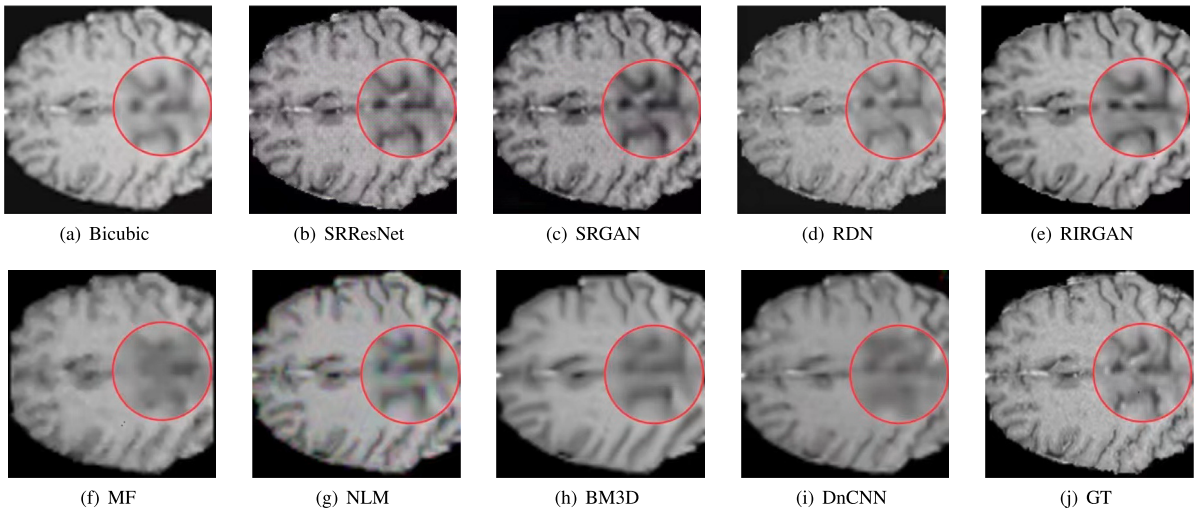


Fig. 9. Visualized results of multi-task learning RIRGAN and single task learning SR and DN methods. The experimental inputs of SR methods are LR images (down-sampling to factor $s = 4$), and the inputs of other methods are all LR images with random Gaussian noise. Since the DN model has no amplification function, we will magnify the results after processing the DN model to the same size as GT using Bicubic interpolation. The zoomed-in regions marked by the red circle are embedded in the original image to show the details of the output.

Table 6
Quantitative results of multi-task learning RIRGAN and single task learning SR and DN methods in terms of PSNR, SSIM, and qualitative comparative metric, MOS. Best results are marked bold.

Method	Bicubic	SRResNet	SRGAN	RDN	RIRGAN	MF	NLM	BM3D	DnCNN
Input	LR	LR	LR	LR	LR+noise	LR+noise	LR+noise	LR+noise	LR+noise
Task	SR	SR	SR	SR	MTL	DN	DN	DN	DN
PSNR	25.48	25.40	25.65	**29.92**	28.05	26.54	24.57	26.30	27.19
SSIM	0.8950	0.8762	0.9140	**0.9293**	0.8792	0.8240	0.8323	0.8373	0.8856
MOS	1.207	3.187	3.810	2.837	**4.183/4.921**	1.693	3.418	2.746	2.154

making the model processing more complex. When the input images of RIRGAN are brain MRI with salt&pepper noise, the pixel information of the input images itself has changed due to the influence of salt&pepper noise. After 350k iterations of training on salt&pepper noise input, RIRGAN can already recover the structural information of the images well (reflected by SSIM values), but the restoration of pixel information in the image is not as good as that of additive noise (reflected by PSNR values). Therefore, the experimental results show that the model needs more time to process salt&pepper noise than additive noise.

Whether for additive noise or salt&pepper noise, RIRGAN can remove this type of noise only by training once, which proves that RIRGAN has strong robustness to different kinds of noise. The experiment also proved that the MTL based model has high generalization ability.

To verify whether our RIRGAN can simultaneously handle two different types of noise, we conduct a new experiment. The new experimental input adds random Gaussian noise and salt&pepper noise with down-sampling to factor 4, which makes the input of the network more complex. According to our experiment, this overly complex input has caused confusion in our network. The model needs to complete three complex tasks simultaneously: SR, DN, and noise classification. Unfortunately, our model does not yet have the ability to complete this task.

