UNIVERSITY OF CAMBRIDGE

Deep neural networks for medical image super-resolution

Jin Zhu

Darwin College

This dissertation is submitted on March, 2023 for the degree of Doctor of Philosophy

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text. It is not substantially the same as any that I have submitted, or am concurrently submitting, for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or is being concurrently submitted, for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. This dissertation does not exceed the prescribed limit of 60 000 words.

Jin Zhu
March, 2023

Abstract

Deep neural networks for medical image super-resolution

Jin Zhu

Super-resolution plays an essential role in medical imaging because it provides an alternative way to achieve high spatial resolutions with no extra acquisition cost. In the past decades, the rapid development of deep neural networks has ensured high reconstruction fidelity and photo-realistic super-resolution image generation. However, challenges still exist in the medical domain, requiring novel network architectures, training tricks, and SR image evaluation techniques. This dissertation concentrates on backbone networks for supervised single-image super-resolution tasks on various medical images with challenging magnification scales. Besides incorporating widespread methods designed for natural images, I explore progressive learning, adversarial learning and meta-learning in end-to-end frameworks based on convolution neural networks, generative adversarial networks and vision transformers for robust medical image super-resolution. In addition to general image quality assessments, task-specific objective and subjective evaluation metrics are implemented for comprehensive comparisons. Specifically, the proposed approaches contain three directions, achieving state-of-the-art performance on diverse medical image modalities.

First, I implement progressive and adversarial learning for perceptually realistic texture generation in super-resolution tasks with challenging magnification scales (i.e. $\times 4$). I present a CNN-based multi-scale super-resolution image generator that decomposes the complex mapping problem into simpler sub-problems to avoid over-smoothing the structural information and introducing non-realistic high-frequency textures in super-resolved images. Moreover, it involves a lesion-focused training strategy and an advanced adversarial loss based on the Wasserstein distance for more efficient and stabilised training. This proposed method dramatically improves the perceptual quality of generated images, achieving comparable subjective scores of experienced radiologists with ground truth high-resolution images in the experiments of the brain and cardiac magnetic resonance images. It competed for state-of-the-art perceptual quality in medical image super-resolution in 2019 and became the pioneer of GAN-based medical image research with enduring effects.

Second, I introduce meta-learning and transfer learning to GANs for efficient and robust medical image super-resolution with arbitrary scales (e.g. $(1, 4]$). In the post-upsampling framework, I implement a lightweight network based on EDSR for productive low-resolution feature extraction and a weight prediction module for scale-free feature map upsampling. Compared with existing SISR networks, this framework supports non-integer magnification with no adverse effects of pre-/post- processing. Specifically, this approach achieves comparable reconstruction accuracy and objective perceptual quality performance with much fewer parameters than SOTA methods. Additionally, I robustly transfer the pre-trained SR model of one medical image dataset (i.e. brain MRI) to various new medical modalities (e.g. chest CT and cardiac MR) with a few fine-tuning steps. Moreover, exhaustive ablation studies are conducted to discuss the perception-distortion tradeoff and to illustrate the impacts of residual block connections, hyper-parameters, loss components and adversarial loss variants on medical image super-resolution performance.

Finally, I propose an efficient vision transformer with residual dense connections and local feature fusion to achieve superior single-image super-resolution performances of medical modalities. Due to the improved information flow, this CNN-transformer hybrid model has advanced representation capability with fewer training computational requirements. Meanwhile, I implement a general-purpose perceptual loss with manual control for desired image quality improvements by incorporating prior knowledge of medical image segmentation. Compared with state-of-the-art methods on four public medical image datasets, the proposed method achieves the best PSNR scores of 6 modalities among seven modalities with only 38% parameters of SwinIR (the most recent SOTA method). On the other hand, the segmentation-based perceptual loss increases by +0.14 dB PSNR on average for prevalent super-resolution networks without extra training costs. Additionally, I discuss potential factors for the superior performance of vision transformers over CNNs and GANs and the impacts of network and loss function components in a comprehensive ablation study.

In conclusion, this dissertation represents my research contributions of applying deep neural networks on robust medical image super-resolution tasks, including efficient network architectures, broad applicability training techniques, and clinically meaningful image quality evaluation. When publishing, these proposed approaches perform state-of-the-art on various public and private medical image datasets in simulation experiments. These algorithms potentially apply in hospitals for advanced clinical processes with proper case-specific modifications and supplementary techniques. Moreover, the novel methods and findings of super-resolution may also benefit other low-level image processing tasks, while the discussion and ablation studies provide exciting future research directions.

Acknowledgements

I have spent an unusually long period on the challenging marathon of this doctoral project. I could complete this thesis with the invaluable help and support of my family, friend, colleagues and supervisor.

