OPEN ACCESS
International Journal of Neural Systems, Vol. 31, No. 10 (2021) 2150037 (20 pages)
© The Author(s)
DOI: 10.1142/S0129065721500374

World Scientific
www.worldscientific.com

Int. J. Neur. Syst. 2021.31. Downloaded from www.worldscientific.com
by IMPERIAL COLLEGE LONDON on 10/08/21. Re-use and distribution is strictly not permitted, except for Open Access articles.

Arbitrary Scale Super-Resolution for Medical Images

Jin Zhu*, Chuan Tan and Junwei Yang
Department of Computer Science and Technology
University of Cambridge, Cambridge, CB3 0FD, UK
**jin.zhu@cl.cam.ac.uk*

Guang Yang†,‡
Cardiovascular Research Centre
Royal Brompton Hospital, London, SW3 6NP, UK
National Heart and Lung Institute
Imperial College London, London, SW7 2AZ, UK
g.yang@imperial.ac.uk

Pietro Lio'‡
Department of Computer Science and Technology
University of Cambridge, Cambridge, CB3 0FD, UK

Received 11 June 2021
Accepted 12 June 2021
Published Online 24 July 2021

Single image super-resolution (SISR) aims to obtain a high-resolution output from one low-resolution image. Currently, deep learning-based SISR approaches have been widely discussed in medical image processing, because of their potential to achieve high-quality, high spatial resolution images without the cost of additional scans. However, most existing methods are designed for scale-specific SR tasks and are unable to generalize over magnification scales. In this paper, we propose an approach for medical image arbitrary-scale super-resolution (MIASSR), in which we couple meta-learning with generative adversarial networks (GANs) to super-resolve medical images at any scale of magnification in $(1, 4]$. Compared to state-of-the-art SISR algorithms on single-modal magnetic resonance (MR) brain images (OASIS-brains) and multi-modal MR brain images (BraTS), MIASSR achieves comparable fidelity performance and the best perceptual quality with the smallest model size. We also employ transfer learning to enable MIASSR to tackle SR tasks of new medical modalities, such as cardiac MR images (ACDC) and chest computed tomography images (COVID-CT). The source code of our work is also public. Thus, MIASSR has the potential to become a new foundational pre-/post-processing step in clinical image analysis tasks such as reconstruction, image quality enhancement, and segmentation.

Keywords: Super-resolution; medical image analysis; image processing; generative adversarial networks; meta learning; transfer learning.

*,†Corresponding authors.
‡Guang Yang and Pietro Lio' are senior and last co-authors contributed equally for this study.

This is an Open Access article published by World Scientific Publishing Company. It is distributed under the terms of the Creative Commons Attribution 4.0 (CC BY) License which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

Int. J. Neur. Syst. 2021.31. Downloaded from www.worldscientific.com
by IMPERIAL COLLEGE LONDON on 10/08/21. Re-use and distribution is strictly not permitted, except for Open Access articles.

1. Introduction

Medical images of high quality and resolution are important in the current clinical process. For example, features and biomarkers extracted from functional and structural magnetic resonance (MR) images play a key role in the diagnosis and study of Alzheimer's disease,[1] stroke,[2] autism,[3] and Parkinson's disease.[4] However, the spatial resolution of medical images is limited by constraints such as acquisition time and equipment. Thus, super-resolution (SR) methods, which aim to increase the spatial resolution of digital images as post-processing, have the potential to improve the quality of medical images without additional scanning costs.[5] SR methods consider low-resolution (LR) images as the degradation of high-resolution (HR) images and magnify them by approximating a super-resolving transformation[6–8] to generate new pixels based on LR images and prior knowledge. Depending on the number of input and output images, SR methods can be briefly classified into single image SR (SISR) and multi-image SR. In this paper, we focus on SISR, which recovers one HR image from its LR version. In recent years, deep learning techniques have achieved great success in image processing problems such as human gait recognition[9] and image anomaly detection.[10] Particularly, with the rapid development of convolutional neural networks (CNNs) and generative adversarial networks (GANs), deep-learning-based SR methods have achieved remarkable performance[11–14] on various medical image modalities. However, most of these works are designed for specific magnification scales and treat SR with different scales as independent tasks. Thus, several models have to be trained and stored for different magnification tasks. Furthermore, it is challenging to collect large clinical datasets of HR and LR image pairs to train these SR methods for new applications. The high cost of training and implementation lead to poor clinical applicability and limit their application as a result. In this work, we first seek to apply meta-learning to GANs to tackle scale-free SR in medical images. We implement an end-to-end medical image SR network, which takes one LR image as input and generates corresponding SR images of an arbitrary scale of magnification. We train this model with pixel-wise error, perceptual loss and GAN-based adversarial loss, to improve both the perceptual quality and fidelity of generated images. Moreover, we overcome the cost of modifying well-trained models to new medical modalities using transfer learning. In particular, we focus on SR tasks with arbitrary scales in $(1, 4]$. This range is chosen as it is most commonly used in practice, but larger scales can be tackled using proper training data. The rest of the paper is structured as follows: first, we review the state-of-the-art (SOTA) SISR methods for natural and medical images; second, we describe the proposed method in detail; next, we introduce the design of our experiments, including data, evaluation metrics, and implementation details; and finally we illustrate the results of MIASSR on four medical image datasets and discuss the influence of each component of MIASSR.

2. Related Work

2.1. *Single image super-resolution*

Traditionally, SISR methods increase the resolution of the input LR image using a variety of interpolation,[15] reconstruction,[16] neighbor embedding,[17] and sparse coding.[18] However, these methods are incapable of simulating the nonlinear transformation from LR image to HR space and generate super-resolved images of poor quality. Recent developments in deep learning have led to dramatic improvements. Neural networks such as CNNs make it possible to learn this nonlinear transformation from corresponded LR–HR image pairs. Dong *et al.* first pioneered SRCNN to solve SR problems[19] with natural images, and achieved superior performance than traditional methods. Residual blocks[20] were then used to implement a much deeper network VDSR[21] and further improved the performance. These pre-upsampling methods first up-scaled the LR image with interpolation methods and then refined the primary results. However, processing feature maps in a high-dimensional space has a high computational and memory cost, thus limiting the depth of networks. To avoid this inefficient calculation, post-upsampling architectures moved to process LR feature maps and reconstruct SR images by up-sampling these LR features with transposed convolutions[22] and sub pixel layers.[23] Meanwhile, various skip-connection strategies were introduced to stabilize the training of these deep networks, including residual connections,[21] dense

Int. J. Neur. Syst. 2021.31. Downloaded from www.worldscientific.com
by IMPERIAL COLLEGE LONDON on 10/08/21. Re-use and distribution is strictly not permitted, except for Open Access articles.

connections,[24] and residual dense connections.[25] Despite the high performance on the peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) of these deep networks, their outputs were not photo-realistic. Thus, SRGAN[26] was proposed to overcome this gap and generate more photo-realistic images. Training with VGG-based[27] perceptual loss[28] and GANs,[29] SRGAN successfully generated perceptually realistic SR images which achieved comparable mean-opinion-score (MOS) with ground truth (GT) HR images. Following the same idea, ESRGAN[30] achieved superior performance by training a much deeper network with material recognition model-based perceptual loss and relativistic discrimination.[31] Additionally, methods with channel attention,[32] unsupervised learning,[33] and back projection[34] were also proposed. Most of these methods were designed for specific magnification SR scales and are incapable of varying the up-sampling scale. Thus, MetaSR[35] was proposed for scale-free SR tasks. It introduced a meta-upscale module that predicted a weight matrix and replaced the up-sample convolution layers (e.g. sub-pixel) with matrix multiplication. Instead of learning an up-sample transformation on a specific magnification scale, this new upscale module learns the relationship between up-sample transformations and scales of magnification, thus allowing a single model to super resolve images with arbitrary scales.

2.2. *Medical image super-resolution*

The increasing interest in, and development of, CNN-based SR algorithms has drastically improved their performance on medical image SR tasks.[5] In contrast to natural image SISR tasks, medical image SR is often pipelined by applications such as segmentation, classification and diagnosis, so it is required to preserve sensitive information and to enhance the structures of interest. Tackling the domain-specific SR problems, methods for medical image SR solved practical problems with techniques such as U-Net architecture,[14,36] cycle GAN,[13,37] and 3D convolution.[38] Moreover, medical image SR has embedded benefits for other medical image analysis tasks such as lesion detection,[11,12] and segmentation.[39] However, most of the current methods have poor clinical applicability due to two limitations. First, they were mainly proposed for specific SR scales, which are not able to meet the range of magnification requirements in medical image analysis tasks and diagnosis. Second, modifying the methods designed for one medical image modality to new modalities and tasks is time-consuming and unreliable. Thus, in this paper, we propose a medical image arbitrary scale super-resolution (MIASSR) method, in which we seek to apply meta-learning to GANs to tackle arbitrary scale SR tasks. Our main contributions are as follows:

- We first introduce meta-learning to medical image SR tasks and propose the first scale-free SR method for medical images with SOTA performance. Notably, in this paper, we focus on SR tasks of up to $\times 4$ magnification to meet the most common requirements of medical image SR in practice, but the proposed method can also work for a larger range of SR scales if proper training data is provided.
- We show that, compared to an existing meta-learning-based SR method,[35] MIASSR has fewer parameters (only 26%), but achieves comparable fidelity and perceptual quality, due to its use of adversarial learning and perceptual loss in the training process.
- We successfully extend the proposed method to various medical scans of different purposes, such as single- and multi-modal MR images of the brain, cardiac MR images, and CT scans of COVID patients. With transfer learning, the pre-trained model can be applied to new modalities efficiently.

3. Methods

SISR aims to restore a HR image I_{hr} from one LR observation I_{lr} of the same object. Generally, in the real world, the LR image is modeled as follows[7]:

$$I_{\mathrm{lr}} = (I_{\mathrm{hr}} \otimes \kappa) \downarrow_s + n, \tag{1}$$

where $I_{\mathrm{hr}} \otimes \kappa$ denotes the convolution between a blur kernel κ and the HR image I_{hr}, $\downarrow_s$ is the down-sampling operation with scale s, and n represents the noise. In SISR, we want to inverse the degradation mapping of Eq. (1) to recover a super-resolved image I_{sr} from I_{lr}[8]:

$$I_{\mathrm{sr}} = G(I_{\mathrm{lr}}, s; \theta_G), \tag{2}$$

where G is a CNN-based SR image generator and θ_G denotes its trainable weights. In each step of training, errors between the approximation I_{sr} and the HR GT I_{hr} are measured by a well-designed loss function

Int. J. Neur. Syst. 2021.31. Downloaded from www.worldscientific.com
by IMPERIAL COLLEGE LONDON on 10/08/21. Re-use and distribution is strictly not permitted, except for Open Access articles.