4.9. Comparative results with SR methods

On Section 4.5, we use noisy LR brain MRI as input to compare RIRGAN with SR models, and achieved superior results. But RIRGAN is an MTL model, while these SR methods are STL models. The complex input seems unfair to STL SR methods. To further evaluate the performance of RIRGAN, we conducted new experiments. To be fair, we change the inputs of four SR methods with an LR image only, while the inputs of RIRGAN are still noisy LR image.

In Fig. 9 and Table 6, perception-driven SRResNet and GAN-based SRGAN can achieve good visual effects, but they will lead to the decline of PSNR and SSIM, and produce undesirable artifacts (grids, patches, etc.). The PSNR-oriented RDN use pixel loss only will lead to over blurring, but best results in PSNR and SSIM score. While our RIRGAN gets the second best result in PSNR. However, it is worth noting that RDN still has unstable image output quality in SR experiments. Some images have poor recovery on black backgrounds and where contrast is sharp, which severely imposes the doctor's sense of experience.

We know that a high PSNR and SSIM value will cause the output image to be too smooth, which is completely inconsistent with visual perception [24]. In real medical applications, physicians often focus on the visual effect of medical images. To alleviate this issue, we choose mean opinion score (MOS) as the evaluation index of human subjective perception, which can more realistically reflect the similarity between model results and ground-truth from several aspects, such as brightness, artifacts, details, etc. MOS is a commonly used subjective image quality assessment (IQA) method. MOS assesses the quality of samples by randomly selecting some pairs of ground-truth samples and generated samples, and scoring them manually by several people. Although the MOS seems a faithful method for evaluating the quality of SR images, it has some inherent defects. Owning to the difference in each person's experience and the difference of attention to different features of image may affect the evaluation results, the score of MOS is often different from person to person. However, MOS needs to consume numerous human cost, economic cost and time cost, which is only applicable to small sample sets in practical. Nevertheless, MOS is considered being the most reliable IQA method for accurately measuring perceptual quality [53].

The MOS value in our experiment is based on 20 randomly selected images from the testing set. Each image was processed by Bicubic, SRResNet, SRGAN, RDN and RIRGAN, and then 15 people were invited to give a full blind scores 1 (bad) to 5 (good) for 20 groups of images

in a disordered order. These 15 evaluators are either radiologists with more than three years' clinical experience from the Department of Radiology, Hainan Women and Children's Medical Center, China, or master/Ph.D. students with more than two years research experiences in the areas of medical image processing from the School of Health Sciences and Biomedical Engineering, Hebei University of Technology, China. Therefore, their clinical or academic training and experience in the field of medical image analysis make them have sufficient ability to distinguish the quality of MRI images. We calculated the final MOS value as the arithmetic average over all 300 ratings.

From the PSNR, SSIM, MOS results shown in Table 6 and the visualized results shown in Fig. 9, we can see that SRResNet and SRGAN can generate pleasant texture details under the help of perceptual loss, but they also bring in lots of artifacts and noise, which makes the output images appear to be unclean. Although RDN has the highest PSNR and SSIM values, its visualized results is lacking of necessary high frequency texture details, making it seems to be very blurring. And our proposed RIRGAN has achieved the best balance between the quantitative performances in PSNR and SSIM and the qualitative visualization results to satisfy the human vision. Therefore, our model is more in line with the requirements of doctors in practical applications.

4.10. Comparative results with DN methods

Although our RIRGAN is not a specially designed DN model and the DN is only the related auxiliary task for SR, we still carried out experiments comparing with DN methods: Median Filtering (MF, nonlinear-smoothing-based method) [54], Non-Local Means (NLM, non-local-based method) [55], Block Matching 3D (BM3D, image-block-matching-based method) [56] and DnCNN (deep-learning-based method) [57]. All these methods are proved to be effective DN methods.

