Biomedical Signal Processing and Control 62 (2020) 102120

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

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

A novel 3D medical image super-resolution method based on densely connected network

Wei Lu, Zhijin Song, Jinghui Chu [*]

School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, PR China

ARTICLE INFO

Keywords:
3D medical image super-resolution
3D dilated convolution
3D channel-wise and spatial attention
3D asymmetric convolution

ABSTRACT

High-quality and high-resolution medical images can help doctors make more accurate diagnoses, but the resolution of medical images is often limited by a variety of factors such as device, operation and compression rate. To deal with this issue, in this paper, we propose a novel densely connected network for super-resolution reconstruction of 3D medical images. In order to obtain multiscale information, we first adopt 3D dilated convolution with different dilation rates to extract shallow features. To better handle these hierarchical features, we combine local residual learning with densely connected layers, which apply 3D asymmetric convolution to improve performance without increasing inference time. Meanwhile, an improved attention module, which considers both channel-wise and spatial information, is applied to enhance attention of the channels and regions with more high-frequency details. Finally, a feature fusion module which contains three parallel dilated convolution is applied to fuse hierarchical features. Compared with the state-of-the-art methods, such as SRCNN, FSRCNN, SRResnet, DCSRN, ReCNN and DCED, our experimental results show that the proposed method has better performance in both objective metrics and visual effect.

1. Introduction

With the development of computer vision technology, medical images play a vital role in clinical medical applications. For example, magnetic resonance imaging (MRI) has been widely used in clinic. MRI offers high resolution in vivo imaging and rich functional and anatomical multimodality tissue contrast [1]. It is mainly used for soft tissue [2] and can directly get tomographic images of transverse plane, sagittal plane, coronal plane and various inclined plane. However, the hardware upgrade of MRI relies on physics, which cause a long cycle to carry out technological innovation [2]. In addition to the limitations of physical, technical and economic factors [3], the acquisition time of the MRI and the motility of organs also diminish the image quality. Low resolution images may reduce the visibility of important pathological details and further affect the diagnostic results [3]. Therefore, improving the resolution of medical images is an important research topic and many studies have shown that super-resolution (SR) can provide a relatively cheaper solution.

Image super-resolution (SR) reconstruction is a typical ill-posed problem [4] and its main purpose is to reconstruct high-resolution (HR) images from low-resolution (LR) images. A variety of methods have been proposed. These methods can be roughly divided into three categories: interpolation-based methods, reconstruction-based methods and learning-based methods [2].

The interpolation-based methods generally include nearest neighbor interpolation, bilinear interpolation [5], bicubic interpolation [6], and some subsequent methods [7,8]. These traditional methods are simple and fairly effective, but regularly fail to restore more high-frequency information. In the reconstruction-based methods, P. Purkait et al. [9] proposed the maximum a posterior probability (MAP) method to constrain the solution space using prior information. However, there is little prior information available when the input image size is small.

The traditional machine learning-based methods can be categorized into dictionary learning methods, regression methods and sparse based methods. Rueda et al. [10] and Bhatia et al. [11] proposed methods based on coupling dictionary learning to learn HR and LR dictionaries from MRI to generate SR images. These methods depend on learning dictionaries on external LR–HR patches and benefit from sparse constraints to express the relationship between LR and HR images [12]. In the regression methods, many models are directly used to predict some pixels in HR images. Wu et al. [13] solved the mapping error problem by using kernel partial least squares regression model. In order to reduce the calculation expense, the literatures [14–16] obtained the closed-form solution of SR process by ridge regression which uses L2-norm to regulate sparse coefficients. Yang et al. [17] first used sparse coding (SC) to solve the super-resolution problem and further improved

[*] Corresponding author.
E-mail address: cjh@tju.edu.cn (J. Chu).

https://doi.org/10.1016/j.bspc.2020.102120
Received 1 February 2020; Received in revised form 1 July 2020; Accepted 1 August 2020
Available online 19 August 2020

the performance and made the dictionary simpler. Subsequently, Tian, Yang, Ying, Liu, Ben et al. [18–22] proposed several methods based on the sparse method and obtained better results. Wei et al. [16] proposed a sparse medical image super-resolution method. With the rapid development of deep learning, many deep learning-based methods have been proposed. Dong et al. [23] introduced a super-resolution method using convolutional neural network (SRCNN) which includes feature extraction, nonlinear mapping and reconstruction for the first time. Later, various works have been proposed to improve SR performance via residual learning [24,25], recursive learning [25,26]. However, these methods are more time-consuming because the input size is the same as the final output size. FSRCNN [27] extracted upsampled spatial resolution only at the end of the processing pipeline via a deconvolution layer. Ledig et al. [28] designed a network structure named SRResnet with 16 residual blocks, which was further improved by EDSR [29] via removing the BN layer and using the residual scaling to speedup the training process. The literature [30] introduced dense neural network to the field of image super-resolution and proposed the super-resolution dense network (SRDensNet), which can avoid gradient vanishing, enhance feature propagation, support feature reuse, and reduce parameters.

For the super-resolution reconstruction of 3D medical images, Chen et al. [31] proposed a simple densely connected network (DCSRN) for 3D brain MRI. Pham et al. [32] proposed a deep 3D convolution neural network (ReCNN) for super-resolution of brain MRI data. Du et al. [33] proposed a dilated encoder–decoder network (DCED) to reconstruct high-resolution MRI. These methods generally focus on designing a deeper or wider network to learn more features. Yet, it is a great challenge to restore high-frequency details and fully utilize hierarchical features.

In this paper, we propose a super-resolution method for 3D medical images based on the densely connected layers. The experimental results show that our method has superior performance in both the objective metrics and the visual effect. The main contributions of our method can be summarized as described below.

- The 3D dilated convolution module (DCM) with different dilation rates is adopted to increase the receptive field and obtain multiscale information without extra parameters.
- We introduce a 3D channel-wise and spatial attention module (CSAM) to focus on the more important features and to improve the learning ability of the model.
- We propose a local residual dense attention module (LRDAM) which includes a bottleneck layer, a residual dense module (RDM) and a CSAM. The bottleneck layer can reduce the data dimension so as to further reduce parameters. The RDM can merge hierarchical features so as to further enhance the learning ability of the network. Besides, local residual learning not only transfers abundant image details to the back layers, but also helps gradient flow which can simplify the training of the deep network.
- The 3D asymmetric convolution (AC) is adopted in place of standard convolution. Due to the additivity of convolution, we can fuse 3D asymmetric convolutions into standard convolution before testing, which improve the performance without increasing the time of inference.
- The feature fusion module (FFM), which includes parallel dilated convolution, is used at the end of the network to merge hierarchical features.

The rest of this paper is organized as follows: Section 2 reviews the present work including CNN-based SR models, MRI super-resolution and attention mechanism. Section 3 describes the structure of our proposed network and the key components in detail. Section 4 discusses the experimental results. Finally, Section 5 concludes the paper.

2. Related work

2.1. CNN-based SR models

In recent years, with the development of deep learning, many CNN-based SR models have been proposed. Dong et al. [23] introduced a three-layer convolutional network called SRCNN which implements end-to-end learning and achieves better performance than traditional methods. However, SRCNN can only restore limited information, convergence slowly and lose multiscale features. Later, Kim et al. [24,25] used global residual learning to train a deeper network which contains 20 convolutional layers, and for the first time introduced recursive learning to SR. Inspired by the above literatures, Tai et al. [26] utilized a deep recursive residual network (DRRN) to reuse parameters and make the training process stable. These methods have achieved better performance, but they keep the size of input consistent with the final high-resolution output which results in great time-consuming and memory consumption.