I express my deepest gratitude to my supervisor, Prof. Pietro Lió, for the tireless support, great patience, free research interests and one-in-a-lifetime opportunity to pursue a PhD at such a prestigious institution. His kindness, professional spirit, expertise and valuable advice have benefited my life. I also thank my adviser and co-author, Dr Guang Yang from Imperial College, for the insightful discussions and detailed suggestions on academic writing and medical image knowledge. Without his generous and supportive help, I may rarely find many exciting research topics in this dissertation.

I appreciate the kindness from the members of the AI Group and the support from the staff of the Computer Laboratory, with particular thanks to Mrs Lise Gough for her kindest and warmest advice.

I am lucky and privileged to spend years in such a beautiful city with such brilliant people. I must thank my friends from whom I have learned much in my profession and in person. Among them, my special thanks to Tiago Azevedo, Duo Wang and Leran Cai. You guys help me go forward when struggling with the confusion of the PhD.

I enjoy the research internship in Intuitive Surgical, working in Anthony Jarc's group. Many thanks to Anthony, Aneeq and Kiran for their professional collaborations and insightful discussion, which help me earn more academic-industrial understanding. I am also grateful to Linlin for offering me the accommodation and Jianyu for recommending me for this opportunity. My special thanks to Ziyuan Fang for everything we have enjoyed together in Atlanta and Beijing as close friends.

The last paragraph is reserved for my family for the care, understanding, great support, and much more.

Contents

List of Figures

List of Tables

Abbreviations

ACDC the open-access CT image dataset for automated cardiac diagnosis challenge.

AWGN additive white Gaussian noise.

BraTS the brain tumour segmentation dataset, one of the biggest open-access medical image segmentation benchmarks, including multi-modal brain MR scans and manual labels of tumour components.

CNN convolution neural network.

COVID coronavirus disease 2019.

COVID-CT an open-access dataset of CT scans of patients with COVID infections.

COVID-19 CT another open-access dataset of 3D CT scans of 20 patients with COVID infections and annotations.

CT computed tomography (images/scans).

DNN deep neural network.

DSTB a dense swin (i.e. shifted-window attention) transformer block proposed in Chapter. 5.

DT-CMR diffusion tensor cardiovascular magnetic resonance (images/scans).

EDSR a super-resolution network with enhanced residual blocks by removing the batch normalisation.

FC fully-connected layer, also called as linear layer.

FID Frechet Inception Distance, an automatic perceptual quality evaluation method of images by measuring the distance between the distributions of two groups of images in the feature domain.

FPS inference frame rate (i.e. frame-per-second), a metric to evaluate the model efficiency on throughput: bigger FPS means a faster model.

GAN generative adversarial networks.

GFF global feature fusion.

GPU graphics processing unit.

GT ground truth.

HAN a CNN-based super-resolution network with holistic attention (i.e. channel-spatial attention and layer-wise attention).

HR high-resolution (images/feature maps/matrices etc.).

IQA image quality assessment.

LFF local feature fusion.

LFSR lesion-focused super-resolution, an advanced super-resolution method by introducing lesion detection and ROI-focused training to the original SRGAN.

LEG-CMR Late gadolinium-enhanced cardiovascular magnetic resonance (images/scans).

LR low-resolution (images/feature maps/matrices etc.).

MACs multi-add calculations, a metric to evaluate the model efficiency: fewer MACs means a more efficient model.

MAE mean absolute error (i.e. L1 loss).

MetaRDN a super-resolution network which proposes meta-upscale module for arbitrary scale super-resolution.

MIASSR the proposed method in Chapter 4 for scale-free medical image super-resolution.

MLP multi-layer perceptron.

MOS mean opinion score, an perceptual quality evaluation method of images by manual scores.

MR magnetic resonance (images/scans) while **MRI** denotes MR images.

MSE mean squared error.

MS-GAN the proposed method in Chapter 3 for medical image super-resolution, consisting of a multi-scale image generator.

OASIS the open access series of imaging studies dataset.

PSNR peak signal-to-noise ratio, the most widespread evaluation criteria for image restoration tasks, measuring the pixel-wise accuracy.

RCAN a CNN-based super-resolution network with residual blocks and channel attention.

RDN a super-resolution network with residual dense blocks and local feature fusion.

RDST the proposed method in Chapter 5 for medical image super-resolution with residual dense vision transformer.

RDSTB a residual dense swin (i.e. shifted-window attention) transformer block proposed in chapter 5.

ReLU an activation function, Rectified Linear Unit.

ROI region of interest.

SISR single image super-resolution, an operation/task to apply super-resolution based on the input of only one image.

SOTA state-of-the-art.

SR super-resolution, an operation/task to increase the resolution of images.

SRGAN a super-resolution network with GAN for photo-realistic results.

SRResNet a super-resolution network with residual blocks, the SR image generator proposed in SRGAN.

SSIM structural similarity, widespread evaluation criteria for image restoration tasks, reflecting the global structural information.

STL a swin (i.e. shifted-window attention) transformer layer.

SwinIR a CNN-transformer hybrid super-resolution network with shifted-window attention.