In the comparison experiments with DN group, we did not choose to denoise the original size image directly, because the large size image (128×128) contains much more information than the small size image (32×32), which lacks fairness for our RIRGAN. To ensure that the size of DN network input and output images is consistent with our RIRGAN, we use LR+noise brain MRI slices as inputs. That is to say, the inputs of the DN group are the same with RIRGAN. And we enlarge the output images obtained by DN methods using Bicubic to ensure the size of output images is consistent with the ground-truth. The reason why we use the Bicubic interpolation instead of the more powerful SR methods, e.g., SRGAN and RDN, for enlargement is as follows: SRGAN and RDN are deep-learning-based generative SR methods, so they not only have the ability to enlarge images, but also have the capability to improve and repair the qualities of images, i.e., they also have certain denoising ability with the help of their losses (e.g., perceptual loss for SRGAN and pixel loss for RDN); however, the purpose of this experimental study is not to obtain the best output images but to compare the different denoising capability of RIRGAN and the existing DN methods; therefore, using deep-learning-based generative methods, e.g., SRGAN and RDN, for enlargement, will further narrow the difference in image quality generated by different DN methods (i.e., the generated images will no longer reflect the results of using these DN methods for denoising, but the results of using them and also RDN or SRGAN for two-stage denoising), and thus bring bias in comparing their denoising capability; consequently, here we choose Bicubic, which is an interpolation-based method without denoising capability, to preserve the quality differences between output images as much as possible when enlarging them and avoiding comparative bias. Finally, please note that this setting is only to obtain more intuitive and clearer comparison in denoising experiments in this subsection; actually, using SRGAN and RDN for two-stage denoising will definitely obtain better-denoised images than using Bicubic, so they are better choices in clinical practices if having sufficient computing facilities.

The PSNR, SSIM and MOS values are obtained in the same way as those described in the SR group in Section 4.9. The results are also shown in Table 6 and Fig. 9. From the results of Table 6 and Fig. 9, we can see that the DN methods with small size input are difficult because very little pixel information can be provided for model to learn. Furthermore, what we cannot deny is that the impact of image blurring caused by Bicubic amplification on DN group results. But we still believe that our RIRGAN has achieved good results in the restoration of low-level vision brain MRI.

Table 7
The performances of RIRGAN using different number of blocks and parameters in SR and DN multi-task learning.

Method	Block	Params	PSNR	SSIM
RIRGAN	$N = 8$	3.3M	28.26	0.8833
RIRGAN*	$N = 10$	4.1M	22.89	0.8199

Table 8
The impact of the number of parameters on the performances of RIRGAN in SR single-task learning.

Block	$N = 2$	$N = 4$	$N = 8$	$N = 16$	$N = 32$
Params	1.1M	1.8M	3.3M	6.3M	12.2M
PSNR	17.10	18.42	24.44	24.55	23.71
SSIM	0.6002	0.6203	0.8667	0.8992	0.7898

4.11. The impact of parameter reduction on the performances of RIRGAN

Additional experiments are conducted to investigate the impact of parameter reduction on the performances of RIRGAN. We first increase the number of blocks of RIRGAN from N = 8 to N = 10 in multi-task learning, which thus increases the number of parameters from 3.3M to 4.1M. As shown in Table 7, the performances of RIRGAN* with N = 10 are surprisingly worse than those of RIRGAN with N = 8. This may be because more trainable parameters lead to more challenges in optimization and make the model feasible in over-fitting [31]; since multi-task learning is more complex than single task learning, increasing the number of blocks and parameters of RIRGAN make it more difficult to achieve a balance between generator and discriminator.

Then we also investigate the impact in single-task learning of super resolution (SR), where the inputs of training are LR medical images without additional noise and the noise-based total variation (TV) is thus removed from RIRGAN (setting its weight to 0). As shown in Table 8, with the increase of RIR-Blocks, the number of parameters of RIRGAN also increases consistently; however, the super-resolution performances of RIRGAN first increases from N = 2 to N = 16, but decreases from N = 16 to N = 32. This thus proves our argument again that increasing the number of blocks and parameters of RIRGAN makes it more difficult to achieve balance between generator and discriminator, so it does not always results in the rise of performance. In addition, by comparing the performances results of RIRGAN in N = 8 and N = 16, we notice that by doubling the number of parameters, the performance increase is not significant, especially for PSNR. Consequently, according to the results in Tables 7 and 8, we believe that the increase of parameter will not result in significant performance improvements for RIRGAN, so setting the number of blocks to N = 8 not only light-weights the model but also achieves satisfactory performances in both multi-task and single-task learning scenarios.