To solve the aforementioned problems, two kinds of methods were further proposed, i.e. deconvolution and transpose convolution. Inspired by that, Lim et al. [29] proposed a very deep network called enhanced deep super resolution (EDSR) by removing the BN layer and stacking residual blocks. Zhang et al. [34] combined the idea of residual learning and densely connection and then proposed residual dense network (RDN) to focus on the hierarchical features which can obtain more information for reconstruction. These methods tended to make the network deeper and wider to get better performance at the cost of time and memory consumption. As the network becomes deep, features from different convolution layers have different levels of receptive field, but these methods ignore the inherent correlations among the features from different convolutional layers.

2.2. MRI super-resolution methods

For the super-resolution reconstruction of medical images, J. Song et al. [35,36] proposed an adaptive quad-tree decomposition method using K-means to determine the clustering center and using the mapping of LR image patches and HR image patches to acquire SR images. However, it is time-consuming and hard to select a suitable value of k. Dou et al. [2] proposed a method based on minimum error regression using a random forest which learns the internal relations from the same cluster and then chooses the best model. Nevertheless, many models belonging to this method occupy great amount of memory, and overfitting may occur due to high noise level.

In the deep learning field, Yang et al. [37] employed a deep convolutional neural network for single medical image super-resolution, which uses randomized rectified linear unit (RReLU) and Nesterov's accelerated gradient (NAG) to get better performance. A new deep network model was proposed which takes full advantage of the fact that all medical images basically have distinct repetitive structure and a large black border without any texture information [38]. It added a convolution layer to carry out secondary feature extraction and overlapped pooling layers to highlight the important features. To make the process faster, Zhang et al. [39] proposed a fast medical image super-resolution (FMISR) method, which extracts the features from three hidden layers and uses different activation functions in the same structure. However, none of the above methods can obtain multiscale information. Therefore, a multiscale super-resolution model based on deep residual networks was then proposed [36], which can reconstruct different types of medical images with different scales. Nevertheless, these methods do not make full use of the features from different layers and they treat the features equally thus increasing the representational burden of CNN models.

2.3. Attention mechanism

Attention is an inherent signal processing mechanism in the human brains, which quickly selects areas that need to be focused on from visual signals, and then focuses on the details within these areas. Through attention mechanism, valuable information can be picked out from a large amount of information, only using limited resource of brain. In the same way, attention mechanism can be used in neural networks in order to focus on the specific channels or regions and recalibrate the most informative and important parts of the inputs.

Recently, attention modules have been adopted in various tasks. In image captioning, Xu et al. [40] proposed an attention model which includes a hybrid "soft" deterministic and a "hard" stochastic attention mechanism. Chen et al. [41] proposed a cascade spatial attention and channel attention module, in which the second attention is operated in the first attention, and thus the network will pay more attention to the channels and regions which have more high-frequency details. In the visual question answering tasks, Yang et al. [42] and Xu et al. [43] proposed models, in which the second attention is based on the feature mapping of the first attention. By studying the relationship of the convolution feature channels in the network, Hu et al. [44] proposed a squeeze and excitation (SE) block for image classification, which is able to adaptively assign different weights to different channels, thus further improve the network performance. Inspired by SE network, Zhang et al. [45] proposed a very deep residual network with channel attention in the SR tasks. However, the attention mechanism of the most aforementioned methods is only applicable to the last layer, which means the difference between each receptive field is very limited. As a result, the spatial attention is not obvious. Therefore, we combine the channel attention and the spatial attention into the local residual dense module to accelerate the network convergence and further improve the performance of the network.

3. The proposed method

3.1. Network structure

As shown in Fig. 1, our network mainly consists of four parts: shallow feature extraction module (SFEM), dilated convolution module (DCM), local residual dense attention module (LRDAM) and feature fusion module (FFM). I_{LR} and I_{SR} are the input and output of the network respectively. First, we use a convolutional layer to extract shallow features. This process can be denoted as

$$I_0 = H_{SFEM}(I_{LR}), \tag{1}$$

where $H_{SFEM}(\cdot)$ represents $3 \times 3 \times 3$ convolution operation. Then, the shallow feature I_0 is fed to the DCM which consists of six dilated convolutional layers whose dilation rates are 1, 2, 3, 1, 2, 3 respectively. In addition, the DCM also introduces residual learning to help gradient flow, denoted as

$$I_1 = I_0 + H_{DCM}(I_0), \tag{2}$$

where $H_{DCM}(\cdot)$ represents several dilated convolution operation. The dilated convolution can be considered to be helpful to increase the receptive field without increasing the computational complexity and also helpful to obtain multiscale information. Then I_1 is fed to the LRDAM to acquire deeper features, denoted as

$$I_{LRDAM} = H_{LRDAM}(I_1), \tag{3}$$

where I_{LRDAM} refers to the output of the LRDAM. The LRDAM consists of a bottleneck layer, a RDM and a CSAM. Thus, our method can obtain hierarchical features from different stages, which can alleviate the problems of gradient vanishing problem and enhance the representation ability of the network. The FFM consists of three parallel dilated convolutional layers and the global residual learning is used to make the training stable. The input of the FFM can be denoted as

$$I_{IN} = I_{LRDAM} + I_0. \tag{4}$$

Thus, the output of the FFM is the final output, which can be denoted as

$$I_{SR} = H_{FFM}([H_1 I_{IN}, H_2 I_{IN}, H_3 I_{IN}]), \tag{5}$$

where $H_1(\cdot)$, $H_2(\cdot)$, $H_3(\cdot)$ represent dilated convolution operation and $H_{FFM}(\cdot)$ represents $3 \times 3 \times 3$ convolution operation. The network is optimized by minimizing the difference between HR images I_{HR} and SR images I_{SR}. There are various loss functions to measure the difference, such as L_1 loss function, L_2 loss function, etc. Here, we use L_2 loss function as our metrics, denoted as

$$L = \frac{1}{n}\sum_{v=1}^{n}(I^{HR} - I^{SR})^2, \tag{6}$$

where I^{HR} refers to the HR images, I^{SR} refers to the SR images, and n is the batch size.

3.2. Dilated convolution module

It is well-known that the receptive field is an important factor in the SR task, and its size can be formulated as

$$l_k = l_{k-1} + ((f_k - 1) \times \prod_{i=1}^{k-1} s_i), \tag{7}$$

where l_{k-1} is the receptive field of the k-1-th layer, f_k is the filter size of the kth layer and s_i is the stride. A larger receptive field means that more comprehensive features can be extracted. In image super-resolution tasks, the context information contributes to generate realistic details. There are usually two methods to enlarge the receptive field which are increasing the filter size, and increasing the depth of the network [46], but they will introduce more parameters and increase the computational burden. Lu et al. [47] used dilated convolution instead of general convolution in SR models to increase the receptive field, and finally obtained much better performance. The actual convolution kernel size of dilated convolution K can be described as

$$K = k + (k - 1)(r - 1), \tag{8}$$

where k is the original convolution kernel size, and r is the dilation rate. In a sense, the dilated convolution with a dilation rate greater than 1 can be interpreted as a sparse filter. As shown in Fig. 2, similar to 2D dilated convolution, the receptive field of the $3 \times 3 \times 3$ convolution with the dilation rate of 1 equals $3 \times 3 \times 3$, the receptive field of the $3 \times 3 \times 3$ convolution with the dilation rate of 2 equals $7 \times 7 \times 7$ and the receptive field of the $3 \times 3 \times 3$ convolution with the dilation rate of 4 equals $15 \times 15 \times 15$. In summary, the dilated convolution can increase the receptive field and obtain multiscale context information. Therefore, we set the dilation rates with sawtooth values, such as the loop [1, 2, 3, 1, 2, 3].