$\mathcal{L}_{\mathrm{SR}}$, and applied to the whole network using back-propagation to calculate gradients and update the weights θ_G[8]:

$$\hat{\theta_G} = \arg\min_{\theta_G} \mathcal{L}_{\mathrm{SR}}(G(I_{\mathrm{lr}}), I_{\mathrm{hr}}). \tag{3}$$

The SR image generator in our approach consists of a feature extractor $\mathcal{F}$ which extracts the feature maps of the LR image and a meta-upscale module $\mathcal{M}$ which up-samples the feature maps with arbitrary scales:

$$I_{\mathrm{sr}} = G(I_{\mathrm{lr}}) = \mathcal{M}(\mathcal{F}(I_{\mathrm{lr}}), s). \tag{4}$$

Note that the input and output layers that normalize and convert the images to the feature domain, and vice versa in the generator, are ignored to simplify this equation. The rest of this section will introduce each component of the proposed MIASSR (Fig. 1) in detail: firstly the feature extraction, then the meta-upscale module, and finally the losses for training.

3.1. *Feature extraction*

We use an Enhanced Residual Block (ERB) $\mathcal{B}_{\mathrm{ERB}}$-based feature extractor, namely EDSR-lite in MIASSR. This residual block (Fig. 2) is first proposed in EDSR,[40] which consists of two convolution layers $\mathcal{C}$, a nonlinear activation layer σ, a residual connection, and no batch normalization layer:

$$\begin{aligned} F_{\mathrm{out}} &= \mathcal{B}_{\mathrm{ERB}}(F_{\mathrm{in}}) \\ &= F_{\mathrm{in}} + \mathcal{C}_1^w(\sigma(\mathcal{C}_2^w(F_{\mathrm{in}}; \phi_{\mathcal{C}_2})); \phi_{\mathcal{C}_1}) \times \alpha, \\ &\quad \phi_{\mathcal{C}_1}, \phi_{\mathcal{C}_2} \in \theta_G, \end{aligned} \tag{5}$$

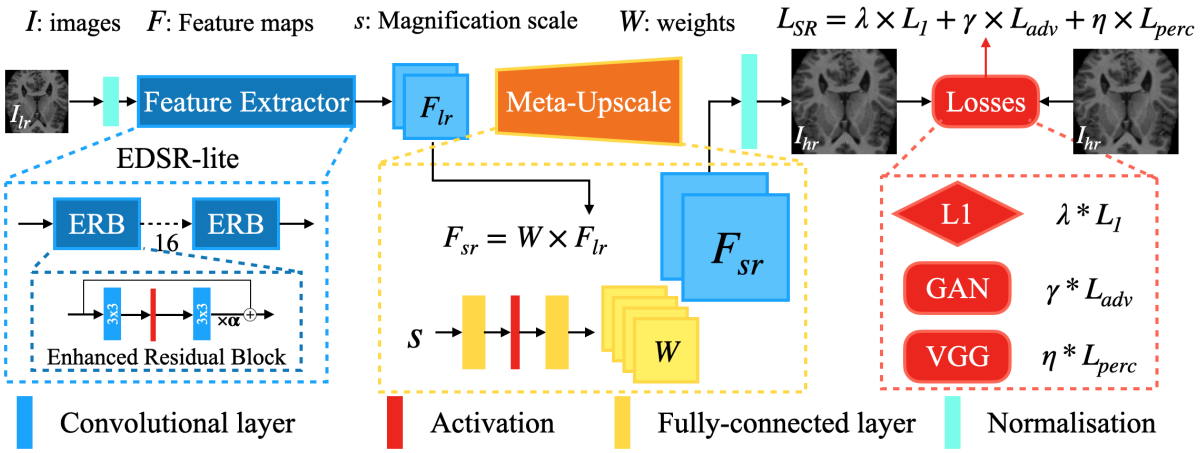


Fig. 1. The proposed MIASSR consists of an EDSR-lite-based low dimension feature extractor and a meta-upscale module. The feature extractor comprises 16 enhanced residual blocks, each of which includes two convolution layers and a nonlinear activation layer. It extracts low dimension feature maps F_{lr} of the input LR image I_{lr}. The meta-upscale module consists of two fully-connected layers and an activation layer. It predicts a group of weights from the input SR scale and achieves the feature map magnification with matrix multiplication. A super-resolved image I_{sr} is then generated from the enlarged feature maps F_{sr}. The whole model is trained end-to-end with a combined loss function including L1 loss, adversarial loss and VGG-based perceptual loss.

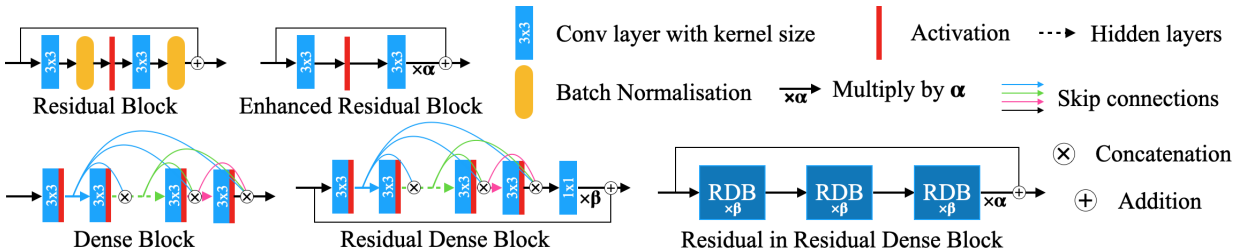


Fig. 2. Residual and dense blocks. These blocks consist of convolution layers and activation layers in common but have various skip connections. They are the basic unit of SOTA SISR networks: Residual Block is used in SRResNet and SRGAN;[26] Enhanced Residual Block is used in EDSR and MDSR;[40] Dense Block is used in SRDenseNet;[24] Residual Dense Block is proposed in RDN;[25] Residual in Residual Dense Block is proposed in ESRGAN.[30] Notice that residual scales α and β are introduced to stabilize the training of very deep neural networks (e.g. ESRGAN).

Int. J. Neur. Syst. 2021.31. Downloaded from www.worldscientific.com
by IMPERIAL COLLEGE LONDON on 10/08/21. Re-use and distribution is strictly not permitted, except for Open Access articles.

where F_{in} and F_{out} are the input and output feature maps, w presents the width of each convolutional layer, and $\{\phi_{\mathcal{C}_1}, \phi_{\mathcal{C}_2}\}$ are trainable parameters of the convolution layers. Padding is applied according to the kernel size in all convolution layers such that feature maps maintain their spatial dimensions. Thus, it becomes very convenient to build deeper neural networks by simply stacking several residual blocks, and low dimension feature maps F_{lr} can be extracted from the input LR image I_{lr}:

$$F_{\text{lr}} = \mathcal{F}(I_{\text{lr}}) = \mathcal{B}^d(I_{\text{lr}}; \phi_{\mathcal{B}}), \quad \phi_{\mathcal{B}} \in \theta_G. \tag{6}$$

In our approach, we use a lite version of EDSR, which consists of 16 enhanced residual blocks (i.e. Eq. (5)), and 64 convolution kernels with a size of 3×3 for each convolutional layer.

3.2. *Meta-upscale module*

To generate HR output from LR input, the low dimension feature maps F_{lr} extracted by $\mathcal{F}$ in Eq. (6) need to be upscaled:

$$F_{\text{sr}} = \mathcal{U}_s(F_{\text{lr}}; \phi_s), \tag{7}$$

where F_{sr} is the super-resolved feature maps, $\mathcal{U}_s$ is an up-sampler, and ϕ_s is the group of parameters for the magnification option with scale s. In comparison with common single scale up-samplers such as sub-pixel[23] which only learn one group of parameters for a specific SR scale s, the meta-upscale module[35] $\mathcal{M}$ in our approach learns to predict a group of parameters for each SR scale. Particularly, one pixel (i, j) of the super-resolved feature maps F_{sr} is calculated as a weighted sum of all pixels in F_{lr}:

$$F_{\text{sr}}(i, j) = \mathbf{v}_{i,j} \times F_{\text{lr}}, \tag{8}$$

where F_{lr} with original shape $(H_{\text{in}}, W_{\text{in}})$ is flattened to $(H_{\text{in}} \times W_{\text{in}}, 1)$ and $\mathbf{v}_{i,j}$ is a vector with dimensions $(1, H_{\text{in}} \times W_{\text{in}})$. The parameters of $\mathbf{v}_{i,j}$ are predicted by a weights prediction network $\mathcal{W}$ in $\mathcal{M}$ according to the scale s and the pixel's location (i, j):

$$\mathbf{v}_{i,j} = \mathcal{W}\left(\frac{i}{s} - \left\lfloor \frac{i}{s} \right\rfloor, \frac{j}{s} - \left\lfloor \frac{j}{s} \right\rfloor, \frac{1}{s}\right), \quad (i, j) \in F_{\text{sr}}. \tag{9}$$

Accordingly, all pixels in F_{sr} can be achieved via matrix multiplication:

$$F_{\text{sr}} = W_s \times F_{\text{lr}}, \tag{10}$$

where $W_s = \mathcal{W}(s)$ denotes the magnification matrix consists of $\mathbf{v}_{i,j}$ of all $(i, j) \in F_{\text{sr}}$ with scale s. Thus, the meta-upscale module is represented as

$$F_{\text{sr}} = \mathcal{M}(F_{\text{lr}}, s) = \mathcal{W}(s; \phi_{\mathcal{W}}) \times F_{\text{lr}}, \quad \phi_{\mathcal{W}} \in \theta_G, \tag{11}$$

The weights prediction network $\mathcal{W}$ consists of only three-layers, including two fully-connected layers and a nonlinear activation layer. This meta-upscale module works with any scales, which is quite different with sub-pixel-based up-samplers. Thus, it becomes possible to train an end-to-end model for SR tasks with arbitrary magnification scales.

3.3. *Loss functions*

We use a combined loss of pixel-wise L1 loss, adversarial loss and perceptual loss as in Eq. (3) to train our model.

3.3.1. *L1 loss*

SISR requires predicting the correct value of each pixel in the super-resolved images. Thus, pixel-wise errors are important for both evaluation and training of SR networks. In our work, we used the L1 loss, also called the mean-absolute-error (MAE) to train the model for good performance on PSNR and SSIM scores. It is defined as

$$\mathcal{L}_1(I_{\text{sr}}, I_{\text{hr}}) = \frac{1}{H * W} \sum_{(i,j) \in I} \|I_{\text{hr}}[i, j] - I_{\text{sr}}[i, j]\|, \tag{12}$$

where H and W are the height and width of the images. L1 Loss is a typical loss function used to train SISR networks. However, as this loss leads to over-smoothing, it has limited ability in generating perceptually realistic textures in medical images, which are important for human beings for visual understanding. Therefore, the perceptual loss and adversarial loss are also introduced to the training of our method.