5. Discussion

The proposed end-to-end lightweight MTL model, RIRGAN, which can concurrently accomplish low-level medical image vision in both SR and DN tasks. This means that when facing clinical noisy LR images, there is no longer to improve resolution and definition, respectively. Through our experiments, RIRGAN has been proven to be able to complete the restoration of low-level vision brain MRI.

5.1. Social impact of RIRGAN

RIRGAN can be widely used in a lot of clinical scenarios. In some image centers, there are some older medical data that have poor image resolution and high noise due to various reasons in the process of collection, storage, transmission, etc. The application of our RIRGAN can help restore these medical image data. In underdeveloped remote areas, due to economic limitations and hardware costs of image acquisition equipment, MRI images may not be able to display detailed information of patient lesions. By applying our proposed RIRGAN to such clinical practice, doctors can obtain images with more detailed information.

Clear and high-quality images can effectively reduce the workload of doctors and improve the efficiency and accuracy of medical image segmentation. We take an MRI assisted cancer diagnosis as an example, where doctors need to accurately delineate the outline of the tumor area on 3D MRI images of patients composed of hundreds of slices as the target area. Under such a large workload, if the image quality is very low, the whole image segmentation process is very time-consuming and laborious, which can easy to cause misdiagnosis. By applying our proposed RIRGAN in such clinical practices, the medical image quality will be improved and the segmentation task will be easier.

5.2. Limitations and future works

Similar to the existing SISR methods, RIRGAN also experiences information loss in areas where image edges and pixels change dramatically. So an interesting future research is to find a new loss function that can pay special attention to the areas with sharp pixel changes, so as to improve the network's ability in generating such areas. As a supervised regression task, RIRGAN requires a large amount of data for training, but the number of high-quality images in medical practice is limited. We can seek some semi-supervised [58] and self-supervised [59] methods to improve our model.

Despite achieving good performances in either additive noise or Salt & Pepper noise, our experiments also find that the performances of the proposed RIRGAN significantly degrades when both types of noises are added. We believe this problem may be due to the limitation of the noise-related loss, i.e., the total variation (TV) loss. Specifically, as stated in Section 3.4. and the existing TV loss related works [17,50], TV loss is a regularization loss item aiming to enhance the image's spatial smoothness and reduce noise. As a regularization loss, the weight of TV loss has to be set to a relatively small value, otherwise the auxiliary task (i.e., the denoising task in our work) will overwhelm the main task (i.e., the super-resolution task) and negatively affect the model's performances in the main task; however, setting its weight to a small value inevitably limits the model's denoising capability, making it unable to handle complex noises well. Consequently, we will continue to investigate how to further improve RIRGAN to achieve satisfactory performances under complex and multi-type noise circumstances in future research works. A Potential research direction may be introducing other kinds of denoising loss and discovering appropriate loss weight combinations to help model achieve good performances in both SR and DN tasks.

6. Conclusion

In this paper, we proposed an end-to-end lightweight MTL model, RIRGAN, which can do SR and DN multi-task learning of low-level medical image vision simultaneously. Compared to the SISR methods based on STL, our RIRGAN has achieved good results in both quantitative evaluation of numerical metrics and qualitative evaluation of human vision. Specifically, the technical improvements of RIRGAN are threefold. First, RIR-G with SSC, LSC and GSC improves the hierarchical feature extraction ability, allowing our RIRGAN to reach a deeper network than SRGAN easily with a smaller number of parameters and focus on learning high-frequency information of medical images. Second, we use RaD to replace the original discriminator loss, which can judge whether an image is more realistic than another one, greatly improving the stability and accuracy of our model. Third, the use of hybrid loss function enables RIRGAN to achieve MTL of SR and DN, and to focus on different performance of images in a more balanced way, so as to achieve a more robust MTL effect. All these improvements make our RIRGAN different from the current popular STL models for SR and DN tasks.