3.3. Local residual dense attention module

There are N LRDAMs in our proposed network. As shown in Fig. 3, each LRDAM contains a bottleneck layer, a residual dense module (RDM) and a channel-wise and spatial attention module (CSAM). Supposing the output of the nth LRDAM is F_n, the LRDAM can be denoted as

$$F_n = H_n([I_1, F_{n-1}]), \tag{9}$$

where I_1 refers to the output of the dilated convolution module, $H_n(\cdot)$ represents the function of the nth LRDAM, and F_{n-1} refers to the output of the (n-1)-th LRDAM. The bottleneck layer refers to the leftmost $1 \times 1 \times 1$ convolutional layer, which can reduce the data dimension

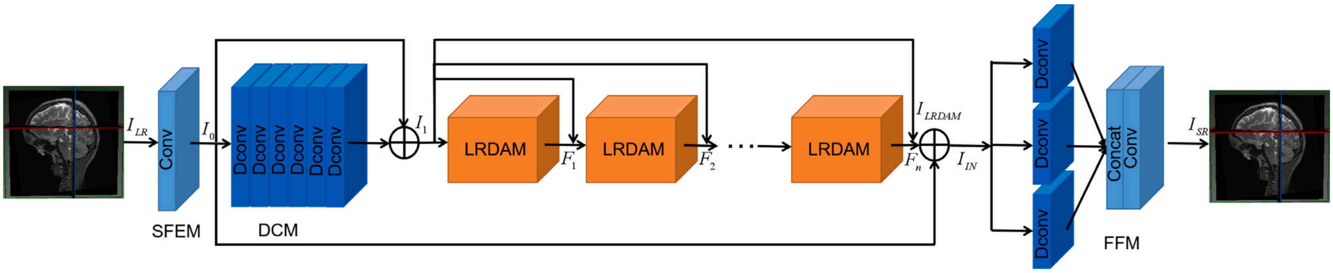

Fig. 1. The architecture of our network which mainly consists of four parts: shallow feature extraction module (SFEM), dilated convolution module (DCM), local residual dense attention module (LRDAM), and feature fusion module (FFM).

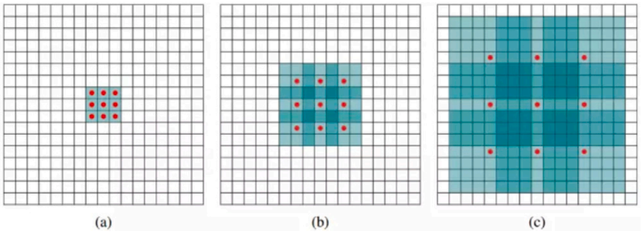

Fig. 2. Dilated convolution with different dilation rate. (a) shows the convolution of 3 × 3 and the dilation rate is 1. (b) shows the convolution of 3 × 3 with the dilation rate being 2, and the receptive field equals 7 × 7. (c) shows the convolution of 3 × 3 with the dilation rate being 4, and the receptive field equals 15 × 15.

and increase the non-linearity. The output of the bottleneck layer can be denoted as $F_{n,-1}$, and this process can be described as

$$F_{n,-1} = H_0(F_{n-1}), \tag{10}$$

where H_0 represents the 1 × 1 × 1 convolution operation.

As shown in Fig. 4(b), a standard convolution can be split into four convolutions: one 3 × 3 × 3 convolution, one 3 × 3 × 1 convolution, one 1 × 3 × 3 convolution, and one 3 × 1 × 3 convolution. Then, the 4-way outputs can be combined together by an element-wise addition. After training, the 3D asymmetric convolution (AC) can be converted into a standard convolution without extra calculation. As shown in Fig. 4(a), there are k densely connected layers in each RDM. The output of the kth dense connected layer can be formulated as

$$F_{n,k} = H([F_{n,0}, F_{n,1}...F_{n,k-1}]), \tag{11}$$

where $H(\cdot)$ denotes the nonlinear function which contains a BN, a ReLu and a convolutional layer. $[F_{n,0}, F_{n,1}...F_{n,k-1}]$ represents the concatenation of the feature maps produced by the (k-1)-th densely connected layer. This process is beneficial for extracting local dense features.

With the fusion of features, the network will grow larger and larger, and consequently it will be more difficult to train. We hereby introduce a 1 × 1 × 1 convolutional layer at the end of a RDM to control the output information. The final output of the nth RDM can be obtained by

$$F_{n,dm} = H^{n}_{BL}([F_{n,0}, F_{n,1}...F_{n,k}]), \tag{12}$$

where $[F_{n,0}, F_{n,1}...F_{n,k}]$ represents the concatenation of the feature maps produced by the kth densely connected layer. H_{BL} represents the 1 × 1 × 1 convolution operation. Then, there is a local skip connection which can provide a fast and stable training, and thus the output of a RDM can be denoted as

$$F_{n,RDM} = F_{n,dm} + F_{n,-2}, \tag{13}$$

As shown in Fig. 3, a RDM is followed by a CSAM, which includes a channel-wise attention module (CAM) and a spatial attention module (SAM). Supposing the output of CSAM is $F_{n,CSA}$, which can be denoted as

$$F_{n,CSA} = H_{CSA}(F_{n,RDM}), \tag{14}$$

where $H_{CSA}(\cdot)$ represents the function of CSAM. More details will be given in Section 3.4. In order to improve information flow and alleviate network degradation, local residual learning is introduced in the nth LRDAM. The output of LRDAM can be formulated as

$$F_n = [I_1, F_{n-1}] + F_{n,CSA}, \tag{15}$$

where $[I_1, F_{n-1}]$ represents the input of LRDAM.

3.4. Channel-wise and spatial attention module

Usually, different regions of an image contain different kinds of information. For example, the edges have more high-frequency information, whereas the smooth regions have more low-frequency information. Consequently, different kinds of information have different contributions to the image super-resolution reconstruction, and therefore they should be treated differently so as to enhance the learning ability of the network and get better performance.