U-Net a network proposed in 2015 for medical image segmentation, the U-shape architecture also benefits image enhancement tasks.

VGG a network proposed in 2014 for objection detection, popularly used in perceptual loss function design.

ViT vision transformer.

WGAN Wasserstein GAN.

WGAN-GP Wasserstein GAN with gradient penalty.

Notations

α a scalar (so are β, γ, η and λ), normally indicate the scale of loss function components.

$\mathcal{A}$ unified form of the single-head attention operation in vision transformers.

$\boldsymbol{a}_{i,j}$ the normalised attention vector across the i-th and j-th tokens in vision transformers.

$\boldsymbol{b}_d$ a vector/matrix with indicated shape (e.g. d), as the trainable bias of one layer/node in neural networks.

$\mathcal{B}$ a block in neural networks, consisting of batch normalisation layers, convolution layers, activation layers etc. $\phi_{\mathcal{B}}$ indicates its trainable parameters. The potential superscript indicates a successive connection of blocks (e.g. $\mathcal{B}^n$ means n stacked blocks) and the subscript indicates the type of block (e.g. $\mathcal{B}_R$ denotes a residual block).

$\mathcal{T}_{d\to g}(\cdot)$ a bottleneck/local feature fusion module consisting of convolution/FC layers in neural networks. It compresses the dimension of feature maps from d to g (so-called the growth rate in dense blocks.

$\mathcal{C}$ a convolution layer, $\phi_{\mathcal{C}}$ indicates its trainable parameters.

$\star$ convolution operation, modifies the shape and values of one function (e.g. an image or feature maps) by kernels. This is the basic calculation of convolution neural networks..

$\star\boldsymbol{\kappa}_i$ convolution operation with the i-th kernel $\boldsymbol{\kappa}_i$.

$\varrho(\cdot)$ the patch cropping operation in the meta-upscale module.

d a scalar, normally indicates the embedding dimension of vision transformers.

$\mathbb{D}$ a set of label and image pairs, as the training/testing dataset. Additionally, $\mathbb{X}$ indicates the set of inputs while $\mathbb{Y}$ indicates the set of outputs..

$\star\boldsymbol{\delta}$ image degradation during capturing process.

D a network, normally indicates the discriminator in GAN.

$\downarrow_s$ down-sampling with scale s.

$\mathcal{D}(\cdot)$ a dense swin transformer block proposed in chapter 5.

$\mathbb{E}$ the expectation.

$\mathcal{F}(\cdot)$ the feature extraction module in super-resolution networks.

$\boldsymbol{F}$ the feature maps of one layer/module, potential subscript and superscript indicate the shape (e.g. H_{out}, W_{out}, c) and order (e.g. in or out).

$\boldsymbol{f}_i$ the i-th feature map calculated by the i-th kernel in a convolution layer.

Ψ embedding function in transformers.

$\boldsymbol{n}_\sigma$ Gaussian noise with standard deviation σ.

G a network, normally indicates the SR image generator.

$\mathcal{H}(\cdot)$ the CNN-based shallow feature extraction head in SwinIR and RDST.

H, W height and width of images/feature maps.

$\boldsymbol{I}_{hr}$ a high-resolution image obtained from the training/testing dataset (i.e. the ground truth).

$\boldsymbol{I}_{lr}$ a low-resolution image obtained from the training/testing dataset (i.e. the input of the super-resolution networks).

$\mathbb{I}$ a set of images.

$\mathbb{L}, \mathbb{H}$ the low-dimension and high-dimension spaces of images.

$\boldsymbol{I}_{sr}$ a high-resolution image super-resolved by super-resolution networks (i.e. the output of the SR networks). The potential superscript indicates the magnification scale (e.g. $\boldsymbol{I}_{sr}^{\times 4}$ denotes a $\times 4$ super-resolution result).

$\boldsymbol{K}_{h,w,c}$ the convolution kernels of one convolution layer with shape h, w, c, where c indicates the number of kernels/channels (i.e. layer width) and h, w indicate the shape of each 2D kernel.

k a scalar, the kernel size.

$\boldsymbol{L}$ ground truth labels in segmentation tasks.

LD a lesion detection network.

ι a scalar, normally indicates the learning rate in gradient descent.

$\mathcal{L}$ a loss function.

$\mathcal{M}(\cdot)\uparrow_s$ the up-sampling module in super-resolution networks for feature maps magnification with scale s. Its simplified version is $\mathcal{M}_s$.

$\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}$ the query, key and value matrices for all tokens of one attention layer in vision transformers.

$\mathcal{N}$ a normalisation layer (e.g. batch normalisation or layer normalisation), $\phi_{\mathcal{N}}$ indicates its trainable parameters.

$\mathcal{P}(\cdot)$ a multi-layer perceptron.

$\boldsymbol{P}$ predicted labels by pre-trained U-Net in segmentation tasks.

s a scalar, the magnification scale in super-resolution tasks.

$\mathcal{S}(\cdot)$ swin transformer layer.