3.3.2. *VGG-based perceptual loss*

The VGG-based perceptual loss $\mathcal{L}_{\text{perc}}$ was first introduced by Johnson *et al.*,[28] and had been widely used in SR tasks.[11,12,26,30] It presents the mean-square-error (MSE) between the super-resolved images and the HR GT images in the feature domain:

$$\mathcal{L}_{\text{perc}}(I_{\text{sr}}, I_{\text{hr}}) = \mathbb{E}(\|\mathcal{V}_l(I_{\text{hr}}) - \mathcal{V}_l(I_{\text{sr}})\|^2), \tag{13}$$

Int. J. Neur. Syst. 2021.31. Downloaded from www.worldscientific.com
by IMPERIAL COLLEGE LONDON on 10/08/21. Re-use and distribution is strictly not permitted, except for Open Access articles.

where $\mathcal{V}$ is a pre-trained VGG19 model and l denotes the feature maps of the specific layer of $\mathcal{V}$. Following the conclusion in ESRGAN,[30] we use the feature maps before activation of earlier layers to provide more textural information.

3.3.3. *Adversarial loss*

To generate more perceptually realistic images, we apply Wasserstein GAN-based adversarial training in our method. In addition to the SR image generator G, a GAN consists of a discriminator D, which assists the training of G by playing a game: the discriminator aims to distinguish generated fake images from real GT images, while the generator aims to fool the discriminator. The networks are trained jointly, so the discriminator should be able to recognize any image as real or fake correctly, while the generator should be able to produce as real as possible SR images. Thus, the basic adversarial loss function is defined as follows[29]:

$$\mathcal{L}_{\text{GAN}} = -\mathbb{E}_{I_{\text{hr}}}[\log D(I_{\text{hr}})] - \mathbb{E}_{I_{\text{lr}}}[\log(1 - D(G(I_{\text{lr}})))]. \quad (14)$$

Whilst this basic version of adversarial loss, so-called vanilla GAN loss, has been successfully used in natural image[26] and medical image[11] SR tasks, it is susceptible to problems of training instability and mode collapse. Thus, Wasserstein GAN was proposed by Arjovsky *et al.*[41] to resolve these issues. Instead of the binary classification loss in vanilla GAN, the Wasserstein distance between the distribution of real and generated images was introduced as the adversarial loss:

$$\mathcal{L}_{\text{WGAN}} = \mathbb{E}_{I_{\text{lr}}}[D(G(I_{\text{lr}}))] - \mathbb{E}_{I_{\text{hr}}}[D(I_{\text{hr}})]. \quad (15)$$

One important trick of Wasserstein GAN is to clip all the weights of the discriminator to a constant range $[-c, c]$, to meet the condition of derivable Wasserstein distance. However, with the clipping strategy, weights of discriminator tend to be either the minimum or maximum values. This causes the discriminator to behave like a binary network, and depresses the nonlinear simulating abilities of GANs. Thus, the gradient penalty was proposed to replace the clipping operation.[42] The new trick restricts the gradients of D to not change rapidly, by adding a new term in the adversarial loss:

$$\mathcal{L}_{\text{adv}} = \mathcal{L}_{\text{WGAN}} + \mathbb{E}_I[\| \bigtriangledown_I D(I)\|_p - 1]^2, \quad (16)$$

where $\| \|_p$ is the p-norm.

In summary, the loss that we use to solve equation (3) is defined as

$$\mathcal{L}_{\text{SR}} = \lambda \times \mathcal{L}_1 + \gamma \times \mathcal{L}_{\text{adv}} + \eta \times \mathcal{L}_{\text{perc}}, \quad (17)$$

where λ, γ, and η are scale factors to balance each part of the loss function.

4. Experiments

In our simulation experiments, the proposed method has been successfully applied to four different medical image datasets (Sec. 4.1) in SR tasks with arbitrary scales of magnification in $(1, 4]$. HR GT images and corresponding LR images were generated from the original slices (Sec. 4.1). To evaluate our method, metrics including PSNR, SSIM and Frechet Inception Distance (FID)[43] (Sec. 4.2) were used to measure the differences between the super-resolved images and GT images in the test set. Afterwards, the mean performance over all scales in $(1, 4]$ was compared with bicubic interpolation and seven SOTA SISR methods (Sec. 4.4).

4.1. *Data and pre-processing*

LR–HR image pairs for training, validation and testing were generated from the original slices. HR images were achieved by removing the pure-black background margin of original slices, whilst LR images were generated by down-sampling the corresponding HR images and blurred with a 3×3 Gaussian kernel. We focused on the central regions of each slice as the pure-black background regions has little information content and including this area only slows the training process. In the experiments, a suitable margin size for each dataset was carefully chosen to ensure no nonzero values were removed.

4.1.1. *OASIS-brains*

The OASIS-brains[44] dataset consists of a cross-sectional collection of 416 subjects including individuals with early-stage Alzheimer's Disease (AD). For each subject, 3 or 4 individual T1-weighted MRI scans obtained within a single imaging session are included (Fig. 3). For the single modality SR

Int. J. Neur. Syst. 2021.31. Downloaded from www.worldscientific.com
by IMPERIAL COLLEGE LONDON on 10/08/21. Re-use and distribution is strictly not permitted, except for Open Access articles.

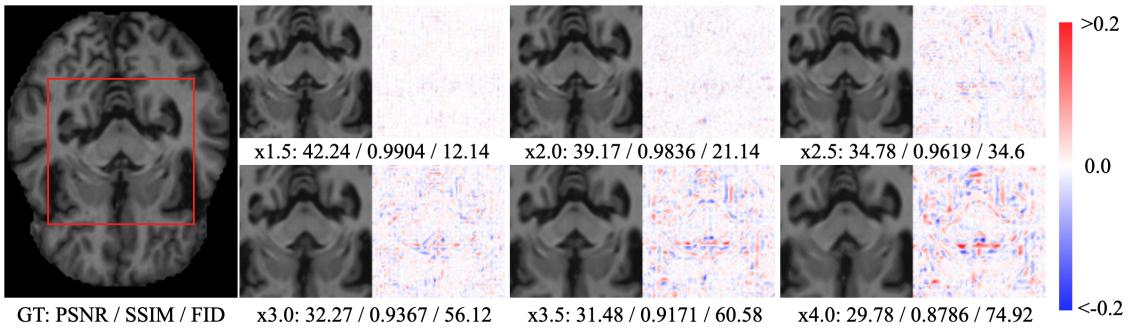


Fig. 3. An example of super-resolved images with different scales of magnification by our proposed MIASSR. The slice is randomly selected from the test set of OASIS-brains. Here, we only illustrate SR images with six scales $(1.5, 2.0, 2.5, 3.0, 3.5, 4.0)$, and the method could generate SR images with arbitrary scales in $(1, 4]$. All images are converted to $[0, 1]$. Differences between SR images and GT images are rendered with the color-bar and measured by PSNR, SSIM and FID. Higher PSNR and SSIM indicate better fidelity quality, while lower FID represents better perceptual quality.

experiments, the brain-masked version of an atlas-registered gain filed-corrected image, namely T88-111, was used. Due to the limitations of computing resources, from the whole dataset, we randomly select 30 subjects for training, 3 subjects for validation and another 9 subjects for testing. Note that the validation dataset was only used for hyper-parameters searching. The original size of each subject was $[176 \times 208 \times 176]$, and only a central area of $[144 \times 180]$ was used. Our experiments were applied on the axial plane.

4.1.2. *BraTS*

The brain tumor segmentation dataset (BraTS)[45–47] provides multi-modal MRI scans of 210 patients with glioblastoma (GBM/HGG) and the other 75 patients with lower grade glioma (LGG). Each BraTS multi-modal scan includes 4 MR modalities: native (T1), contrasted enhanced T1-weighted (T1ce), T2-weighted (T2), and T2 Fluid Attenuated Inversion Recovery (T2-FLAIR) volumes. We randomly selected 50 scans (35 HGG and 15 LGG) for training, and 10 scans (7 HGG and 3 LGG) for testing. The original image shape was $[240 \times 240 \times 155]$. Slices on axial plane were cropped to $[180 \times 170]$ to focus the training on brain area.

4.1.3. *ACDC*

The Automated Cardiac Diagnosis Challenge (ACDC)[48] dataset includes 1.5T and 3.0T MR scans of 150 subjects, including 30 health people and 120 patients with previous myocardial infarction, dilated cardiomyopathy, hypertrophic cardiomyopathy and abnormal right ventricle. The whole dataset was randomly assigned to 80 cases for training and 19 cases for testing. We applied MIASSR on the transverse plane, where the slices had various shapes from $[174 \times 208]$ to $[184 \times 288]$. To standardize the slice shapes, only the center areas with a size of $[128 \times 128]$ were cropped and used for training and testing.

4.1.4. *COVID-CT*

The COVID-CT dataset[49] includes CT scans of 632 patients with COVID infections, from which images of 199 patients were used for training, and images of another 25 patients were used for testing. The original image shape is $[512 \times 512]$. After removing the background, only the center area of $[412, 332]$ was used for either training or testing.

4.2. *Metrics*

Three objective image quality assessment methods were used to measure both the fidelity and perceptual quality of the generated SR images in our experiments. First, we use the PSNR, which is defined as follows[8]:

$$\mathrm{PSNR}(I_{\mathrm{sr}}, I_{\mathrm{hr}}) = 10 \cdot \log_{10} \times \left(\frac{L^2}{\frac{1}{N}\sum_{i=1}^{N}(I_{\mathrm{sr}}(i) - I_{\mathrm{hr}}(i))^2} \right), \tag{18}$$

where L denotes the maximum pixel ($L = 1.0$ in our case), and N is the number of all pixels in I_{sr} and I_{hr}.

Int. J. Neur. Syst. 2021.31. Downloaded from www.worldscientific.com
by IMPERIAL COLLEGE LONDON on 10/08/21. Re-use and distribution is strictly not permitted, except for Open Access articles.

PSNR relates to the pixel-level mean squared error, and is the most widely used evaluation criteria for SR models.[5,7,8] Meanwhile, another popular metric, structural similarity (SSIM), is also used. It is defined as follows[7]:

$$\mathrm{SSIM}(x, y) = \frac{2\mu_x\mu_y + \kappa_1}{\mu_x^2 + \mu_y^2 + \kappa_1} \cdot \frac{\sigma_{xy} + \kappa_2}{\sigma_x^2 + \sigma_y^2 + \kappa_2}, \quad (19)$$

where x, y denote two images, 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. SSIM assumes that the human visual system is highly adapted to extract image structures, and measures the structural similarity between images based on the comparisons of luminance, contrast, and structures.

However, both metrics are limited to measure the fidelity quality, but cannot evaluate perceptual quality. Over-smoothed images were reported[12,26] to achieve higher PSNR and SSIM scores than texture-rich images, but they might be less perceptually realistic. Thus, we also calculated the Frechet Inception Distance (FID),[43] which is widely used to evaluate the perceptual quality of generated images by GANs. It measures the difference of high-level global features of generated SR images and GT images, by calculating the distance between the distributions of both groups of images in the latent space of a pre-trained image classification model Inception-V3.[50] Note that higher scores of PSNR and SSIM represent better fidelity quality, while lower FID indicates more perceptually realistic images have been generated.