3.4.1. Channel-wise attention module

Many CNN methods lack the ability to distinguish different types of information and instead treat them equally which results in limited representation ability of the model [48]. In order to make full use of all kinds of features, the channel-wise attention was proposed. The essence of channel-wise attention is the scaling process, which means the outputs of different channels are multiplied by different weights, so that more attention could be paid to the key channels. The structure is shown in Fig. 3. For channel-wise attention module (CAM), we first squeeze each feature map $F_{n,RDM} = [F_{n,RDM_1}, F_{n,RDM_2}...F_{n,RDM_c}]$ by global average pooling to obtain the channel feature $v = [v_1, v_2...v_c]$. This process can be described as

$$v_c = \frac{1}{HWL}\sum_{i=1}^{H}\sum_{j=1}^{W}\sum_{k=1}^{L} F_{n,RDM_c}(i,j,k), \tag{16}$$

where $F_{n,RDM_c}(i,j,k)$ represents the value of the cth channel, c is the number of channels, and H, W and L are the height, width and length of (i,j,k). Then the channel feature is fed to two convolutional layers and the corresponding activation layers. The CAM is described as

$$F_{n,CA} = F_{n,RDM} \otimes \sigma(W_{CA2})\delta(W_{CA1}v), \tag{17}$$

where W_{CA1} and W_{CA2} represent the weights of the first and second layers, σ and δ denote the sigmoid function and ReLu function respectively, and $\otimes$ denotes element-wise product. The first convolutional layer uses dimension reduction ratio of 16 to reduce the channel dimension, and then obtains the feature map of H × W × L × C/16. The second convolutional layer increases the feature map by ratio 16, and then obtain the feature map with dimension H × W × L × C. After two convolutional and corresponding activation layers, the module can adaptively adjust the channel features according to the information of the input channel, and can enhance the channel with high-frequency details while suppressing the channel with low-frequency contents.

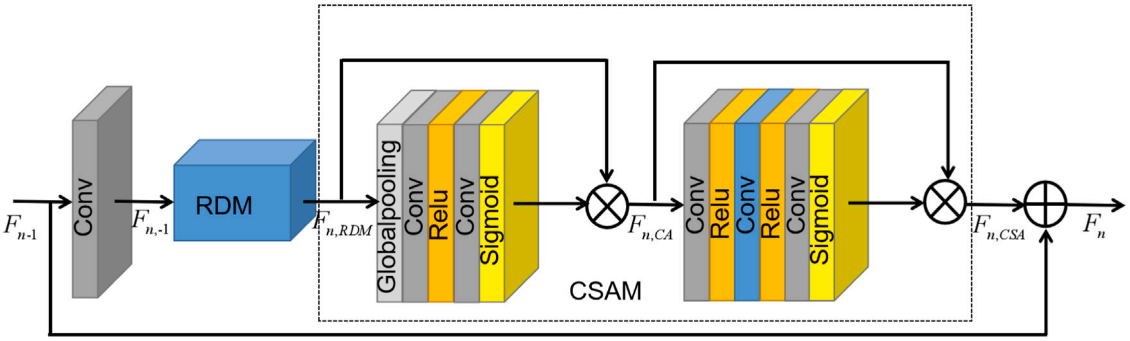

Fig. 3. The structure of local residual dense attention module (LRDAM) which contains a bottleneck layer, a residual dense module (RDM) and a channel-wise and spatial attention module (CSAM).

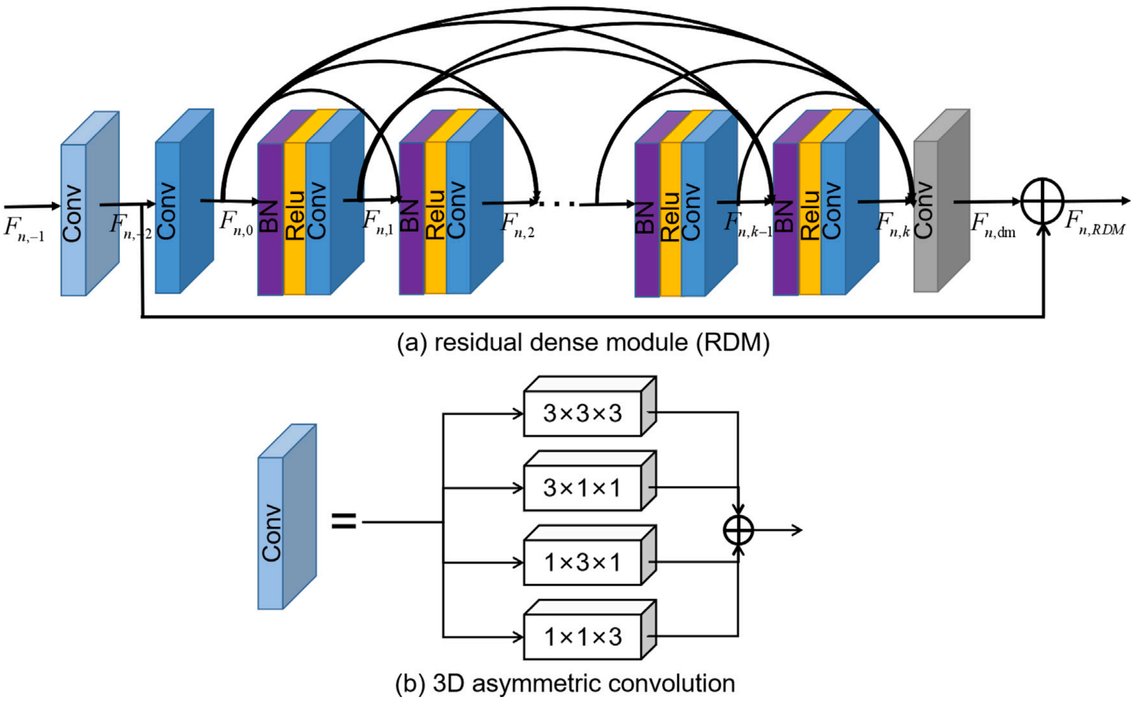

Fig. 4. The structure of residual dense module (RDM) whose first convolution is replaced by 3D asymmetric convolution.

3.4.2. Spatial attention module

Spatial attention focuses more on the regions which are more important to reconstruct [48]. As shown in Fig. 3, we first use a 1 × 1 × 1 convolutional layer to increase the dimension, and then use a 3 × 3 × 3 convolutional layer to obtain feature maps. Finally, a 1 × 1 × 1 convolutional layer is used to get a weight matrix, and a sigmoid function is used to normalize it to [0,1]. The definition of the spatial attention module (SAM) is as

$$F_{n,CSA} = F_{n,CA} \otimes \sigma(W_{SA}F), \tag{18}$$

$$F = \delta(W_{SA2}\delta(W_{SA1}F_{n,CA})), \tag{19}$$

where $F_{n,CA}$ represents the input of the spatial attention module. W_{SA1}, W_{SA2}, W_{SA} are the parameters of the 1 × 1 × 1 convolutional layer, the 3 × 3 × 3 convolutional layer and the 1 × 1 × 1 convolutional layer respectively. δ denotes the ReLu function, σ denotes the sigmoid function, and $\otimes$ denotes element-wise product.

4. Experimental results

In this section, we first introduce the datasets we used and the preprocessing method. Second, we provide the implementation details including the settings of the experimental environment and the parameters of the network. Third, we compare the different performances of various component combinations. Finally, we compare our method with the state-of-the-art methods.

4.1. Datasets and preprocessing method

In our experiments, we used the HCP (Human Connectome Project) dataset [49] and the ADNI (Alzheimer's Disease Neuroimaging Initiative) dataset (see www.adni-info.org for details) to conduct comparative studies.