Our experimental studies first show that RIRGAN not only achieves better performances than the classic SR methods, Bicubic, SRResNet, and SRGAN, but also outperforms the state-of-the-art SISR baseline, RDN, in multi-task learning. Then, we randomly select 40 images and show that the quality of high-level medical images generated by RIRGAN is more stable than the baselines. Ablation studies are also conducted to show that all the above improvements of RIRGAN are effective and essential for the RIRGAN to achieve the superior performances. Finally, Some additional experiments are also conducted to show that the multi-task learning results of RIRGAN are comparable with the single task learning results of the baselines and the parameter reduction in RIRGAN will not bring significant performance loss. All of these experimental results sufficiently prove the superiority and stability of the proposed RIRGAN in the tasks of Brin MRI super-resolution and denoising.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under the grants 62276089 and 61906063, by the Natural Science Foundation of Hebei Province, China, under the grant F2021202064, by the S&T Program of Hebei, China, under the grants 215676146H and 225676163GH, by the Key Research and Development Project of Hainan Province, China, under the grant ZDYF2022SHFZ015, and by the Natural Science Foundation of Hainan Province, China, under the grant 821RC1131. This work was also supported by the AXA Research Fund, France.

References

[1] C.H. Pham, C. Tor Díez, H. Meunier, N. Bednarek, R. Fablet, N. Passat, F. Rousseau, Multiscale brain MRI super-resolution using deep 3D convolutional networks, Comput. Med. Imaging Graph. 77 (2019) 101647.
[2] N. Ichijo, S. Matsuno, T. Sakai, Y. Tochigi, M. Kaminoyama, K. Nishi, R. Misumi, S. Nishiyama, Resolution enhancement of electrical resistance tomography by iterative back projection method, J. Vis. 19 (2) (2016) 183–192.
[3] J. Liu, W. Yang, X. Zhang, Z. Guo, Retrieval compensated group structured sparsity for image super-resolution, IEEE Trans. Multimed. 19 (2) (2016) 302–316.
[4] B. Lim, S. Son, H. Kim, S. Nah, K.M. Lee, Enhanced deep residual networks for single image super-resolution, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1132–1140.
[5] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, Y. Fu, Residual dense network for image restoration, IEEE Trans. Pattern Anal. Mach. Intell. 43 (7) (2021) 2480–2495.
[6] W. Li, K. Zhao, Q. Lu, L. Lu, N. Jiang, J. Lu, J. Jia, Best-buddy GANs for highly detailed image super-resolution, 2021, arXiv preprint arXiv:2103.15295v3.
[7] X. Luo, R. Chen, Bi-GANs-ST for perceptual image super-resolution, in: Proceedings of the European Conference on Computer Vision, 2019, pp. 20–34.
[8] D. Mahapatra, B. Bozorgtabar, Progressive generative adversarial networks for medical image super resolution, 2019, arXiv preprint arXiv:1902.02114v2.
[9] J. Wang, Y.H. Chen, Y. Wu, J. Shi, J. Gee, Enhanced generative adversarial network for 3D brain MRI super-resolution, 2019, arXiv preprint arXiv:1907.04835v2.
[10] Y. Xia, N. Ravikumar, J. Greenwood, F. A., Super-resolution of cardiac MR cine imaging using conditional GANs and unsupervised transfer learning, Med. Image Anal. 71 (2021) 17.