4.3. *Implementation details*

We use PyTorch to implement our method, NiBabel to load medical data, and OpenCV-python for image resize and blur operations. All experiments were performed on an Nvidia Quadro RTX 8000 GPU. Following the instructions of our open-source implementation,[a] researchers could reproduce our experiments and test new experiments easily. The following hyper-parameters and details of training tricks were also released as config files.

The generator consists of $d = 16$ enhanced residual blocks, in which each convolution layer has $w = 64$ feature maps. A residual scale $\alpha = 1.0$ was used for the residual connections in ERB. The discriminator is similar as in DCGAN[52] and SRGAN.[26] It consists of 4 down-sample blocks, each of which has one convolution layer with stride = 1 for feature expanding and one convolution layer with stride = 2 for down-sampling. No batch normalization layers were used in either the generator or the discriminator, while leaky-ReLU with negative-slope = 0.2 was chosen as the nonlinear activation function. In the experiments without transfer learning, both networks were initialized with Kaiming-uniform.[53]

During training, LR and HR images were randomly cropped into sample patches. The original path size was set to $H_p, W_p = (96, 96)$ for HR patches, but either the size of LR patches or the size of the HR patches were adjusted to match the magnification scale s:

$$\begin{aligned} H_{\mathrm{lr}}, W_{\mathrm{lr}} &= \left\lfloor \frac{H_p}{s} \right\rfloor, \left\lfloor \frac{W_p}{s} \right\rfloor; \\ H_{\mathrm{hr}}, W_{\mathrm{hr}} &= \lfloor sH_{\mathrm{lr}} \rfloor, \lfloor sW_{\mathrm{lr}} \rfloor. \end{aligned} \quad (20)$$

For each training step, a batch of 16 random patches, with the same SR scale was fed to the model. Initially, in pre-training, the generator of MIASSR was trained with only $\mathcal{L}_1$ for 1×10^5 steps, because this 'warm-up' has been found to stabilize the training of GANs.[30] Then we train both generator and discriminator with $\lambda = 1$, $\gamma = 0.001$ and $\eta = 0.006$ for 1×10^5 steps. Adam optimizer,[54] with an initial learning rate $l_r = 0.0001$, momentum $m = 0.9$, and betas $b = (0.9, 0.999)$ was used for backpropagation. Every 5×10^4 steps, the learning rate was halved. To avoid a gradient explosion, losses above 1×10^8 were discarded.

All the above hyper-parameters were chosen based on the validation performance in our experiments on the OASIS-brains dataset. In the transfer learning experiments of ACDC and COVID-CT datasets, the model which had been pre-trained on OASIS-brain was fine-tuned for 1×10^4 steps with $\mathcal{L}_{\mathcal{SR}}$. Meanwhile, to make MIASSR work with multi-modal scans in the experiments with the BraTS dataset, the single-channel input and output layers were modified to 4-channel. Particularly, the four modalities of BraTS, T1ce, T1, T2 and Flair, were stacked during both training and testing, and the loss function $\mathcal{L}_{\mathrm{SR}}$ was first calculated on each modality independently and then averaged.

[a]https://github.com/GinZhu/MIASSR.

Int. J. Neur. Syst. 2021.31. Downloaded from www.worldscientific.com
by IMPERIAL COLLEGE LONDON on 10/08/21. Re-use and distribution is strictly not permitted, except for Open Access articles.

4.4. *Comparison with SOTA methods*

We compared the proposed method with the bicubic interpolation and 7 SOTA SISR methods: SRGAN,[26] EDSR,[40] SRDenseNet,[24] RDN,[25] MDSR,[40] ESRGAN-L1,[30] and MetaRDN.[35] All these methods were designed for natural images, which have much bigger dimensions than the medical images we used. To make them work well with our experiments, we retrained the models with the medical image datasets with smaller patches. All models were trained with the same steps and learning rate decay policy, without model embedding and data augmentation for a fair comparison. We used the original loss functions to train most of the models, but the material recognition-based[55] perceptual loss in ESRGAN hindered training with medical images, so we used the L1-loss-based ESRGAN (so-called ESRGAN-L1) for our comparison.

All these SISR methods (except MetaRDN) were designed only for specific integer magnification scales ($\times 2$, $\times 3$, and $\times 4$). To evaluate their performance on SR tasks with arbitrary scales, we used an up-and-down strategy, consisting of two steps: in the up-sampling step, the well-trained model with the ceiling scale $\lceil s \rceil$ is used to infer an over-magnified SR image; then, in the down-sampling step, the over super-resolved image was resized correspondingly using the bicubic interpolation.

5. Results and Discussion

In this section, we first present the results of MIASSR with four datasets, and compare the performance with SOTA SR methods. Then, we systemically analyze the influence of each component of the model. The ablation study is important to understand the how the hyper-parameters and model architecture affect the final SR performance, as in other image processing tasks.[56]

5.1. *MIASSR performance*

First of all, in the experiments with the OASIS-brains dataset (Fig. 3), MIASSR was compared with SOTA methods with arbitrary SR scales. On average, it achieved the third-best mean PSNR and SSIM scores, and the best FID with the fewest parameters (Fig. 4, and Table 1). Particularly, with all SR scales, MIASSR generated images with comparable fidelity

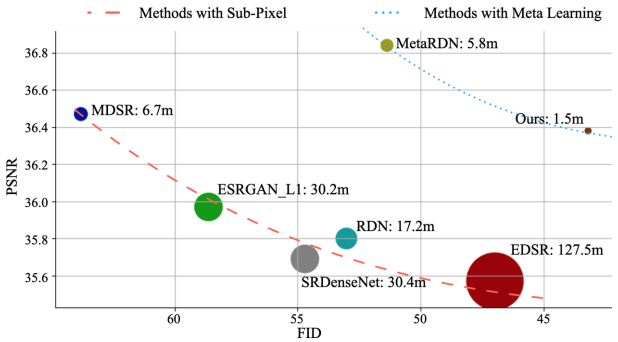


Fig. 4. Performance and model size of our method compared to SOTA methods. Mean PSNR and mean FID on SR scales between $(1, 4]$ are used to evaluate both the fidelity and perceptual quality of generated images. The bubble size denotes the number of parameters of each model. SRGAN is not plotted because it has much worse performance ($\text{meanPSNR} = 28.15, \text{meanFID} = 128.21$) compared to all other methods. Higher PSNR scores denote better fidelity quality, and lower FID represents better perceptual quality. Our proposed method has achieved the best FID and the third-best PSNR with the smallest model size. Perception-distortion trade-off[51] is reflected in both groups: methods with meta-learning, and methods with sub-pixel. High fidelity and perceptual quality are at odds, and impossible to be improved simultaneously.

Table 1. Comparing our method with SOTA methods in SR tasks with arbitrary scales in $(1, 4]$ on the OASIS-brains dataset. Higher PSNR and SSIM denote better fidelity quality, while lower FID means better perceptual quality. The best performance is in bold. The unit of the number of parameters of each model is million. Along with the smallest model size out of all CNN-based methods, our proposed method achieves the best FID and comparable PSNR/SSIM.

	BiCubic	SRGAN[26]	EDSR[40]	ESRGAN(L1)[30]	MDSR[40]	RDN[25]	SRDenseNet[24]	MetaRDN[35]	Ours
PSNR	31.32	28.79	35.71	36.11	36.63	35.95	35.83	**36.84**	36.46
SSIM	0.8600	0.6380	0.9541	0.9568	0.9595	0.9574	0.9548	**0.9627**	0.9576
FID	144.9	117.7	44.00	54.36	58.95	49.37	50.90	51.36	**39.85**
Params	—	4.5M	127.5M	30.2M	6.7M	17.2M	30.4M	5.8M	**1.5M**

Int. J. Neur. Syst. 2021.31. Downloaded from www.worldscientific.com
by IMPERIAL COLLEGE LONDON on 10/08/21. Re-use and distribution is strictly not permitted, except for Open Access articles.

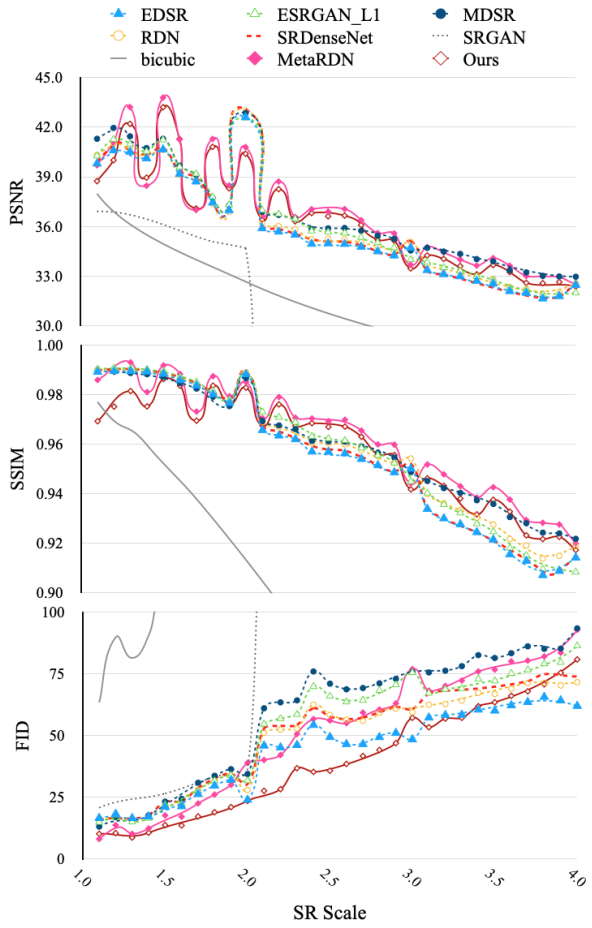


Fig. 5. Comparing our proposed method with bicubic interpolation and SOTA methods in SR tasks with arbitrary scales in $(1, 4]$. Higher PSNR and SSIM denote better fidelity quality, while lower FID represents better perceptual quality. Results of bicubic interpolation and SRGAN are not fully plotted because of their poor performance. Our proposed method has achieved comparable performance with SOTA methods with all SR scales.

and perceptual quality with SOTA methods (Fig. 5). Considering the balance between cost and performance, the mean PSNR score and FID were plotted with the number of parameters of each model. Particularly for SISR methods which only supported one SR scale (i.e. EDSR, RDN, SRDenseNet and ESRGAN), the model size denoted the number of all parameters in three models $\times 2$, $\times 3$, and $\times 4$, because they were all required in inference. Methods designed for multi-scale (i.e. MDSR) and arbitrary scales (i.e. MetaRDN and Ours) greatly decreased the effective model size by reducing the required models to one.