The HCP dataset consists of 3D MPRAGE images from 67 volunteers. The matrix size of these scans are 320 × 320 × 256 and spatial resolution is 0.7 mm isotropic. The scans were acquired from Siemens 3T platforms. The ADNI dataset consists of 3D MPRAGE images from 230 volunteers aged from 50 to 90. These MRI scans were acquired from Siemens platforms and Philips Medical System Scanners. These two datasets were randomly divided into training set, validation set and test set at the ratio of 8:1:1. Due to the limited memory size of the GPU, each 3D MPRAGE image was split into 64 × 64 × 64 patches which include randomly selected location and were randomly rotated by 90° for data enhancement.

Table 1
Comparison of dilated convolution parameters with [1, 2, 3, 1, 2, 3] and [1, 2, 5, 1, 2, 5].

Parameters	PSNR	SSIM
1-2-5-1-2–5	34.9479	0.9508
1-2-3-1-2–3	35.0812	0.9510

In general, the magnetic resonance data collected by the instrument are not a time domain function, but a frequency domain function. Then, IFT is performed on the data in k-space to obtain the time domain function, and the spatial localization of each proton is resolved to obtain the magnetic resonance image. Therefore, the acquisition of k-space data is directly related to the spatial resolution of magnetic resonance images. In the sampling process, more the number of sampling points mean more pixels in the frequency and phase encoding direction. Practically, the number of sampling points may be reduced in order to save the scan time, and meanwhile the signal-to-noise ratio will decrease. If F (U, V, W) is limited along the W-axis, the resulting image will have an isotropic digital resolution on the xy-plane and a low resolution on the z-axis [50]. For a more realistic simulation of the data acquisition process, we can get a low-resolution image through the k-space truncation.

4.2. Implementation details

We trained our models with the 64 × 64 × 64 patches randomly selected from HR images. Each patch is normalized to percentile values and its gray values are clipped to the range of [0,1]. Then LR patches are obtained by the k-space truncation. We used a Fast Fourier Transform (FFT) to zero along three axes, and then applied an inverse FFT. The whole network has six LRDAMs and each RDM has four densely connected layers. The training process can used the ADAM (Adaptive Moment Estimation) optimizer for network optimization, and the settings are β_1=0.9, β_2=0.999, $\epsilon = 10^{-8}$. The initial learning rate is set to 0.0001, and decays to 0.95 at every 26500 iterations. The epoch is set to 50. During training, we saved the model checkpoint which has the lowest validation loss. To ensure the uniformity of the experimental environment, the models were implemented with TensorFlow 1.13.1 on Ubuntu 18.04 operating system with CUDA 10.0 and CUDNN 7.0 libraries, and trained on a RTX 2080Ti GPU. The workstation has 64 GB system memory and 11 GB video memory. In addition, we used peak signal-to-noise ratio (PSNR) and the structural similarity (SSIM) to measure our model's performance. Higher PSNR and SSIM mean better image quality.

4.3. Model analysis

(1) Dilated Convolution Module

As shown in Fig. 1, the first convolution is followed by the dilated convolution module (DCM) to extract more comprehensive features and obtain multiscale information without increasing computational complexity. To compare the contribution of different parameters of the DCM, we show our study in Table 1. It can be observed that the DCM with the parameters of [1, 2, 3, 1, 2, 3] has higher PSNR and SSIM values compared with that with [1, 2, 5, 1, 2, 5]. Our explanation is as follows: (1) The adjacent pixels obtained by a certain layer come from a mutually independent subset of the previous layer, and there is no interdependence between them, thus there is no correlation between the convolution results of this layer. That is to say, continuous information is lost. (2) The use of a greater dilation rate is more harmful to small objects. In summary, it proves that applying the DCM with appropriate parameters can improve the performance of our network.

(2) Channel-wise Attention Module

Channel-wise attention can fit complex functions, and further improve the representation ability of the network. The channel-wise attention module (CAM) firstly uses the global average pooling to squeeze channel features to obtain feature channel vectors, and then sends the feature channel vectors to two convolutional layers and two activation layers. The activation layers adopt ReLu function and sigmoid function respectively. The reasons for using two convolutional layers are as follows: First, it can introduce more nonlinearities, and thus can better fit complex correlations among channels. Second, it minimizes the amount of parameters and calculations. Finally, it can obtain normalized weights in the range of [0, 1]. As shown in Fig. 5, we compared the performances of CAM at different positions of the network. As shown in Table 2, InM outperforms OuM and IOM by 0.5668 dB (PSNR) and 0.0051 (SSIM), and 0.2849 dB (PSNR) and 0.0022 (SSIM) respectively. Theoretically, the channel-wise attention inside the module can pay different attention to hierarchical features and obtains more important information.

Table 2
Comparison of different position of channel-wise attention module.

CAM location	PSNR	SSIM
Outside the LRDAM (OuM)	34.1321	0.9430
Inside the LRDAM (InM)	34.6989	0.9481
Inside and Outside the LRDAM (IOM)	34.4140	0.9459

Table 3
Comparison of different components combination in our network.

Components	Different combinations of components				
+ CAM	√	√	√	√	√
+ DCM	×	√	√	√	√
+ SAM	×	×	√	√	√
+ AC	×	×	×	√	√
+ FFM	×	×	×	×	√
PSNR	34.6989	35.0812	35.2342	35.4633	35.6413
SSIM	0.9481	0.9510	0.9535	0.9553	0.9569

(3) Contribution of Different Components

In this part, we analyze the contributions of different components, namely channel-wise attention module (CAM), dilated convolution module (DCM), spatial attention module (SAM), asymmetric convolution (AC) and feature fusion module (FFM). Table 3 compares the performance of different combinations of the aforementioned components. The first column presents the different components and the metrics of PSNR and SSIM. The other five columns present different combinations of these components. First, we only added CAM after the RDM. Compared with the CAM added outside the LRDAM, there is an improvement of 0.5668 dB (PSNR) and 0.0051 (SSIM). Then, we added 3D dilated convolution with dilation rates of [1, 2, 3, 1, 2, 3] based on the first step. It has an improvement of 0.3823 dB (PSNR) and 0.0029 (SSIM). This can result from the expanded receptive field which can help to obtain more comprehensive features. Besides, it indicates that dilated convolution plays a key role for the performance. Third, we added SAM after the CAM, and there is an improvement of 0.1530 dB (PSNR) and 0.0025 (SSIM). This proves that the SAM helps to improve the performance by paying more attention to the regions which are more important to reconstruct. Fourth, we used AC to replace standard convolution in the RDM. It has an improvement of 0.2291 dB (PSNR) and 0.0018 (SSIM). We can conclude from Table 3, that fusing different branch features can get better performance. In addition, it does not introduce additional parameters during testing, because the convolution kernel is additive. Finally, FFM is used to make feature fusion, which has an improvement of 0.178 dB (PSNR) and 0.0016 (SSIM).