[11] Y. Chen, R. Xia, K. Zou, MFFN: image super-resolution via multi-level features fusion network, Vis. Comput. (2023).
[12] Z. Wang, G. Gao, J. Li, Y. Yu, H. Lu, Lightweight image super-resolution with multi-scale feature interaction network, in: Proceedings of the IEEE International Conference on Multimedia and Expo, 2021, pp. 1–6.
[13] S. Vandenhende, S. Georgoulis, W.V. Gansbeke, M. Proesmans, D. Dai, L. Gool, Multi-task learning for dense prediction tasks: A survey, IEEE Trans. Pattern Anal. Mach. Intell. 44 (7) (2020) 3614–3633.
[14] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, Y. Fu, Image super-resolution using very deep residual channel attention networks, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 294–310.
[15] J.M. Alexia, The relativistic discriminator: a key element missing from standard gan, 2018, arXiv preprint arXiv:1807.00734.
[16] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, C.C. Loy, Y. Qiao, X. Tang, ESRGAN: Enhanced super-resolution generative adversarial networks, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 63–79.
[17] H. Pan, Y.W. Wen, H.M. Zhu, A regularization parameter selection model for total variation based image noise removal, Appl. Math. Model. 68 (2019) 353–367.
[18] J. Zhu, G. Yang, P. Lio, How can we make GAN perform better in single medical image super-resolution? A lesion focused multi-scale approach, in: Proceedings of the IEEE International Symposium on Biomedical Imaging, 2019, pp. 8–11.
[19] B.H. Menze, A. Jakab, S. Bauer, K.C. Jayashree, K. Farahani, J. Kirby, Y. Burren, et al., The multimodal brain tumor image segmentation benchmark (BRATS), IEEE Trans. Med. Imaging 34 (10) (2015) 1993–2024.
[20] Y.D. Wang, R.T. Armstrong, P. Mostaghimi, Boosting resolution and recovering texture of micro-CT images with deep learning, 2019, arXiv preprint arXiv:1907.07131v3.
[21] C. Tor-Díez, C.-H. Pham, H. Meunier, S. Faisan, I. Bloch, N. Bednarek, N. Passat, F. Rousseau, Evaluation of cortical segmentation pipelines on clinical neonatal MRI data, in: Proceedings of the International Engineering in Medicine and Biology Conference, 2019, pp. 6553–6556.
[22] C. Dong, C.-L. Chen, K. He, X. Tang, Image super-resolution using deep convolutional networks, IEEE Trans. Pattern Anal. Mach. Intell. 38 (2) (2016) 295–307.
[23] K. Umehara, J. Ota, T. Ishida, Application of super-resolution convolutional neural network for enhancing image resolution in chest CT, J. Dig. Imag. 31 (2017) 441–450.
[24] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, W. Shi, Photo-realistic single image super-resolution using a generative adversarial network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 105–114.
[25] I. Sánchez, V. Vilaplana, Brain MRI super-resolution using 3D generative adversarial networks, 2018, arXiv preprint arXiv:1812.11440.
[26] C. Tan, J. Zhu, P. Lio, Arbitrary scale super-resolution for brain MRI images, 2020, arXiv preprint arXiv:2004.02086v1.
[27] M. Zhao, Y. Wei, K.L. Wong, A generative adversarial network technique for high-quality super-resolution reconstruction of cardiac magnetic resonance images, Magn. Reson. Imaging 85 (2022) 153–160.
[28] T. Tong, G. Li, X. Liu, Q. Gao, Image super-resolution using dense skip connections, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 4809–4817.
[29] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[30] Y. Chen, A.G. Christodoulou, Z. Zhou, F. Shi, Y. Xie, D. Li, MRI super-resolution with GAN and 3D multi-level DenseNet: Smaller, faster, and better, 2020, arXiv preprint arXiv:2003.01217v1.
[31] J. Zhu, C. Tan, J. Yang, G. Yang, P. Lio', MIASSR: An approach for medical image arbitrary scale super-resolution, 2021, arXiv preprint arXiv:2105.10738.
[32] Y. Chen, R. Xia, K. Zou, K. Yang, FFTI: Image inpainting algorithm via features fusion and two-steps inpainting, J. Vis. Commun. Image Represent. 91 (2023) 103776.
[33] Y. Chen, R. Xia, K. Yang, K. Zou, DARGS: Image inpainting algorithm via deep attention residuals group and semantics, J. King Saud. Univ. Comput. Inf. Sci. 35 (6) (2023) 101567.
[34] K. Liang, X. Liu, S. Chen, J. Xie, H.K. Lee, Resolution enhancement and realistic speckle recovery with generative adversarial modeling of micro-optical coherence tomography, Biomed. Opt. Express 11 (12) (2020) 7236.
[35] Y. Chen, R. Xia, K. Zou, K. Yang, RNON: image inpainting via repair network and optimization network, Int. J. Mach. Learn. Cybern. (2023).