Notably, MIASSR had the fewest parameters which were only 26% of the existing meta-learning-based method (i.e. MetaRDN) and fewer than 1% of EDSR. Interestingly, multi-task methods (i.e. MDSR, MetaRDN, our MIASSR) achieved much better PSNR scores than single-scale SR methods, although they were learning more challenging transformations. Instead of approximating one mapping with a certain SR scale, their parameters were shared in approximations of mappings with various SR scales. However, the parameter sharing between tasks and the diversity of LR–HR image pairs with different magnification scales helped the models converged better in the training process. Interestingly, if we divided all the methods into two groups — methods with meta-learning and methods without meta-learning — the perception-distortion trade-off[51] was clearly observed in each group. We found that the stem of workflow determined the overall performance, while the details of architectures only affect the balance between fidelity and perceptual quality.

The proposed method has also performed well with various medical image modalities (Table 2). Transfer learning, which helps to decrease the training cost on medical data processing,[57] also helped to modify the OASIS-brains pre-trained model to new single-modality medical image datasets efficiently and effectively. Compared to the bicubic interpolation method, the proposed MIASSR significantly improves performance in experiments with ACDC and COVID-CT. Comparing with EDSR and MetaRDN, MIASSR generated images of comparable fidelity and better perceptual quality, with only one-fifth of the training steps. Indeed, transfer learning reduced the required training steps from 1×10^5 to 2×10^4. Besides, MIASSR is straightforwardly extended to multi-modality images: by simply modifying the input and output layers, it successfully worked for the cross-modality SR task of the BraTS dataset. Comparing with SOTA methods that work for a single modality, it achieved comparable performance on all four modalities (T1, T2, T1ce and Flair). The additional visualization results of ACDC, COVID-CT and BraTS are attached in the appendix.

In summary, the proposed method has good clinical applicability. In our experiments, it has been successfully applied to various medical image SR tasks, with different situations of modalities and

Int. J. Neur. Syst. 2021.31. Downloaded from www.worldscientific.com
by IMPERIAL COLLEGE LONDON on 10/08/21. Re-use and distribution is strictly not permitted, except for Open Access articles.

Table 2. Comparing MIASSR with EDSR, MetaRDN, and bicubic interpolation on multi-modal brain images (BraTS), cardiac MR scans (ACDC), and chest CT images of COVID patients (COVID-CT). Higher PSNR represents better fidelity quality of SR images, while lower FID denotes better perceptual quality. The best performance is in bold. Notice that MIASSR requires only one fifth training steps of EDSR and MetaRDN with transfer learning. Moreover, unlike EDSR and MetaRDN which only work with a single modality, it works with multi-modality data in a single model and further reduces the cost of extending to new SR tasks.

PSNR / FID	BraTS-T1	BraTS-T1ce	BraTS-T2	BraTS-Flair	ACDC	COVID-CT
BiCubic	32.83/151.6	33.13/139.7	29.87/125.1	31.52/145.8	27.42/267.7	38.62/141.0
EDSR[40]	36.54/76.42	36.34/74.18	33.69/51.93	35.27/79.32	30.82/179.3	**46.75**/39.53
MetaRDN[35]	**37.27**/78.16	**37.22**/72.93	**34.76**/53.55	**36.05**/78.88	**31.27**/154.7	43.27/38.96
MIASSR(ours)	36.89/**62.28**	36.86/**61.06**	34.28/**46.74**	35.91/**68.45**	30.94/**153.5**	43.21/**37.66**

diseases. It works with the brain and cardiac MR images (i.e. OASIS-brains, ACDC), chest CT images (i.e. COVID-CT), and cross-modality scans (BraTS). Comparing with SOTA methods, MIASSR can generate SR images with comparable fidelity and better perceptual quality with much smaller model size. Additionally, with transfer learning, MIASSR can be extended to new datasets effectively and efficiently.

5.2. *Generator architectures*

Referring to Eqs. (5) and (6), the SR image generator is influenced by three factors: the block structure $\mathcal{B}$, the number of blocks (namely the depth d), and the number of convolution kernels of each layer (namely the width w). Broadly, the width and depth of the network determine the size of the network, while the block structure represents the connection of these layers. They resolve the capability of feature extraction jointly. To understand how each factor influences the final performance, we have designed an ablation study of generator architectures.

First, we have compared a wide range of architectures for LR feature extraction in MIASSR. Six SOTA networks, which have been widely used in computer vision and achieved high performance on SISR tasks, were tested. These networks, SRResNet, EDSR, MDSR, SRDenseNet, RDN and ESRGAN, were built with various residual and dense blocks (Fig. 2) and behaved differently on simulating the LR-to-HR transformation. To match MIASSR, we replaced their scale specific up-samplers with the meta-upscale module, as in MetaSR,[35] and kept the original settings (i.e. depth and width). To compare the performance of these generators, we trained them

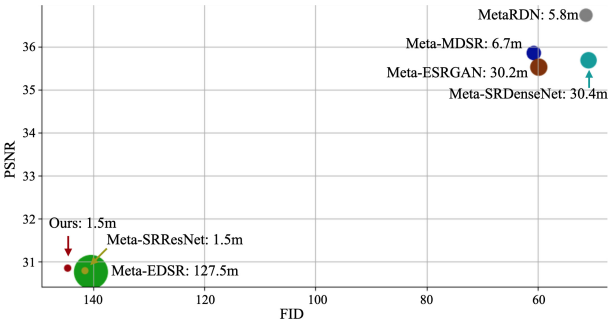
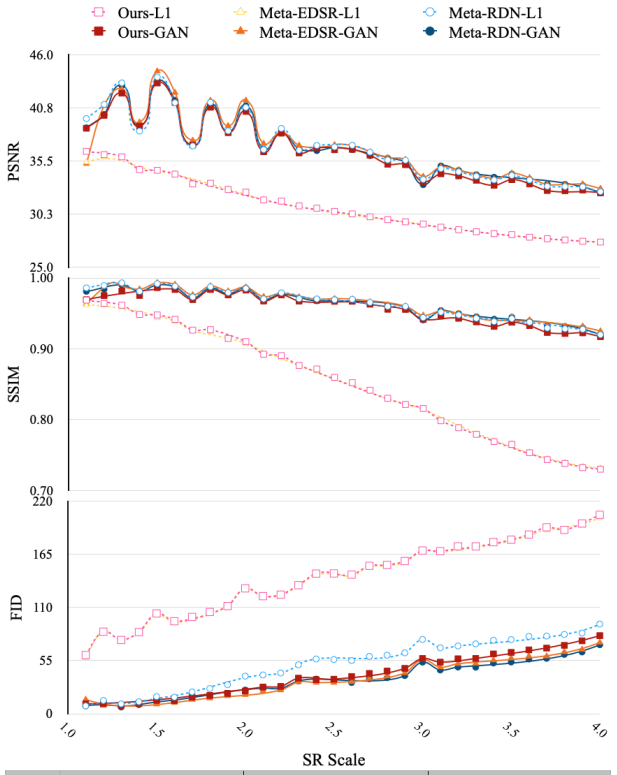


Fig. 6. Comparison of SR image generators in MIASSR (training with $\mathcal{L}_1$). Higher PSNR indicates better fidelity quality, while lower FID represents more perceptually realistic SR images. Bubble size denotes the number of parameters. The basic block design, such as skip connections, rarely affects the final performance, but the depth of networks impact the performance a lot. Deeper networks (MetaRDN, Meta-MDSR, Meta-ESRGAN, and Meta-SRDenseNet) perform better.

with the L1 loss $\mathcal{L}_1$ for 1×10^5 steps with the same training settings. Both fidelity and visual qualities of the generated SR images were evaluated (Fig. 6).

When training with only fidelity loss $\mathcal{L}_1$, we found it very interesting that these variations could be divided into two groups. Methods with more layers always performed better, because the deeper structure provided stronger ability in simulating nonlinear transformations. Other structures, such as skip connections in dense blocks and the width of models, made no big differences. However, suitable structures for minimizing pixel-wise errors might not fit the needs of generating perceptually realistic textures. Thus, we further compared RDN, EDSR and EDSR-lite with adversarial learning (Fig. 7).

Int. J. Neur. Syst. 2021.31. Downloaded from www.worldscientific.com
by IMPERIAL COLLEGE LONDON on 10/08/21. Re-use and distribution is strictly not permitted, except for Open Access articles.



	Ours (EDSR-lite)			Meta-EDSR			Meta-RDN		
	L1	+ GAN	rasie	L1	+ GAN	raise	L1	+ GAN	raise
PSNR	30.98	36.46	5.48	30.91	36.92	6.01	36.84	36.76	-0.08
SSIM	0.8520	0.9576	0.1056	0.8509	0.9638	0.1129	0.9627	0.9623	-0.0004
FID	141.5	39.85	101.7	140.4	35.80	104.60	51.36	35.22	16.14

Fig. 7. The sensitivity analysis of SR image generators to GAN-based adversarial learning. Three networks, EDSR-lite (ours), EDSR and RDN are trained with $\mathcal{L}_1$ only, and with additional $\mathcal{L}_{\text{adv}}$ and $\mathcal{L}_{\text{perc}}$, and compared. Low FID represents good perceptual quality of generated images, while high PSNR and SSIM indicate good fidelity quality. Adversarial learning has significantly improved the performance of EDSR-based methods.

Although MetaRDN achieved the best performance with L1 loss, it was not improved with further adversarial learning. Oppositely, the other two were distinctly benefited by the adversarial and perceptual loss. However, comparing the original EDSR with the EDSR-lite, extra feature maps of each convolution layer only slightly improved the performance, although nearly one hundred times more parameters were used.

Additionally, we investigated the influence of the width and depth of the network over the final SR performance. Technically, deeper and wider networks should have more capability of approximation, which would result in better SR performance. However, more trainable parameters also led to more challenging optimization of Eq. (3) and over-fitting. Meanwhile, the cost and performance balance should also be considered, because the number of parameters grows linearly with the depth and quadratically with the width. In our experiments of the width of the network, the model that consisted of 64 convolution kernels in each layer achieved the best PSNR and FID (Table 3). Wider networks did not improve the performance (e.g. $w = 128$), and the optimization in networks training even failed to converge when $w = 256$. Similarly in the experiments of the depth of the network, additional residual blocks over 16 rarely helped (Table 4).

Network architectures, such as skip connections, depth and width have significant impacts on the final performance of medical image SR tasks. First of all, extra skip connections are a double-edged sword. These connections in dense blocks such as RDB and RRDB add more pathways in the networks. With these highways of loss information, gradients could be passed more efficiently and effectively to each layer during backpropagation. Thus, the model (e.g. RDN) could easily achieve a very good performance, especially for minimizing simple and clear errors such as L1 loss. However, the structure of the dense connections also made the model liable to getting stuck in certain points, and made the model insensitive to uncertain losses such as GANs. As a result, smaller models with fewer connections, such as EDSR-lite, could be comparable with them. Second, wider models are not necessary for medical images. Both RDN and original EDSR had many more feature maps than EDSR-lite in each convolutional layer, which made them more powerful in simulating and extracting features of nature images in SR tasks. However, for medical images with limited size and relatively lower contrast information, too many feature maps tended to be overqualified. In summary, we decided to use EDSR-lite, consisting of 16 residual blocks and 64 convolution kernels in each layer, because it had the fewest parameters and achieved equal performance with bigger models.