$\boldsymbol{\theta}$ trainable weights of a network (e.g. $\boldsymbol{\theta}_G$ indicates the parameters of the network G).

$\boldsymbol{t}_i$ the i-th token in transformers after embedding.

U a U-Net trained for medical image segmentation. It supports the segmentation-based perceptual loss in chapter 5. Specifically, $U[E_i](\cdot)$ indicates the output feature maps of the i-th encoder block and $U[D](\cdot)$ indicates the output feature maps of the decoder. $\boldsymbol{\theta}_U$ denotes its trainable parameters.

$\varphi(\cdot)$ a non-linear activation layer in neural networks.

$\boldsymbol{q}_i, \boldsymbol{k}_i, \boldsymbol{v}_i$ the query, key and value vectors of the i-th token in vision transformers after embedding.

$\mathcal{V}$ the pre-trained VGG network, while $\mathcal{V}_l(\cdot)$ indicates the output feature maps of the l-th layer in the network..

$\boldsymbol{w}_c$ weights of a convolution kernel.

$\boldsymbol{W}_{d,k}$ a matrix with the indicated shape (e.g. $[d, k]$), as the weights of one layer/operation.

$\boldsymbol{w}_q, \boldsymbol{w}_k, \boldsymbol{w}_v$ matrix with shape $d \times d$ for embedding operations of query, key and value vectors $\boldsymbol{q}, \boldsymbol{k}, \boldsymbol{v}$ in vision transformers, where d is the embedding dimension.

$\mathcal{W}(\cdot)$ the weights prediction network in the meta-upscale module.

x a scalar, the input of a function.

$\boldsymbol{x}$ a vector/matrix as the input of a function/layer/network, potential subscript may indicate the index or the shape (e.g. $\boldsymbol{x}_i$ means the i-th sample while $\boldsymbol{x}_d$ is a 1-dimension vector with the length of d).

y a scalar, the output of a function.

$\bar{\boldsymbol{y}}$ a vector/matrix, the ground truth label/values/image in the training/testing dataset.

$\boldsymbol{y}$ a vector/matrix as the output of a function/layer/network, potential subscript may indicate the shape or index (e.g. $\boldsymbol{y}_i$ means the i-th sample while $\boldsymbol{y}_d$ is a 1-dimension vector with the length of d).

CHAPTER 1

Introduction

Download the high-quality PDF of this dissertation with uncompressed figures *here*[1].

1.1 Motivation

Medical images are crucial in the current clinical process, including early detection, staging, guiding intervention procedures and surgeries, radiation therapy, and monitoring disease recurrence [1]. For example, computed tomography (CT) and magnetic resonance (MR) scans are widely used in the diagnosis and study of Alzheimer's disease [2], stroke [3], autism [4], Parkinson's disease [5] and coronavirus disease (COVID) [6]. Although people have witnessed the importance of in-vivo radiology images, their spatial resolution is subject to the scan time, body motion, dose limit and hardware configurations. Thus, super-resolution methods are introduced as alternative post-processing to achieve higher resolution and better image quality without extra acquisition costs [7].

Image super-resolution (SR) is a process to recover an image of high-resolution (HR) from low-resolution (LR) versions. Depending on the number of input and output images, there are two main categories: single image super-resolution (SISR) and multi image super-resolution. With the rapid development of deep learning algorithms, SISR methods with neural networks achieve superior performance on natural images than the previous interpolation-based, reconstruction-based, and learning-based methods [8–10]. However, introducing deep neural networks to medical image super-resolution tasks is still an open problem. In the clinic, super-resolved images always proceed to medical image analysis tasks, and the datasets are relatively small [11–13]. Thus, super-resolution methods for medical images require novel mechanisms and modifications on training datasets, loss functions, evaluation metrics and network architecture design to preserve sensitive information

[1] `https://www.dropbox.com/sh/8juz2jv9bp09hsg/AAC332wIaVX2M06_hEZyDadHa?dl=0`

and to enhance the structures of interest for radiologists and physicians [7]. It is worth discussing the capacity and limitations of deep learning methods in this typical low-level medical image enhancement task. Additionally, successful characteristics and novel findings in SR tasks may potentially benefit other pixel-to-pixel medical image analysis tasks such as reconstruction [14], synthesis [15] and denoising [16]. The fundamental goal of these tasks is non-linear transform approximation from one image space to another with high pixel-wise accuracy and global perceptual fidelity. Thus, they share common mechanisms with super-resolution solutions, such as network framework, loss function and evaluation assessment design.

Furthermore, the image quality assessment (IQA) of medical images is different to natural images [17]. Using IQA metrics in medical SR tasks must be more reliable with domain-specific prior knowledge. For example, artefacts of medical images are unique because they are related to the imaging system (e.g. ghosting in MR images) and scanning process (e.g. blurring caused by patient motion) [18]. Meanwhile, medical images usually attach importance to focal lesions and may sacrifice the overall image quality. Thus, the bias, efficiency and robustness of widespread IQA metrics are worth discussing in medical image SR tasks. Although the quality measurement of medical images does not equal diagnostic accuracy [19], radiologists and medical consultants always prefer high-quality images for accurate diagnosis. Thus, involving related medical image analysis topics for task-based evaluation is necessary to benefit the super-resolution pipeline [20].