4.4. Comparison with the state-of-the-arts

We further compared our method with several state-of-the-art methods which have obtained great performance on SR tasks, including SRCNN [23], FSRCNN [27], DCSRN [31], ReCNN [32], DCED [33]

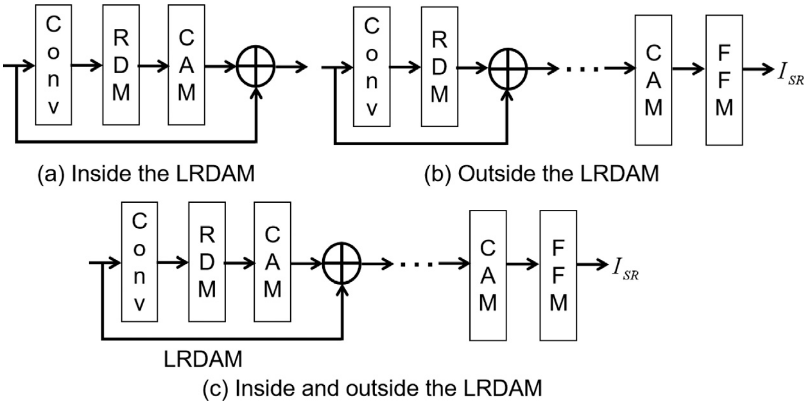

Fig. 5. The channel-wise attention module (CAM) can be inserted at different position of the network.

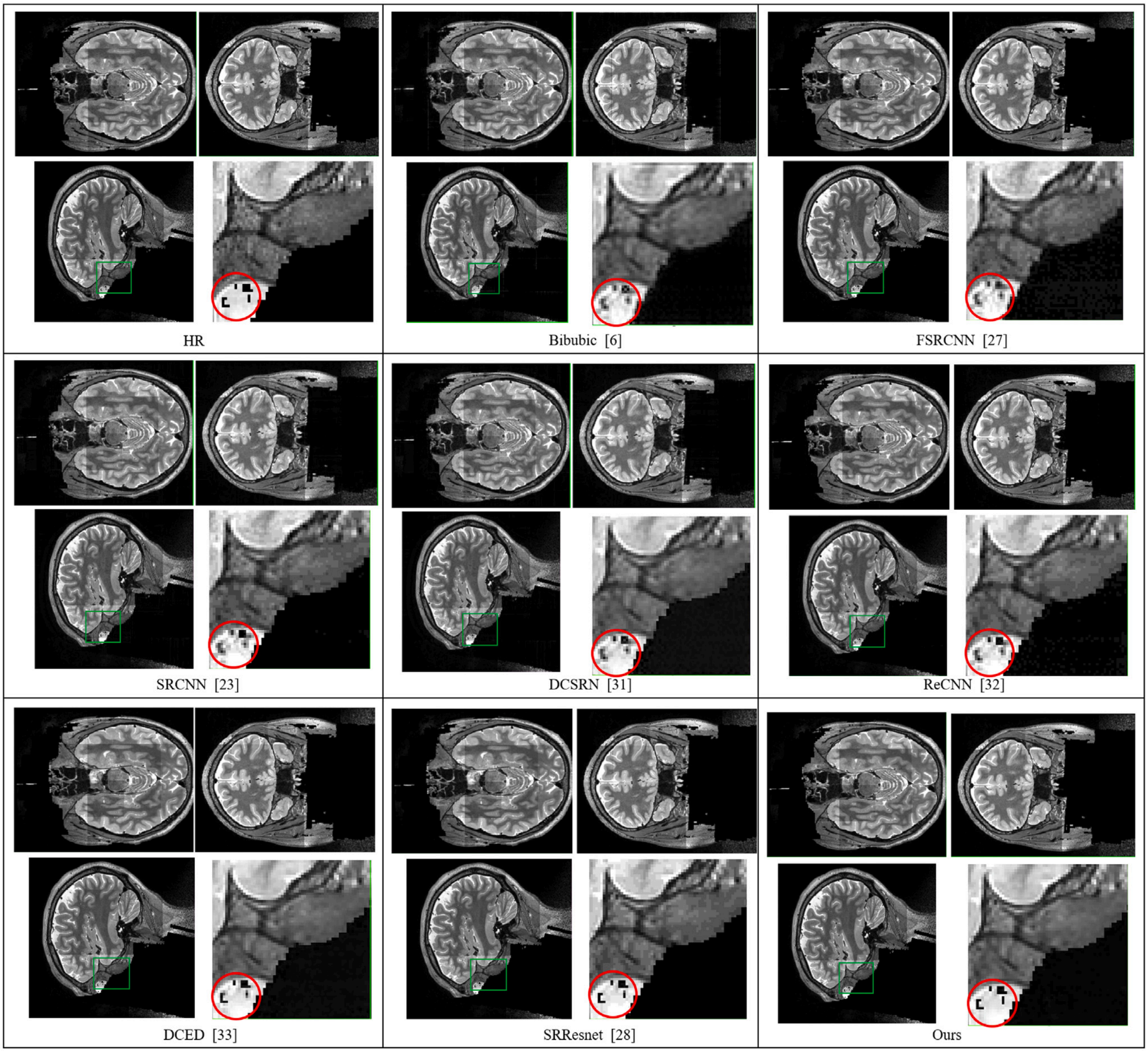

Fig. 6. Visual evaluation on the HCP dataset. The 3D medical images consist of sagittal plane, coronal plane and transverse plane. The red circle denotes the enlarged display of a region inside the green box.