5.3. *Loss functions*

Referring to (3) and (17), the joint loss function $\mathcal{L}_{\text{SR}}$ plays an important role in the training of MIASSR.

Int. J. Neur. Syst. 2021.31. Downloaded from www.worldscientific.com
by IMPERIAL COLLEGE LONDON on 10/08/21. Re-use and distribution is strictly not permitted, except for Open Access articles.

Table 3. Effects of the width of the network w, which represents the number of convolution kernels in each layer. Increasing width leads to additional parameters quadratically. Higher PSNR, SSIM and lower FID denote better performance. Bold texts represent the best performance, and the grey column denotes the hyper-parameter of our method. We finally choose $w = 64$ because it has achieved the best PSNR and FID.

$w =$	4	8	16	32	64	128
PSNR	34.58	34.83	35.22	35.89	**36.46**	36.44
SSIM	0.9346	0.9399	0.9469	0.9542	0.9576	**0.9592**
FID	84.17	73.49	58.23	49.57	**39.85**	43.84
Params	**0.016M**	0.041M	0.12M	0.42M	1.5M	5.8M

Table 4. Effects of the depth of the network d, which represents the number of residual blocks in the network. The unit of the number of parameters is million. Higher PSNR, SSIM and lower FID denote better performance. Bold texts represent the best performance, and the grey column denotes the hyper-parameter of our method. We finally choose $d = 16$, because extra blocks rarely improve the performance, but lead to more parameters linearly.

$d =$	2	4	8	16	32	64
PSNR	35.09	35.48	36.03	36.46	**36.73**	36.46
SSIM	0.9438	0.9490	0.9549	0.9576	**0.9622**	0.9621
FID	62.47	54.44	49.61	39.85	37.46	**37.39**
Params	**0.48M**	0.63M	0.93M	1.5M	2.7M	5.1M

Table 5. Effects of perceptual loss. In this experiment, we set $\lambda = 1$ and $\gamma = 0.001$, and test different values of η. Particularly $\eta = 0$ means no perceptual loss, while $\eta = \infty$ means only perceptual loss is used for training. Higher PSNR, SSIM and lower FID denote better performance. Bold texts represent the best performance, and the grey column denotes the hyper-parameter of our method.

$\eta =$	0	0.006	0.01	0.1	1	∞
PSNR	36.52	36.46	**36.64**	36.34	36.11	34.94
SSIM	**0.9611**	0.9576	0.9607	0.9583	0.9554	0.9532
FID	47.87	**39.85**	40.61	43.59	45.79	46.61

Table 6. Effects of adversarial loss. In this experiment, we set $\lambda = 1$ and $\eta = 0.006$, and test different values of γ. Particularly, $\gamma = 0$ means no adversarial loss, while $\gamma = \infty$ means only adversarial loss is used for training. Higher PSNR, SSIM and lower FID denote better performance. Bold texts represent the best performance, and the grey column denotes the hyper-parameter of our method.

$\gamma =$	0	0.001	0.01	0.1	1	∞
PSNR	**36.60**	36.46	35.74	36.26	30.92	30.52
SSIM	**0.9592**	0.9576	0.9555	0.9540	0.8798	0.8768
FID	41.94	39.85	**37.47**	44.17	114.9	103.1

Int. J. Neur. Syst. 2021.31. Downloaded from www.worldscientific.com
by IMPERIAL COLLEGE LONDON on 10/08/21. Re-use and distribution is strictly not permitted, except for Open Access articles.

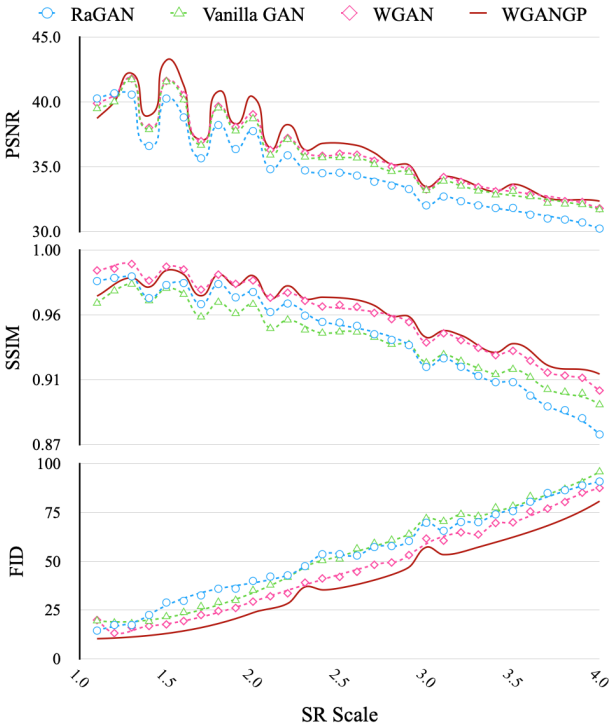
In the three components, $\mathcal{L}_1$ represents the pixel-wise errors, while $\mathcal{L}_{\text{perc}}$ and $\mathcal{L}_{\text{adv}}$ denote the visual dissimilarity of the entire images. Particularly, the perceptual loss $\mathcal{L}_{\text{perc}}$ considers the general visual features of images, because it is based on a well-trained VGG network with big plenty of normal images. In contrast, the adversarial loss $\mathcal{L}_{\text{adv}}$ is trained with the model, so it focuses on the inner features of the training dataset more. To achieve the best performance and to understand each component well, we have tested various weights of each loss. We first set $\lambda = 1$, then tested different values of γ and η. Regarding the perceptual loss, we have tested $\eta = (0, 0.006, 0.01, 0.1, 1)$ and infinity (Table 5). Particularly, $\eta = 0$ means no perceptual loss, while $\eta =$ infinity means only perceptual loss is used. Similarly, $\gamma = (0, 0.001, 0.01, 0.1, 1)$ and infinity were tested for the adversarial loss (Table6).

In our experiments, none of these values could lead to the best PSNR, SSIM and FID simultaneously, but when $\gamma = 0.001$ and $\eta = 0.006$, it performed well on both fidelity and perceptual evaluations.

Regarding the variations of adversarial loss, we also compared four popular GAN variations, which had been successfully used in SOTA SISR studies: vanilla GAN,[26,29] RaGAN,[30,31] WGAN,[12,41] and WGANGP.[12,42] They were trained with the same hyper-parameters from the same start point of a pre-trained generator, but with different designs of $\mathcal{L}_{\text{adv}}$. Vanilla GAN is using Eq. (14); WGAN is using Eq. (15); WGANGP is using Eq. (16); RaGAN is using the following equation:

$$\mathcal{L}_{\text{RaGAN}} = \mathbb{E}_{I_{\text{hr}}}[\log D(I_{\text{hr}}) - \mathbb{E}_{I_{\text{lr}}}[\log D(G(I_{\text{lr}}))]] + \mathbb{E}_{I_{\text{lr}}}[\log(D(G(I_{\text{lr}}))) - \mathbb{E}_{I_{\text{hr}}}[\log D(I_{\text{hr}})]]. \quad (21)$$

In our experiments (Fig. 8), WGANGP helped MIASSR achieve the best performance with all metrics. Therefore, we choose WGANGP in MIASSR.



	RaGAN	Vanilla GAN	WGAN	WGANGP
PSNR	34.72	35.82	36.06	**36.46**
SSIM	0.9422	0.9412	0.9564	**0.9576**
FID	53.34	53.27	46.15	**39.85**

Fig. 8. Comparing GAN variations in MIASSR. Higher PSNR and SSIM indicate better fidelity quality, while lower FID denotes more perceptual realistic images have been generated. WGANGP has not only achieved the best mean performance (the table at bottom) but also performed the best with almost all SR scales.

5.4. *Training tricks*

We also found tricks of data processing that can affect model training with significant impacts on the final performance. First, to transfer the well-trained model on the OASIS dataset to new datasets could accelerate the training process, although it could not improve the final performance. Second, the networks, including both the generator and discriminator, must be initialized by a variety of uniform function (e.g. Kaiming-Uniform[53]) when they are trained from scratch. In our experiments, initializing networks with a normal distribution (e.g. Kaiming-Normal[53]) crashed the training process. Third, batch normalization[58] should not be used, although it has succeeded in a wide range of image processing tasks. SRGAN, as the only method with batch normalization, performed poorly in the

Int. J. Neur. Syst. 2021.31. Downloaded from www.worldscientific.com
by IMPERIAL COLLEGE LONDON on 10/08/21. Re-use and distribution is strictly not permitted, except for Open Access articles.

patient-wise experiment (Table 1). This is because the normalization operation distorted the patient-wise contrast information during training, and led to poor performance on the test set. Finally, hyper-parameters such as patch size and image size can influence the final performance. In Fig. 5, although all methods tended to perform better with smaller magnification scales than larger ones in general, the best PSNR and SSIM scores always appeared with the x1.5 SR task. This is probably because 1.5 is the smallest SR scale to fit both the training patch size $[96 \times 96]$ and testing image shape $[144 \times 180]$.

5.5. *Future work*

MIASSR can be extended in several potential directions. First, we believe that MIASSR can be extended to a wider range of medical image modalities and data types. In this work we focus on 2D slice SR because it is more general. In the future, to modify MIASSR with techniques, 3D convolution and recurrent networks may extend it to 3D images,[59] temporal scans,[9] and other 3D data such as mesh[60] and point cloud.[61] Second, we have had a taste of cross modality with naive transfer learning (ACDC and COVID-CT) and multi-task training (BraTS), and it would be interesting to investigate MIASSR further with more specific cross-modality applications. Thirdly, the ablation study of SR image generators and the adversarial learning help us to understand each component of MIASSR well, and the conclusions and findings may also benefit other medical image analysis research studies, such as synthesis, reconstruction and segmentation. Finally, Fig. 4 has reflected the balance between perception and distortion in either the sub-pixel-based SR methods and meta-upscale module SR methods. Further research of this trade-off will be useful in both research and clinical application development. Furthermore, to find a more straightforward measurement of medical images in SR tasks is also desired. For example, instead of measuring the quality of generated images, to evaluate the images in specific tasks such as AD diagnosis[1] and Parkinson disease identification[62] could be a potential way.

6. Conclusions

In this paper, we have proposed MIASSR for medical image SR tasks with arbitrary scales. To our best knowledge, this is the first attempt to develop a meta-learning scheme for this problem. We have first reduced the model size (only 26% parameters as compared to existing meta-SR methods) by using a lite EDSR model as the LR image feature extractor and achieved comparable fidelity quality of SR images with SOTA methods. Moreover, we have introduced GANs to meta-learning to improve the perceptual quality of the generated images. The proposed method has obtained good practical applicability. In our experiments, we have successfully applied MIASSR with T1-weighted brain MR images and multi-modal brain MR scans. Furthermore, with transfer learning, the pre-trained model on brain images has been fine-tuned to cardiac MR scans and CT scans. We have also discussed our findings and understanding of model architecture design, training tricks, and adversarial learning in the comparison studies.