[36] J. Kim, J.K. Lee, K.M. Lee, Accurate image super-resolution using very deep convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1646–1654.
[37] S. Liu, E. Johns, A.J. Davison, End-to-end multi-task learning with attention, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 1871–1880.
[38] Z. Yan, Y. Lu, J. Wen, C. Li, Super resolution SPECT reconstruction with non-uniform attenuation, Comput. Biol. Med. 42 (2012) 651–656.
[39] Z. Zhang, S. Yu, W. Qin, X. Liang, Y. Xie, G. Cao, Self-supervised CT super-resolution with hybrid model, Comput. Biol. Med. 138 (2021) 104775.
[40] Q. Zhang, J. Sun, G.S.P.M. Mok, Low dose SPECT image denoising using a generative adversarial network, 2019, arXiv preprint arXiv:1907.11944.
[41] H. Chen, Y. Wang, T. Guo, C. Xu, Y. Deng, Z. Liu, S. Ma, C. Xu, C. Xu, W. Gao, Pre-trained image processing transformer, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2021, pp. 12294–12305.
[42] C. You, G. Li, Y. Zhang, X. Zhang, H. Shan, M. Li, S. Ju, Z. Zhao, Z. Zhang, W. Cong, M.W. Vannier, P.K. Saha, E.A. Hoffman, G. Wang, CT super-resolution GAN constrained by the identical, residual, and cycle learning ensemble (GAN-CIRCLE), IEEE Trans. Med. Imaging 39 (1) (2020) 188–203.
[43] Y. Chen, R. Xia, K. Yang, K. Zou, DGCA: high resolution image inpainting via DR-GAN and contextual attention, Multimedia Tools Appl. (2023).
[44] D. Yuan, Y. Liu, Z. Xu, Y. Zhan, J. Chen, T. Lukasiewicz, Painless and accurate medical image analysis using deep reinforcement learning with task-oriented homogenized automatic pre-processing, Comput. Biol. Med. 153 (2023) 106487.
[45] Z. Xu, X. Zhang, H. Zhang, Y. Liu, Y. Zhan, T. Lukasiewicz, EFPN: Effective medical image detection using feature pyramid fusion enhancement, Comput. Biol. Med. 163 (2023) 107149.
[46] Z. Xu, S. Liu, D. Yuan, L. Wang, J. Chen, T. Lukasiewicz, Z. Fu, R. Zhang, ω-Net: Dual supervised medical image segmentation with multi-dimensional self-attention and diversely-connected multi-scale convolution, Neurocomputing 500 (2022) 177–190.
[47] D. Yuan, Z. Xu, B. Tian, H. Wang, Y. Zhan, T. Lukasiewicz, μ-Net: Medical image segmentation using efficient and effective deep supervision, Comput. Biol. Med. 160 (2023) 106963.
[48] J. Liang, H. Zeng, L. Zhang, Details or artifacts: A locally discriminative learning approach to realistic image super-resolution, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2022, pp. 5647–5656.
[49] Y. Chen, Y. Bai, W. Zhang, T. Mei, Destruction and Construction Learning for Fine-Grained Image Recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 5152–5161.
[50] J.A. Iglesias, G. Mercier, Influence of dimension on the convergence of level-sets in total variation regularization, ESAIM Control Optim. Calc. Var. (2018).
[51] Y. Ma, B. Wei, P. Feng, P. He, X. Guo, G. Wang, Low-dose CT image denoising using a generative adversarial network with a hybrid loss function for noise learning, IEEE Access 8 (2020) 67519–67529.
[52] J.M. Wolterink, T. Leiner, M.A. Viergever, I. Išgum, Generative adversarial networks for noise reduction in low-dose CT, IEEE Trans. Med. Imaging 36 (12) (2017) 2536–2545.
[53] Z. Wang, J. Chen, S.C.H. Hoi, Deep learning for image super-resolution: A survey, IEEE Trans. Pattern Anal. Mach. Intell. 43 (10) (2021) 3365–3387.
[54] R. Huang, K. Qing, D. Yang, K.S. Hong, Real-time motion artifact removal using a dual-stage median filter, Biomed. Signal Process. Control 72 (2022) 103301.
[55] C. Liu, L. Guo, Y. Liu, Y. Zhang, Z. Zhou, Seismic random noise attenuation based on adaptive nonlocal median filter, J. Geophys. Eng. (2) (2022) 157–166.
[56] E.M. Eksioglu, Decoupled algorithm for MRI reconstruction using nonlocal block matching model: BM3D-MRI, J. Math. Imaging Vis. 56 (3) (2016) 430–440.
[57] T.S. Sharan, R. Bhattacharjee, S. Sharma, N. Sharma, Evaluation of deep learning methods (DnCNN and U-net) for denoising of heart auscultation signals, in: Proceedings of the International Conference on Communication System, Computing and IT Applications, 2020, pp. 151–155.
[58] S. Zhang, J. Zhang, B. Tian, T. Lukasiewicz, Z. Xu, Multi-modal contrastive mutual learning and pseudo-label re-learning for semi-supervised medical image segmentation, Med. Image Anal. 83 (2023) 102656.
[59] J. Zhang, S. Zhang, X. Shen, T. Lukasiewicz, Z. Xu, Multi-ConDoS: Multimodal contrastive domain sharing generative adversarial networks for self-supervised medical image segmentation, IEEE Trans. Med. Imaging (2023) 1–20, Early Access.