In summary, medical image super-resolution is essential in the clinic process and medical image analysis fields. It can lead to high-quality medical images, decreasing scan costs and improving user experience. At the same time, the novel approaches and findings of super-resolution may benefit a wide range of pixel-wise low-level image processing tasks. As a domain-specific task, innovations in training datasets, loss functions, model architectures and evaluation metrics for medical images are necessary for superior results, besides incorporating deep learning-based methods for natural images. It is also worth exploring the connection between super-resolution and related medical image analysis tasks and its applicability to various medical image modalities.

1.2 Contribution

The fundamental aim of this dissertation is to explore deep neural networks (DNNs) for robust medical image super-resolution with better performance and efficiency, which can apply to a broad range of medical modalities. In the past decades, the rapid development of neural networks has led to dramatic performance improvement in a wide range of

computer vision tasks [21], especially on natural image super-resolution enhancement tasks [9]. The following chapters will present how my proposed DNN-based methods achieve state-of-the-art (SOTA) results in comparison studies with existing SISR algorithms on various public and private medical datasets. I explore supervised methods based on convolution neural networks (CNNs), generative adversarial networks (GANs) and vision transformers (ViTs) for single image super-resolution tasks with specific magnification scales and arbitrary scales. The main directions include efficient network architectures, training tricks for better performance with broad applicability, and more straightforward evaluation. Here, I elaborate on four aspects and highlight the main novel contributions of my research work:

1. **Network architectures** Novel networks based on CNNs and vision transformers are proposed for robustness, efficiency and performance improvement. Avoiding unstable training and unrealistic textures is demanding when adversarial learning applies to super-resolution tasks with large magnification scales (e.g. $\times 4$). Thus, I develop a CNN-based multi-scale SR image generator (so-called MS-GAN) in **Chapter 3**, which decomposes the challenging $\times 4$ SR task into a series of two $\times 2$ SR tasks. It results in the high perceptual quality of SR images without generating unrealistic textures, equalling high-resolution ground truth images. Meanwhile, I conduct a comprehensive comparison study of CNN-based SR image generators in **Chapter 4**. Based on the feature extraction network architectures and hyper-parameter discussions, I implement an efficient CNN model for scale-free image super-resolution (so-called MIASSR). Additionally, I develop a residual dense vision transformer (so-called RDST) in **Chapter 5** by incorporating dense connection and local feature fusion of CNNs to shifted-window attention transformers. It achieves state-of-the-art SR performance with a rapid reduction of trainable parameters.

2. **Training tricks** Although perceptual loss and adversarial loss are popularly used in the super-resolution of natural images, operating them correctly on medical images is difficult. Because medical datasets are relatively small, incorporating GANs in medical image SR causes time-consuming hyper-parameter searching and network warm-up training. Thus, I conduct comprehensive comparison studies of GAN variants, the proportion of loss function components and training processes on integral-scale and scale-free super-resolution tasks in **Chapter 3** and **4**. Based on these conclusions, I provide a guideline for using Wasserstein GAN with gradient penalty (WGAN-GP)[22] in medical image super-resolution tasks, which dramatically improves the perceptual quality with declining the time-consuming warm-up. Additionally, I restrict the super-resolution network to the regions of interest to avoid the adverse effects of backgrounds by involving lesion detection as pre-processing

of super-resolution in **Chapter 3**. Moreover, existing perceptual losses primarily depend on natural images, while I highlight the limitation of applying them to medical images. Instead, I develop a novel perceptual loss based on medical image segmentation in **Chapter 5**. By incorporating relevant prior knowledge of high-level medical image analysis networks, the segmentation-based perceptual loss variants robustly work with SOTA CNN and transformers, leading to a notable advancement of SR image quality.

3. **Evaluation metrics** In a supplementary manner of widely-used image quality assessment metrics of SR image reconstruction fidelity, I evaluate perceptual quality for human viewers and corresponding medical image analysis tasks. In **Chapter 3**, I apply an opinion scoring metric to measure the quality of human perception by experienced radiologists with manual grades and markers of artefacts. Meanwhile, I introduce the objective perceptual quality metric of generated natural images to medical image SR evaluation and discuss the perception-distortion trade-off in **Chapter 4**. Since one primary purpose of medical image SR is to benefit related medical image analysis tasks with higher resolution, task-based evaluation also applies in this dissertation. The performances achieved in the downstream tasks, such as feature maps reconstruction and segmentation, are used as metrics of the machine perception in **Chapter 3** and **5**.