Acknowledgments

This work was supported in part by China Scholarship Council (Grant No. 201708060173), in part by the British Heart Foundation [TG/18/5/34111, PG/16/78/32402], in part by the European Research Council Innovative Medicines Initiative on Development of Therapeutics and Diagnostics Combatting Coronavirus Infections Award 'DRAGON: rapiD and secuRe AI imaging based diaGnosis, stratification, fOllow-up, and preparedness for coronavirus paNdemics' [H2020-JTIIMI2 101005122], in part by the AI for Health Imaging Award 'CHAIMELEON: Accelerating the Lab to Market Transition of AI Tools for Cancer Management' [H2020-SC1-FA-DTS-2019-1 952172], in part by the Hangzhou Economic and Technological Development Area Strategical Grant [Imperial Institute of Advanced Technology], in part by the UK Research and Innovation (MR/V023799/1), and in part by the GO-DS21 EU grant proposal. The authors would like to thank Ben Day for constructive criticism of the paper. Guang Yang and Pietro Lio are senior and last co-authors contributed equally for this study.

Appendix A. Extend MIASSR to Various Medical Image Modalities

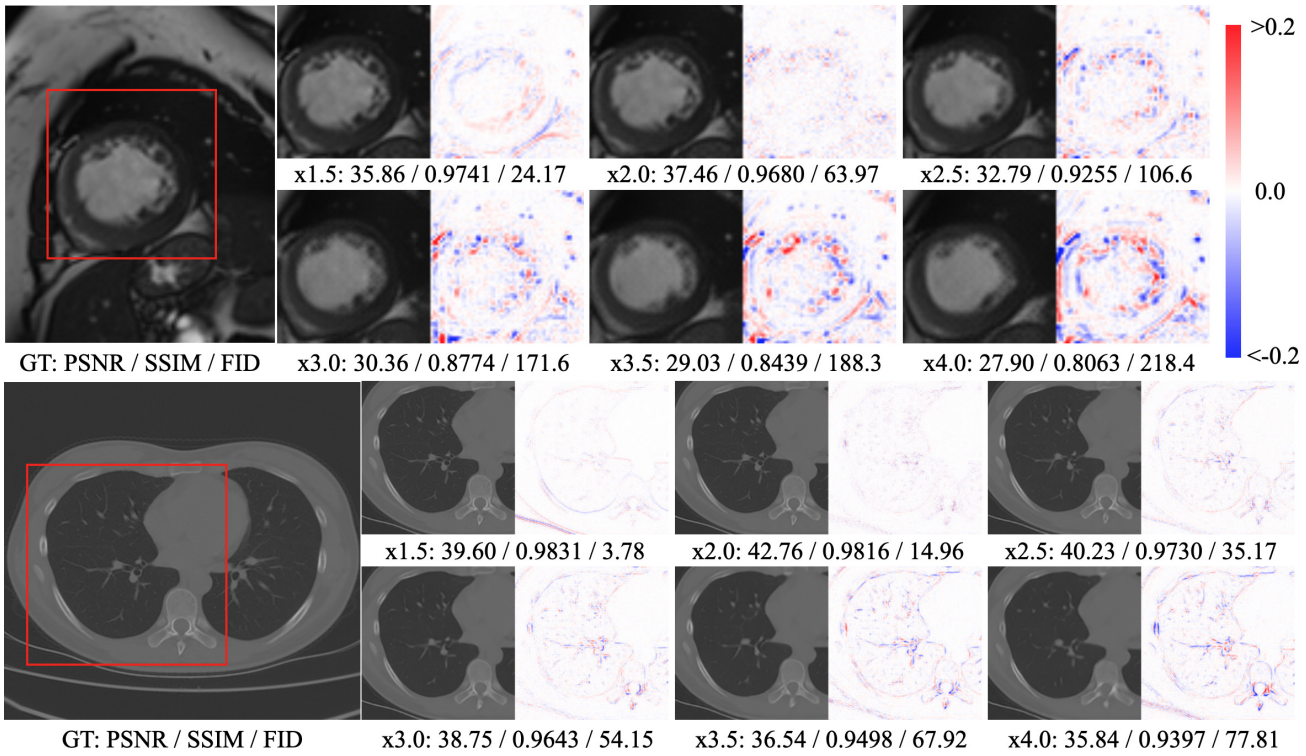


Fig. A.1. With transfer learning, MIASSR can be applied to new medical image modalities conveniently and effectively. We have extended it to cardiac MR images (top), and CT images of COVID-CT patients (bottom). GT image is plotted on the left, while SR images with various magnification scales and the pixel-wise errors are on the right.

Int. J. Neur. Syst. 2021.31. Downloaded from www.worldscientific.com
by IMPERIAL COLLEGE LONDON on 10/08/21. Re-use and distribution is strictly not permitted, except for Open Access articles.

Int. J. Neur. Syst. 2021.31. Downloaded from www.worldscientific.com
by IMPERIAL COLLEGE LONDON on 10/08/21. Re-use and distribution is strictly not permitted, except for Open Access articles.

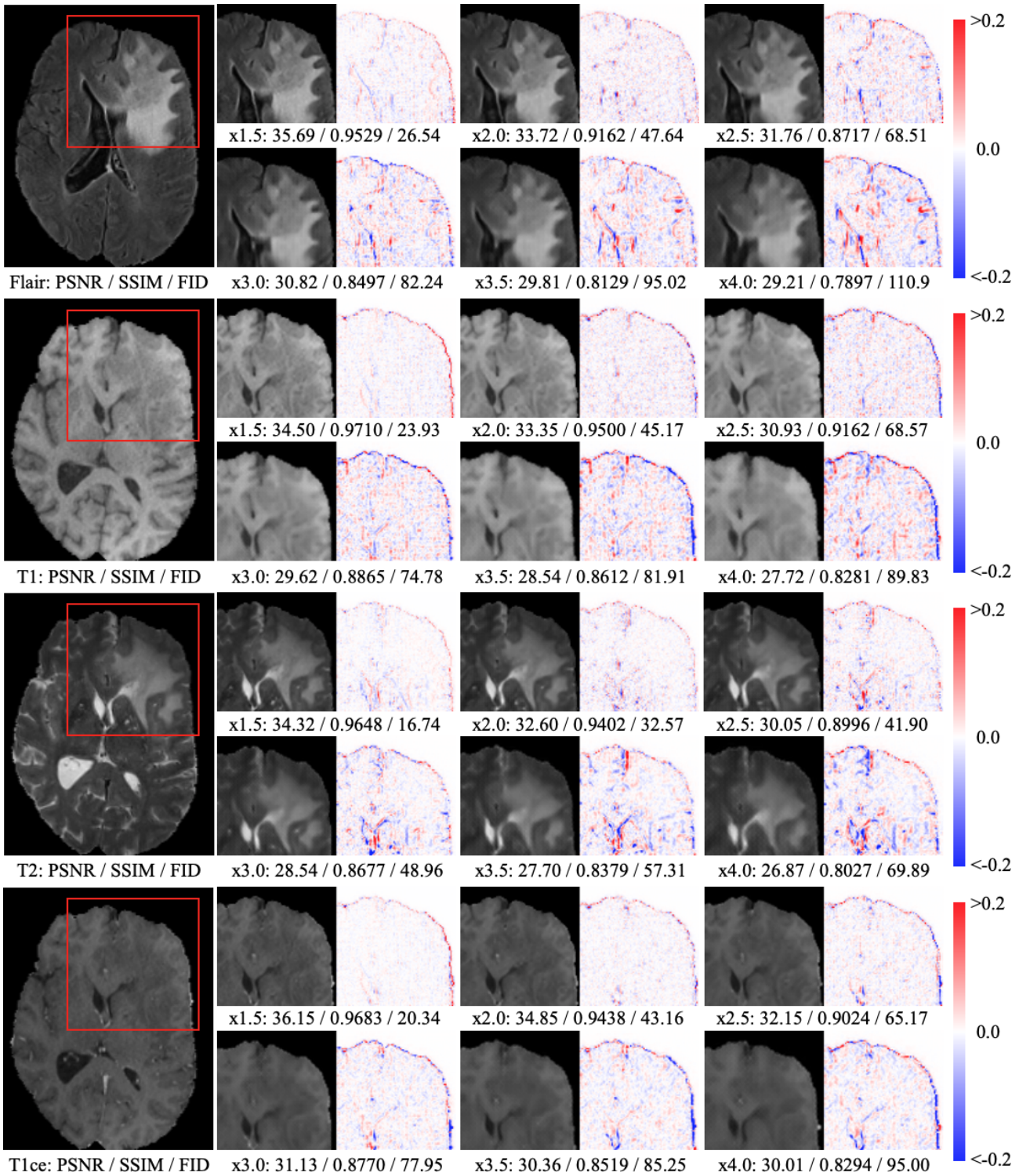


Fig. A.2. MIASSR has successfully worked with multi-modal brain MR image dataset BraTS. All images are converted to [0, 1]. Differences between SR images and GT images are rendered.

Int. J. Neur. Syst. 2021.31. Downloaded from www.worldscientific.com
by IMPERIAL COLLEGE LONDON on 10/08/21. Re-use and distribution is strictly not permitted, except for Open Access articles.