Miao Yu is currently a PHD student in the State Key Laboratory of Reliability and Intelligence of Electrical Equipment, School of Health Sciences and Biomedical Engineering, Hebei University of Technology, Tianjin, China. She received Master degree in Hebei University of Technology, China, in 2018. Her research interests lie in medical image processing using deep learning methods.

Miaomiao Guo received the M.E. and PhD. degrees from the School of Electrical Engineering, Hebei University of Technology, in 2012 and 2016, respectively. From 2014 to 2015, she visited Johns Hopkins University as a visiting scholar. She is currently an associate professor at the School of Health Sciences and Biomedical Engineering,Hebei University of Technology, China. Her current research focuses on neuroregulatory techniques, brain computer interface and its application in rehabilitation, signal processing of EEG, brain networks, etc.

Shuai Zhang received a M.E. in Biomedical Engineering from Hebei University of Technology, China, in 2005, and a PhD. in Electrical Engineering from Hebei University of Technology, China, in 2009. From 2013 to 2014, he worked as a visiting research associate at the Department of Biomedical Engineering, University of Minnesota, USA. He is now a full professor at the School of Health Sciences and Biomedical Engineering, Hebei University of Technology, China. He has published more than fifty papers in conferences and journals. His current research focuses on electromagnetics, neural modulation, intelligent medical image analysis, and deep learning.

Yuefu Zhan is currently an Associate Chief Physician at Department of Radiology, Hainan Women and Children's Medical Center. He received a Doctor of Medicine degree from the West China Medical School, Sichuan University, China, in 2022, and a Master of Medicine degree from the Xiangya School of Medicine, Central South University, China, in 2010. His current research interests mainly focus on imaging diagnosis, minimally invasive intervention, and AI-based medical image analysis.

Mingkang Zhao received a D.Phil in Biomedical Engineering from Kyunghee University, Republic of Korea, in 2014. He is now a lecturer in the School of Health Sciences and Biomedical Engineering,Hebei University of Technology, China. His current research focuses on bio-impedance imaging and intelligent medical image analysis.

Thomas Lukasiewicz is a Professor at Institute of Logic and Computation, TU Wien, Vienna, Austria, and Department of Computer Science, University of Oxford, UK. He currently holds an AXA Chair grant on "Explainable Artificial Intelligence in Healthcare" and a Turing Fellowship at the Alan Turing Institute, London, UK, which is the UK's National Institute for Data Science and Artificial Intelligence. He received the IJCAI-01 Distinguished Paper Award, the AIJ Prominent Paper Award 2013, the RuleML 2015 Best Paper Award, and the ACM PODS Alberto O. Mendelzon Test-of-Time Award 2019. He is a Fellow of the European Association for Artificial Intelligence (EurAI) since 2020. His research interests are especially in artificial intelligence and machine learning.

Zhenghua Xu received a M.Phil. in Computer Science from The University of Melbourne, Australia, in 2012, and a D.Phil in computer Science from University of Oxford, United Kingdom, in 2018. From 2017 to 2018, he worked as a research associate at the Department of Computer Science, University of Oxford. He is now a professor at the Hebei University of Technology, China, and an awardee of "100 Talents Plan" of Hebei Province. He has published more than 30 papers in top AI or database conferences and journals, e.g., Nature Neuroscience, IEEE TMI, MedIA, NeurIPS, AAAI, IJCAI, ICDE, etc. His current research focuses on intelligent medical image analysis, deep learning, reinforcement learning and computer vision.