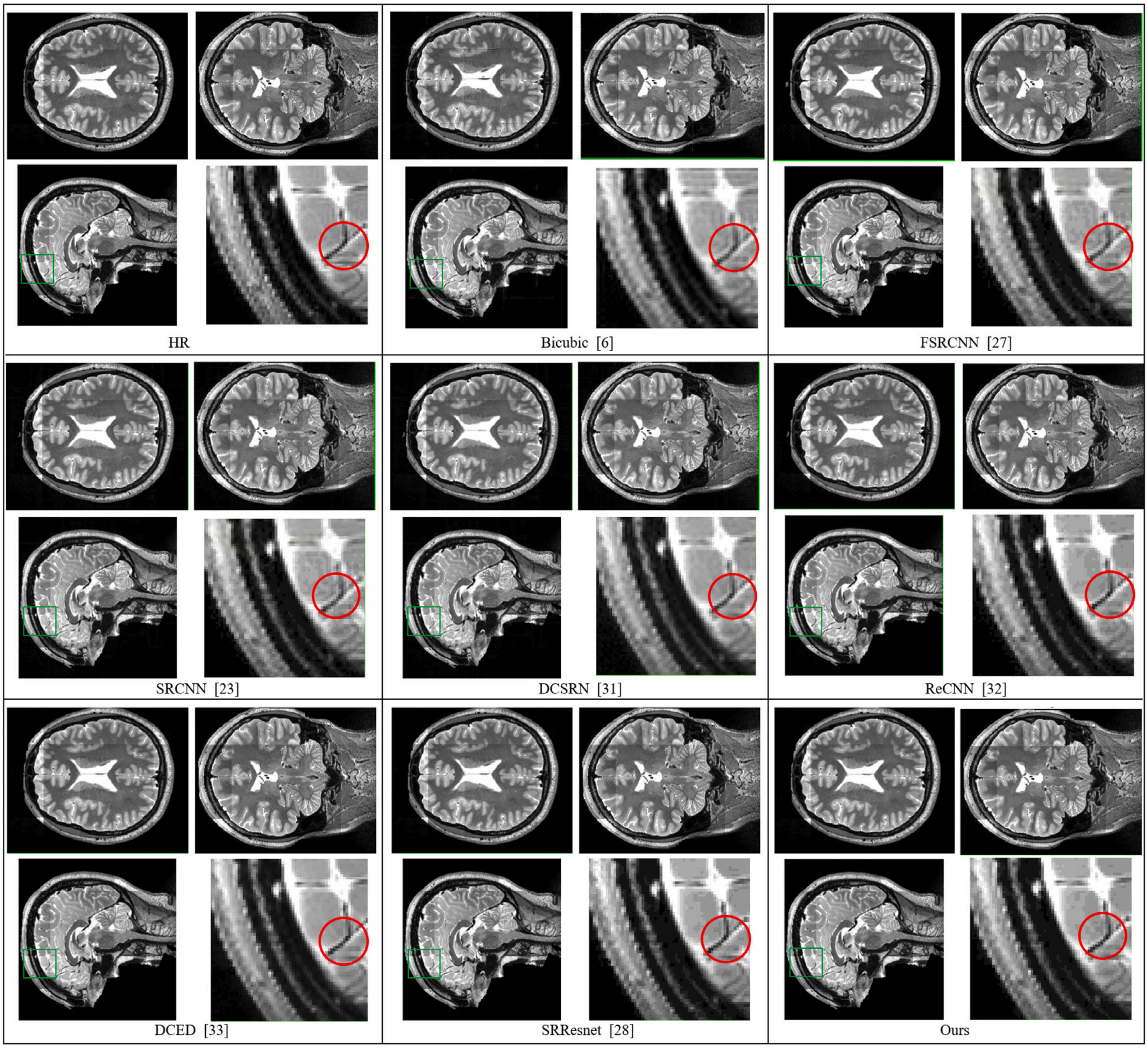

Fig. 7. Visual evaluation on the HCP dataset. The 3D medical images consist of sagittal plane, coronal plane and transverse plane. The red circle denotes the enlarged display of a region inside the green box.

Table 4
Comparison with the state-of-the-arts.

Methods	HCP PSNR/SSIM	ADNI PSNR/SSIM
3D Bicubic [6]	31.0214/0.8773	31.8573/0.9175
3D FSRCNN [27]	32.1622/0.9067	33.3404/0.9376
3D SRCNN [23]	32.9607/0.9290	33.5428/0.9477
DCSRN [31]	33.7876/0.9394	34.4251/0.9541
ReCNN [32]	33.8348/0.9407	35.0056/0.9619
DCED [33]	34.7935/0.9504	35.1022/0.9627
3D SRResnet [28]	35.2545/0.9534	35.3522/0.9642
Ours	35.6413/0.9569	35.7631/0.9680

and SRResnet [28], in terms of quantitative evaluation and visual quality. The settings of these methods are based on the original paper, except that two-dimensional convolution are replaced by three-dimensional convolution. To ensure the credibility of the experiment, the preprocessing of the data all uses k-space truncation.

The quantitative evaluations for the aforementioned methods are shown in Table 4. In Table 4, the first column lists the methods, the second column presents the evaluation index on the HCP dataset, and the third column presents the evaluation index on the ADNI dataset. It can be observed that our proposed network outperformed SRResnet by 0.3868 dB (PSNR) and 0.0035 (SSIM). Besides, our network has 6798923 parameters, whereas SRResnet has 13762356 parameters. In summary, the advantages of our network over SRResnet are higher image quality and lower network complexity. We attribute the results to the fact that our method can deal with hierarchical features, and can enhance important features while weakening unimportant features. In summary, our proposed network has better representational capability and can obtain better performance.

The 3D medical images mainly consist of three scans: sagittal plane, coronal plane and transverse plane. Figs. 6–7 presents the visual effects of different methods. It can be observed that our method can reconstruct 3D medical images with more high-frequency details. For example, as shown in Fig. 6, the images obtained by Bicubic interpolation [6] look fuzzy. SRCNN [23], FSRCNN [27], DCSRN [31], ReCNN [32], DECD [33] cannot reconstruct the details inside the red circle. Although SRResnet [28] can reconstruct the details more clearly, but blurry artifacts appear. The images obtained by our method has sufficient details and are closest to the original HR image. Similar results can be seen in Fig. 7.

5. Conclusion

For 3D medical image super-resolution, the main problem is that the models are lack of the distinction ability to deal with hierarchical features. In this paper, we propose a novel method for 3D medical image super-resolution based on the densely connected layers. Our network mainly consists of four parts: shallow feature extraction module (SFEM), dilated convolution module (DCM), local residual dense attention module (LRDAM), and feature fusion module (FFM). In DCM, 3D dilated convolution with different dilation rate can expand the receptive field and further obtain more comprehensive features. LRDAM contains a bottleneck layer, a residual dense module (RDM) and a channel-wise and spatial attention module (CSAM). RDM combines local residual learning with densely connected layers and replaces standard convolution with 3D asymmetric convolution (AC) which can deal with hierarchical features and help information flow. An improved attention module (CSAM) is followed by RDM to treat information differently and focus on the important features. The FFM, which includes parallel dilated convolution, is used at the end of the network to merge hierarchical features. Compared with other state-of-the-art methods, such as SRCNN, FSRCNN, DCSRN, ReCNN, DCED and SRResnet, our experimental results show that the proposed method has better performance in both objective metrics and visual quality. In future, we will carry out the research work from three points. The first point is to introduce knowledge distillation, which uses our network as a teacher network to train a lightweight network for 3D medical image super-resolution. The second point is to find a suitable normalization technique for SR. The third point is to find a better loss function, which explores the potential correlation between 3D medical images.

CRediT authorship contribution statement

Wei Lu: Conceptualization, Methodology, Investigation. **Zhijin Song:** Software, Validation, Writing - original draft. **Jinghui Chu:** Writing - review & editing.

Declaration of competing interest

No author associated with this paper has disclosed any potential or pertinent conflicts which may be perceived to have impending conflict with this work. For full disclosure statements refer to https://doi.org/10.1016/j.bspc.2020.102120.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China under Grant 61802277 and China Postdoctoral Science Foundation Funded Project (2019M651038).

References

[1] Y. Huang, L. Shao, A.F. Frangi, Simultaneous super-resolution and cross-modality synthesis in magnetic resonance imaging, in: Deep Learning and Convolutional Neural Networks for Medical Imaging and Clinical Informatics, Springer, 2019, pp. 437–457.

[2] Q. Dou, S. Wei, X. Yang, W. Wu, K. Liu, Medical image super-resolution via minimum error regression model selection using random forest, Sustainable Cities Soc. 42 (2018) 1–12.

[3] 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: 2019 IEEE 16th International Symposium on Biomedical Imaging, IEEE, 2019, pp. 1669–1673.

[4] X. Wang, D. Zhou, N. Zeng, X. Yu, S. Hu, Super-resolution image reconstruction using surface fitting with hierarchical structure, J. Vis. Commun. Image Represent. 53 (2018) 65–75.

[5] R. Keys, Cubic convolution interpolation for digital image processing, IEEE Trans. Acoust. Speech Signal Process. 29 (6) (1981) 1153–1160.

[6] X. Li, M.T. Orchard, New edge-directed interpolation, IEEE Trans. Image Process. 10 (10) (2001) 1521–1527.

[7] L. Shao, M. Zhao, Order statistic filters for image interpolation, in: 2007 IEEE International Conference on Multimedia and Expo, IEEE, 2007, pp. 452–455.

[8] L. Zhang, X. Wu, An edge-guided image interpolation algorithm via directional filtering and data fusion, IEEE Trans. Image Process. 15 (8) (2006) 2226–2238.