References

1. G. Mirzaei, A. Adeli and H. Adeli, Imaging and machine learning techniques for diagnosis of alzheimer's disease, *Rev. Neurosci.* **27**(8) (2016) 857–870.
2. G. Mirzaei and H. Adeli, Resting state functional magnetic resonance imaging processing techniques in stroke studies, *Rev. Neurosci.* **27**(8) (2016) 871–885.
3. M. Leming, J. M. Górriz and J. Suckling, Ensemble deep learning on large, mixed-site FMRI datasets in autism and other tasks, *Int. J. Neural Syst.* **30**(7) (2020) 2050012.
4. D. Castillo-Barnes, F. J. Martinez-Murcia, A. Ortiz, D. Salas-Gonzalez, J. RamÍrez and J. M. Górriz, Morphological characterization of functional brain imaging by isosurface analysis in Parkinson's disease, *Int. J. Neural Syst.* **30**(9) (2020) 2050044.
5. Y. Li, B. Sixou and F. Peyrin, A review of the deep learning methods for medical images super resolution problems, *IRBM* **42** (2020) 120–133.
6. L. Schermelleh, R. Heintzmann and H. Leonhardt, A guide to super-resolution fluorescence microscopy, *J. Cell Biol.* **190**(2) (2010) 165–175.
7. W. Yang, X. Zhang, Y. Tian, W. Wang, J.-H. Xue and Q. Liao, Deep learning for single image super-resolution: A brief review, *IEEE Trans. Multimedia* **21** (2019) 3106–3121.
8. Z. Wang, J. Chen and S. C. Hoi, Deep learning for image super-resolution: A survey, *IEEE Trans. Pattern Anal. Mach. Intell.* (2020), doi:10.1109/TPAMI.2020.2982166.
9. X. Wang and W. Q. Yan, Human gait recognition based on frame-by-frame gait energy images and convolutional long short-term memory, *Int. J. Neural Syst.* **30**(1) (2020) 1950027.
10. P. Mishra, C. Piciarelli and G. L. Foresti, A neural network for image anomaly detection with deep pyramidal representations and dynamic routing, *Int. J. Neural Syst.* **30**(10) (2020) 2050060.
11. J. Zhu, G. Yang and P. Lio, Lesion focused super-resolution, arXiv:1810.06693.
12. J. Zhu, G. Yang and P. Lio, How can we make GAN perform better in single medical image super-resolution? a lesion focused multi-scale approach, *2019 IEEE 16th Int. Symp. Biomedical Imaging* (*ISBI 2019*) (IEEE, 2019).
13. C. You, G. Li, Y. Zhang, X. Zhang, H. Shan, M. Li, S. Ju, Z. Zhao, Z. Zhang, W. Cong *et al.*, CT super-resolution GAN constrained by the identical, residual, and cycle learning ensemble (GAN-circle), *IEEE Trans. Med. Imaging* **39**(1) (2019) 188–203.
14. J. Park, D. Hwang, K. Y. Kim, S. K. Kang, Y. K. Kim and J. S. Lee, Computed tomography super-resolution using deep convolutional neural network, *Phys. Med. Biol.* **63**(14) (2018) 145011.
15. R. Keys, Cubic convolution interpolation for digital image processing, *IEEE Trans. Acoust. Speech Signal Process.* **29**(6) (1981) 1153–1160.
16. H. Ji and C. Fermüller, Wavelet-based super-resolution reconstruction: Theory and algorithm, *Computer Vision – ECCV 2006* (Springer, Berlin Heidelberg, Berlin, Heidelberg, 2006), pp. 295–307.
17. R. Timofte, V. De Smet and L. Van Gool, A+: Adjusted anchored neighborhood regression for fast super-resolution, *Asian Conf. Computer Vision*, (Springer International Publishing, 2014), pp. 111–126.
18. R. Zeyde, M. Elad and M. Protter, On single image scale-up using sparse-representations, *Int. Conf. Curves and Surfaces* (Springer Berlin Heidelberg, 2010), pp. 711–730.
19. C. Dong, C. C. Loy, K. He and X. Tang, Image super-resolution using deep convolutional networks, *IEEE Trans. Pattern Anal. Mach. Intell.* **38**(2) (2015) 295–307.
20. K. He, X. Zhang, S. Ren and J. Sun, Deep residual learning for image recognition, arXiv:1512.03385.
21. J. Kim, J. Kwon Lee and K. Mu Lee, Accurate image super-resolution using very deep convolutional networks, in *Proc. IEEE Conf. Computer Vision and Pattern Recognition* (IEEE, 2016), pp. 1646–1654.
22. M. D. Zeiler and R. Fergus, Visualizing and understanding convolutional networks, arXiv:1311.2901.
23. W. Shi, J. Caballero, F. Huszar, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert and Z. Wang, Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network, arXiv:1609.05158.
24. T. Tong, G. Li, X. Liu and Q. Gao, Image super-resolution using dense skip connections, in *Proc. IEEE Int. Conf. Computer Vision* (IEEE, 2017), pp. 4799–4807.
25. Y. Zhang, Y. Tian, Y. Kong, B. Zhong and Y. Fu, Residual dense network for image super-resolution, in *Proc. IEEE Conf. Computer Vision and Pattern Recognition* (IEEE, 2018), pp. 2472–2481.
26. C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang and W. Shi, Photo-realistic single image super-resolution using a generative adversarial network, arXiv:1609.04802.
27. K. Simonyan and A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv:1409.1556.
28. J. Johnson, A. Alahi and L. Fei-Fei, Perceptual losses for real-time style transfer and super-resolution, arXiv:1603.08155.
29. I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville and Y. Bengio, Generative adversarial networks, arXiv:1406.2661.

Int. J. Neur. Syst. 2021.31. Downloaded from www.worldscientific.com
by IMPERIAL COLLEGE LONDON on 10/08/21. Re-use and distribution is strictly not permitted, except for Open Access articles.

30. X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao and C. Change Loy, ESRGAN: Enhanced super-resolution generative adversarial networks, in *Proc. European Conf. Computer Vision* (*ECCV*), (Springer Science Business Media, 2018).
31. A. Jolicoeur-Martineau, The relativistic discriminator: A key element missing from standard GAN, arXiv:1807.00734.
32. Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong and Y. Fu, Image super-resolution using very deep residual channel attention networks, arXiv:1807.02758.
33. A. Shocher, N. Cohen and M. Irani, "zero-shot" super-resolution using deep internal learning, arXiv:1712.06087.
34. M. Haris, G. Shakhnarovich and N. Ukita, Deep back-projection networks for super-resolution, arXiv:1803.02735.
35. X. Hu, H. Mu, X. Zhang, Z. Wang, T. Tan and J. Sun, Meta-SR: A magnification-arbitrary network for super-resolution, in *Proc. IEEE Conf. Computer Vision and Pattern Recognition* (IEEE, 2019), pp. 1575–1584.
36. O. Ronneberger, P. Fischer and T. Brox, U-net: Convolutional networks for biomedical image segmentation, arXiv:1505.04597.
37. J.-Y. Zhu, T. Park, P. Isola and A. A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, arXiv:1703.10593.
38. Y. Chen, F. Shi, A. G. Christodoulou, Z. Zhou, Y. Xie and D. Li, Efficient and accurate MRI super-resolution using a generative adversarial network and 3D multi-level densely connected network, arXiv:1803.01417.
39. O. Oktay, E. Ferrante, K. Kamnitsas, M. Heinrich, W. Bai, J. Caballero, S. A. Cook, A. De Marvao, T. Dawes, D. P. O'Regan *et al.*, Anatomically constrained neural networks (ACNNS): Application to cardiac image enhancement and segmentation, *IEEE Trans. Med. Imaging* **37**(2) (2017) 384–395.
40. B. Lim, S. Son, H. Kim, S. Nah and K. M. Lee, Enhanced deep residual networks for single image super-resolution, arXiv:1707.02921.
41. M. Arjovsky, S. Chintala and L. Bottou, Wasserstein GAN, arXiv:1701.07875.
42. I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin and A. C. Courville, Improved training of Wasserstein GANs, *Adv. Neural Inf. Process. Syst.* **30** (2017) 5767–5777.
43. M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler and S. Hochreiter, GANs trained by a two time-scale update rule converge to a local nash equilibrium, arXiv:1706.08500.
44. D. S. Marcus, T. H. Wang, J. Parker, J. G. Csernansky, J. C. Morris and R. L. Buckner, Open access series of imaging studies (OASIS): Cross-sectional MRI data in young, middle aged, nondemented, and demented older adults, *J. Cognit. Neurosci.* **19**(9) (2007) 1498–1507.
45. B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby *et al.*, The multimodal brain tumor image segmentation benchmark (BRATS), *IEEE Trans. Med. Imaging* **34** (2015) 1993–2024.
46. S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J. Kirby *et al.*, Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features, *Sci. Data* **4** (2017) 170117.
47. S. Bakas, M. Reyes, A. Jakab, S. Bauer, M. Rempfler, A. Crimi *et al.*, Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge, arXiv:1811.02629.
48. O. Bernard, A. Lalande, C. Zotti, F. Cervenansky, X. Yang, P.-A. Heng, I. Cetin, K. Lekadir, O. Camara, M. A. G. Ballester *et al.*, Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: Is the problem solved?, *IEEE Trans. Medical Imaging* **37**(11) (2018) 2514–2525.
49. P. An, S. Xu, S. Harmon, E. Turkbey, T. Sanford, A. Amalou, M. Kassin, N. Varble, M. Blain, V. Anderson, F. Patella, G. Carrafiello, B. Turkbey and B. Wood, CT images in COVID-19 [data set], *Cancer Imaging Arch.* (2020), https://doi.org/10.7937/TCIA.2020.GQRY-NC81.
50. C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens and Z. Wojna, Rethinking the inception architecture for computer vision, arXiv:1512.00567.
51. Y. Blau and T. Michaeli, The perception-distortion tradeoff, *2018 IEEE/CVF Conf. Computer Vision and Pattern Recognition* (IEEE, 2018).
52. A. Radford, L. Metz and S. Chintala, Unsupervised representation learning with deep convolutional generative adversarial networks, arXiv:1511.06434.
53. K. He, X. Zhang, S. Ren and J. Sun, Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, arXiv:1502.01852.
54. D. P. Kingma and J. Ba, Adam: A method for stochastic optimization, arXiv:1412.6980.
55. S. Bell, P. Upchurch, N. Snavely and K. Bala, Material recognition in the wild with the materials in context database, arXiv:1412.0623.
56. M. Chessa and F. Solari, A computational model for the neural representation and estimation of the binocular vector disparity from convergent stereo image pairs, *Int. J. Neural Syst.* **29**(5) (2019) 1850029.
57. P. Gaur, K. McCreadie, R. B. Pachori, H. Wang and G. Prasad, Tangent space features-based transfer learning classification model for two-class motor

imagery brain–computer interface, *Int. J. Neural Syst.* **29**(10) (2019) 1950025.
58. S. Ioffe and C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, arXiv:1502.03167.
59. A. Ortiz, J. Munilla, F. J. Martínez-Murcia, J. M. Górriz and J. Ramírez, Empirical functional PCA for 3D image feature extraction through fractal sampling, *Int. J. Neural Syst.* **29**(2) (2019) 1850040.
60. Y. Liang, F. He and X. Zeng, 3D mesh simplification with feature preservation based on whale optimization algorithm and differential evolution, *Integr. Comput.-Aided Eng.* **27** (2020) 1–19.
61. Y. Wu, F. He, D. Zhang and X. Li, Service-oriented feature-based data exchange for cloud-based design and manufacturing, *IEEE Trans. Serv. Comput.* **11**(2) (2015) 341–353.
62. O. M. Manzanera, S. K. Meles, K. L. Leenders, R. J. Renken, M. Pagani, D. Arnaldi, F. Nobili, J. Obeso, M. R. Oroz, S. Morbelli *et al.*, Scaled subprofile modeling and convolutional neural networks for the identification of Parkinson's disease in 3D nuclear imaging data, *Int. J. Neural Syst.* **29**(9) (2019) 1950010.

Int. J. Neur. Syst. 2021.31. Downloaded from www.worldscientific.com
by IMPERIAL COLLEGE LONDON on 10/08/21. Re-use and distribution is strictly not permitted, except for Open Access articles.