4. **General applicability** My research focuses on radiology images and involves various public and clinical datasets to illustrate applicability comprehensively. All the proposed methods compare with SOTA SISR methods on various medical image modalities, including brain and cardiac MR images and chest CT scans. Public medical image datasets such as single-/multi- brain MR contains plenty of well-processed data, so ablation studies of each method mainly depend on them for general and reliable conclusions. On the other hand, I also explore the robust performance of modifying pre-trained models to new medical image modalities with transfer learning because large and clean datasets are hard to obtain in the clinic. In summary, this dissertation entangles seven medical image datasets, including two open-accessed multi-modal brain MR datasets w/wo tumours, one public and two clinical cardiac MR datasets, and two released chest CT datasets of COVID patients.

1.3 Organisation

The dissertation is organised as follows (Fig. 1.1):

Chapter 2: In this chapter, I give the necessary background on medical image super-

resolution, including preliminaries of deep neural networks and medical imaging techniques. I also review the development of deep learning-based SISR methods, which have had significant impacts over the past decades, by illustrating the network architectures, loss functions and evaluation metrics. Particularly I summarise the super-resolution applications on medical images with a discussion of their limitations to better understand the motivation of my research work.

Chapter 3: This chapter presents my research contributions to medical image super-resolution with a large magnification scale. I propose a multi-scale SR image generator with a lesion-focused strategy (MS-GAN) by incorporating generative adversarial networks on this challenging task. I conduct simulation experiments on one public and two clinical datasets to compare the proposed and SOTA SR methods by objective metrics, experts' opinion scores and performance in clinical downstream analysis tasks. I also discuss the impacts of lesion-focused training and a guideline for using GANs in medical image super-resolution tasks without unstable training.

Chapter 4: This chapter presents my research contributions to medical image super-resolution with arbitrary magnification scales. I propose an efficient GAN-based SR image generator MIASSR that incorporates the weights prediction method of meta-learning. Simulation experiments are conducted on four public datasets to compare the proposed and SOTA SR algorithms on reconstruction fidelity and human perception. I involve the objective metrics of perceptual quality and discuss the fidelity-perception trade-off in medical image super-resolution tasks. Well-designed ablation studies address the impacts of network architectures, hyper-parameters, adversarial learning and loss function components.

Chapter 5: This chapter presents my research contributions to medical image super-resolution with vision transformers. I propose a residual dense vision transformer (RDST) that incorporates the practical architecture design of CNNs to SOTA vision transformers. A novel perceptual loss is invented that leads to superior SR results by connecting super-resolution with medical image segmentation. In addition to objective metrics of image quality, the performance of SR images in the downstream segmentation task is considered supplementary. Simulation experiments on four public datasets are conducted to compare the proposed and SOTA SR methods on reconstruction fidelity and segmentation accuracy. I also discuss the essential factors of vision transformers performing better than CNNs and the impacts of the segmentation-based perceptual loss and its broad applicability.

Chapter 6: In this chapter, I conclude the contributions made in this dissertation,