[9] P. Purkait, B. Chanda, Super resolution image reconstruction through Bregman iteration using morphologic regularization, IEEE Trans. Image Process. 21 (9) (2012) 4029–4039.

[10] A. Rueda, N. Malpica, E. Romero, Single-image super-resolution of brain MR images using overcomplete dictionaries, Med. Image Anal. 17 (1) (2013) 113–132.

[11] K.K. Bhatia, A.N. Price, W. Shi, J.V. Hajnal, D. Rueckert, Super-resolution reconstruction of cardiac MRI using coupled dictionary learning, in: 2014 IEEE 11th International Symposium on Biomedical Imaging, IEEE, 2014, pp. 947–950.

[12] D. Mahapatra, B. Bozorgtabar, R. Garnavi, Image super-resolution using progressive generative adversarial networks for medical image analysis, Comput. Med. Imaging Graph. 71 (2019) 30–39.

[13] W. Wu, Z. Liu, X. He, Learning-based super resolution using kernel partial least squares, Image Vis. Comput. 29 (6) (2011) 394–406.

[14] R. Timofte, V. De Smet, L. Van Gool, Anchored neighborhood regression for fast example-based super-resolution, in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 1920–1927.

[15] D. Dai, R. Timofte, L. Van Gool, Jointly optimized regressors for image super-resolution, in: Computer Graphics Forum, vol. 34, (2) Wiley Online Library, 2015, pp. 95–104.

[16] S. Wei, X. Zhou, W. Wu, Q. Pu, Q. Wang, X. Yang, Medical image super-resolution by using multi-dictionary and random forest, Sustainable Cities Soc. 37 (2018) 358–370.

[17] J. Yang, J. Wright, T.S. Huang, Y. Ma, Image super-resolution via sparse representation, IEEE Trans. Image Process. 19 (11) (2010) 2861–2873.

[18] X. Yang, W. Wu, K. Liu, W. Chen, P. Zhang, Z. Zhou, Multi-sensor image super-resolution with fuzzy cluster by using multi-scale and multi-view sparse coding for infrared image, Multimedia Tools Appl. 76 (23) (2017) 24871–24902.

[19] J. Ying, H. Lu, Q. Wei, J.-F. Cai, D. Guo, J. Wu, Z. Chen, X. Qu, Hankel matrix nuclear norm regularized tensor completion for n-dimensional exponential signals, IEEE Trans. Signal Process. 65 (14) (2017) 3702–3717.

[20] Z. Liu, E. Blasch, G. Bhatnagar, V. John, W. Wu, R.S. Blum, Fusing synergistic information from multi-sensor images: An overview from implementation to performance assessment, Inf. Fusion 42 (2018) 127–145.

[21] Z. Liu, E. Blasch, V. John, Statistical comparison of image fusion algorithms: Recommendations, Inf. Fusion 36 (2017) 251–260.

[22] X. Ben, W. Meng, K. Wang, R. Yan, An adaptive neural networks formulation for the two-dimensional principal component analysis, Neural Comput. Appl. 27 (5) (2016) 1245–1261.

[23] C. Dong, C.C. Loy, K. He, X. Tang, Learning a deep convolutional network for image super-resolution, in: European Conference on Computer Vision, Springer, 2014, pp. 184–199.

[24] J. Kim, J. Kwon Lee, K. Mu 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.

[25] J. Kim, J. Kwon Lee, K. Mu Lee, Deeply-recursive convolutional network for image super-resolution, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1637–1645.

[26] Y. Tai, J. Yang, X. Liu, Image super-resolution via deep recursive residual network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3147–3155.

[27] D. Chao, C.L. Chen, X. Tang, Accelerating the super-resolution convolutional neural network, 2016.

[28] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, et al., 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. 4681–4690.

[29] B. Lim, S. Son, H. Kim, S. Nah, K. Mu Lee, Enhanced deep residual networks for single image super-resolution, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017, pp. 136–144.

[30] 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. 4799–4807.

[31] Y. Chen, Y. Xie, Z. Zhou, F. Shi, A.G. Christodoulou, D. Li, Brain MRI super resolution using 3D deep densely connected neural networks, in: 2018 IEEE 15th International Symposium on Biomedical Imaging, IEEE, 2018, pp. 739–742.

[32] 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.

[33] J. Du, L. Wang, Y. Liu, Z. Zhou, Z. He, Y. Jia, Brain MRI super-resolution using 3D dilated convolutional encoder–decoder network, IEEE Access 8 (2020) 18938–18950.

[34] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, Y. Fu, Residual dense network for image super-resolution, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 2472–2481.

[35] J. Song, H. Liu, K. Deng, C. Zhang, Super resolution reconstruction of medical image based on adaptive quad-tree decomposition, J. Comput. Methods Sci. Eng. 17 (3) (2017) 411–422.

[36] S. Ren, D.K. Jain, K. Guo, T. Xu, T. Chi, Towards efficient medical lesion image super-resolution based on deep residual networks, Signal Process., Image Commun. 75 (2019) 1–10.

[37] X. Yang, S. Zhant, C. Hu, Z. Liang, D. Xie, Super-resolution of medical image using representation learning, in: 2016 8th International Conference on Wireless Communications & Signal Processing, IEEE, 2016, pp. 1–6.

[38] Y. Gao, H. Li, J. Dong, G. Feng, A deep convolutional network for medical image super-resolution, in: 2017 Chinese Automation Congress, IEEE, 2017, pp. 5310–5315.

[39] S. Zhang, G. Liang, S. Pan, L. Zheng, A fast medical image super resolution method based on deep learning network, IEEE Access 7 (2018) 12319–12327.

[40] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, Y. Bengio, Show, attend and tell: Neural image caption generation with visual attention, in: International Conference on Machine Learning, 2015, pp. 2048–2057.

[41] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, T.-S. Chua, Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 5659–5667.

[42] Z. Yang, X. He, J. Gao, L. Deng, A. Smola, Stacked attention networks for image question answering, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 21–29.

[43] H. Xu, K. Saenko, Ask, attend and answer: Exploring question-guided spatial attention for visual question answering, in: European Conference on Computer Vision, Springer, 2016, pp. 451–466.

[44] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7132–7141.

[45] 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. 286–301.

[46] F. Li, H. Bai, Y. Zhao, Detail-preserving image super-resolution via recursively dilated residual network, Neurocomputing 358 (2019) 285–293.

[47] J. Lu, W. Liu, Unsupervised super-resolution framework for medical ultrasound images using dilated convolutional neural networks, in: 2018 IEEE 3rd International Conference on Image, Vision and Computing, IEEE, 2018, pp. 739–744.

[48] Y. Hu, J. Li, Y. Huang, X. Gao, Channel-wise and spatial feature modulation network for single image super-resolution, IEEE Trans. Circuits Syst. Video Technol. (2019).

[49] D.C. Van Essen, S.M. Smith, D.M. Barch, T.E. Behrens, E. Yacoub, K. Ugurbil, W.-M.H. Consortium, et al., The WU-Minn human connectome project: An overview, Neuroimage 80 (2013) 62–79.

[50] C. Zhao, A. Carass, B.E. Dewey, J.L. Prince, Self super-resolution for magnetic resonance images using deep networks, in: 2018 IEEE 15th International Symposium on Biomedical Imaging, IEEE, 2018, pp. 365–368.