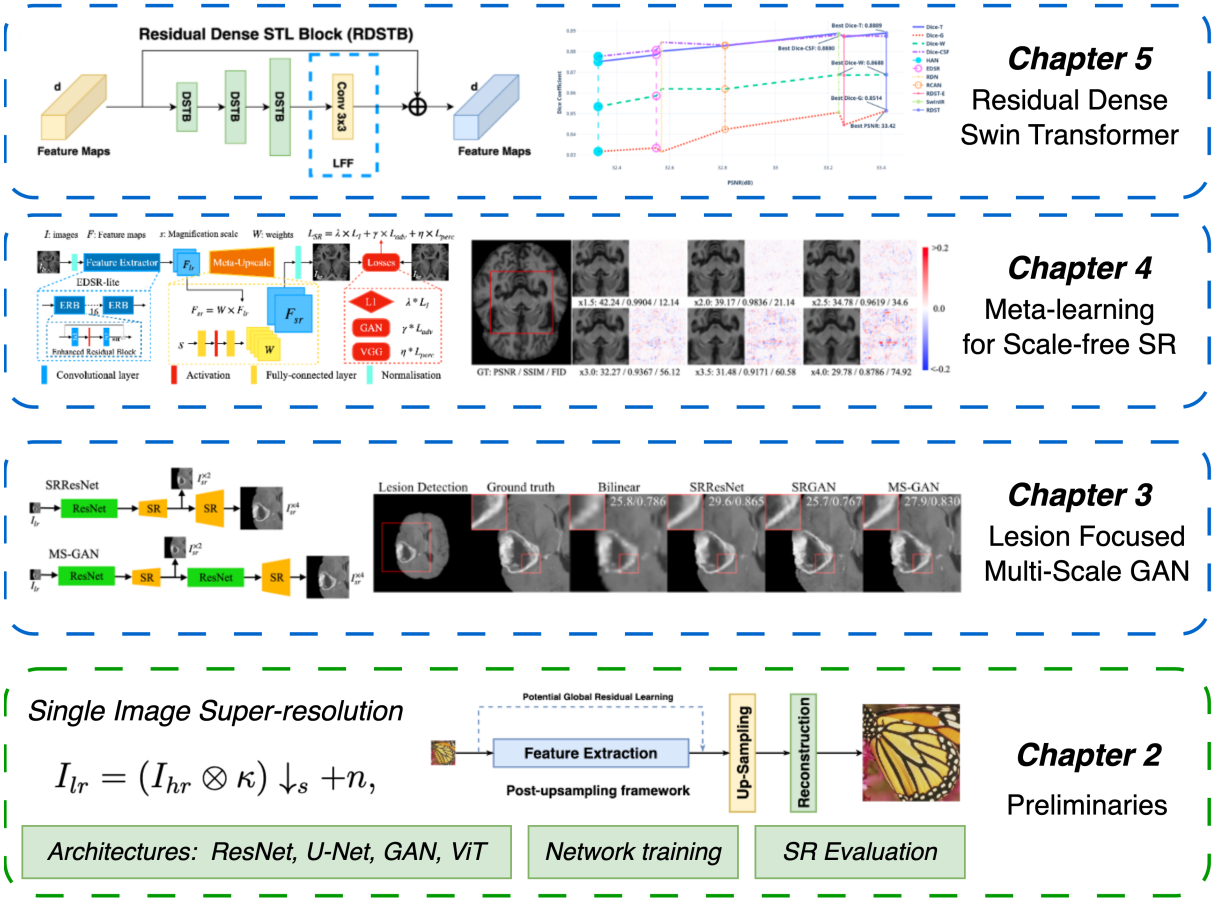


Figure 1.1: An overview of the main contributions presented in this dissertation. Preliminaries of deep neural networks and single image super-resolution are firstly introduced in **Chapter 2** with a review of medical image applications. Then, a CNN-based lesion focused multi-scale GAN is presented **Chapter 3** for perceptually realistic texture generation. Following, weight prediction of meta-learning is introduced to GANs for scale-free magnification in **Chapter 4**. Lastly, a residual dense vision transformer is implemented with a segmentation-based perceptual loss for advanced super-resolution performance in **Chapter 5**.

summarise the results, review the research questions that arise from the work presented, and propose potential research directions in the future.

1.4 Publication List

The work of MS-GAN in **Chapter 3** has been publicly realised on `https://github.com/GinZhu/MSGAN.` with the following publications [23–26]:

1. **Zhu, J.**, Yang, G. and Lio, P., 2019, March. Lesion focused super-resolution. In *Medical Imaging 2019: Image Processing* (Vol. 10949, pp. 401-406). SPIE. `https://arxiv.org/abs/1810.06693`

2. **Zhu, J.**, Yang, G. and Lio, P., 2019, April. 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 (ISBI 2019)* (pp. 1669-1673). IEEE. `https://arxiv.org/abs/1901.03419`

3. **Zhu, J.**, Yang, G., Ferreira, P., Scott, A., Nielles-Vallespin, S., Keegan, J., Pennell, D., Lio, P. and Firmin, D., 2019. A ROI focused multi-scale super-resolution method for the diffusion tensor cardiac magnetic resonance. *Proc Int Soc Magn Reson Med (ISMRM), 1.* `https://cds.ismrm.org/protected/19MProceedings/PDFfiles/0778.html`

4. **Zhu, J.**, Yang, G., Wong, T., Mohiaddin, R., Firmin, D., Keegan, J. and Lio, P., 2019. A single-image super-resolution method for late gadolinium enhancement CMR. *In Proceedings of International Society for Magnetic Resonance in Medicine (ISMRM).* (p. 2028).

The work of MIASSR in **Chapter 4** has been publicly realised on `https://github.com/GinZhu/MIASSR.` with the following publications [27, 28]:

5. **Zhu, J.**, Tan, C., Yang, J., Yang, G. and Lio', P., 2021. Arbitrary scale super-resolution for medical images. *International Journal of Neural Systems*, 31(10), p.2150037. `https://www.worldscientific.com/doi/10.1142/S0129065721500374`

6. Tan, C.*, **Zhu, J.***[2] and Lió, P., 2020, June. Arbitrary scale super-resolution for brain MRI images. In *IFIP International Conference on Artificial Intelligence Applications and Innovations* (pp. 165-176). Springer, Cham. `https://arxiv.org/abs/2004.02086`

The work of RDST in **Chapter 5** has been publicly realised on `https://github.com/GinZhu/RDST.` with the following publications [29]:

7. **Zhu, J.**, Yang, G. and Lió, P., 2023. A residual dense vision transformer for medical image super-resolution with segmentation-based perceptual loss fine-tuning. *arXiv:* `https://arxiv.org/abs/2302.11184`

[2] * indicates equal contributions

Additionally, I have also joined in related medical image analysis tasks such as segmentation and MR reconstruction, leading to the following publications [30–34]:

8. **Zhu, J.**, Yang, G., Wong, T., Mohiaddin, R., Firmin, D., Keegan, J. and Lio, P., USR-Net: A Simple Unsupervised Single-Image Super-Resolution Method for Late Gadolinium Enhancement CMR. `https://cds.ismrm.org/protected/20MProceedings/PDFfiles/3537.html`

9. **Zhu, J.**, Wang, D., Teng, Z. and Lio, P., 2017. A multi-pathway 3d dilated convolutional neural network for brain tumor segmentation. *Proceedings of the International MICCAI BraTS Challenge*, pp.342-347.

10. Wang, D., Zhang, R., **Zhu, J.**, Teng, Z., Huang, Y., Spiga, F., Du, M.H.F., Gillard, J.H., Lu, Q. and Liò, P., 2018, March. Neural network fusion: a novel CT-MR aortic aneurysm image segmentation method. In *Medical Imaging 2018: Image Processing* (Vol. 10574, pp. 542-549). SPIE.

11. Lv, J., **Zhu, J.** and Yang, G., 2021. Which GAN? A comparative study of generative adversarial network-based fast MRI reconstruction. *Philosophical Transactions of the Royal Society A*, 379(2200), p.20200203.

12. Yang, G., Lv, J., Chen, Y., Huang, J. and **Zhu, J.**, 2021. Generative Adversarial Networks (GAN) Powered Fast Magnetic Resonance Imaging–Mini Review, Comparison and Perspectives. *arXiv preprint arXiv:2105.01800.*

CHAPTER 2

Preliminaries

In this chapter, I give the necessary background regarding deep neural networks and single image super-resolution with a brief review of current applications on medical images for a better understanding of my research motivation and contribution in the following three chapters.

2.1 Deep neural netwokrs

In past decades, people have witnessed the revolutionary success of deep learning in a broad range of domains, such as computer vision [21] and natural language processing [35]. Thanks to the rapid development of affordable and efficient computing hardware (e.g. GPU), machine learning platforms (e.g. TensorFlow [36] and PyTorch [37]) and vibrant open-source communities, researchers can implement deep neural networks concisely like playing LEGO blocks. In this section, I introduce the basic concepts of deep neural networks, including several widespread network architectures with their fundamental units and some general training strategies. The fantastic references [38, 39] refer to a thorough and systematic overview of the long history of deep neural networks.

2.1.1 Layers and network units

Any deep neural network can be considered a stack of linear and non-linear layers with corresponding operations. Generally, the output (i.e. feature maps) of one layer serves as the input of the next layer, and sometimes feature maps are fed to other layers by skip-connections. In most cases, the successive layers utilise non-linear activation functions in between to facilitate the representation capability.

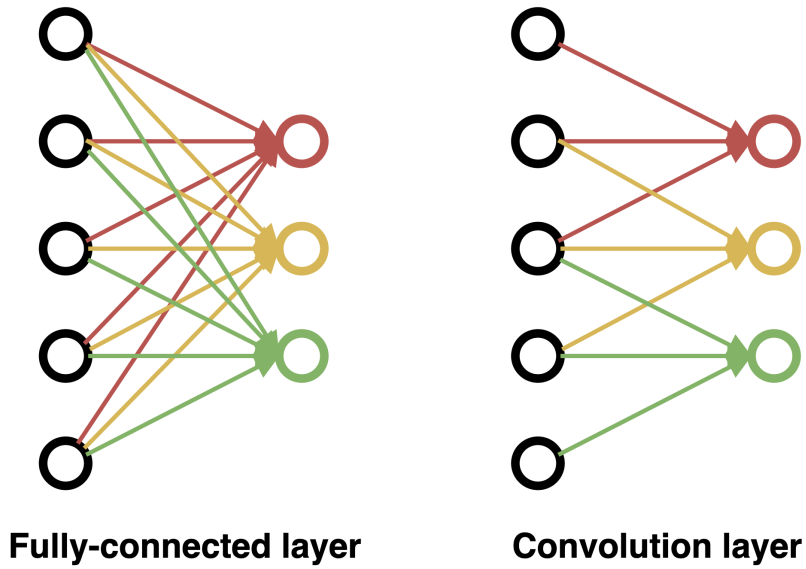


Figure 2.1: Comparison of fully-connected and convolution layers. Compared to FC layers, convolution layers avoid the expanding computational cost with weight sharing and window operations.

Fully connected layers Fully connected (FC) layers are the most basic linear operators in a deep neural network. Mathematically, an FC layer conducts matrix multiplication on the input vector $\boldsymbol{x}_k$:

$$\boldsymbol{y}_d = \boldsymbol{W}_{d,k} \times \boldsymbol{x}_k + \boldsymbol{b}_d, \tag{2.1}$$

where $\boldsymbol{y}_d$ is the output, (d, k) are the input and output dimensions and $\boldsymbol{W}_{d,k}$ and $\boldsymbol{b}_d$ indicate the trainable weights within this FC layer. The matrix multiplication principles show that the dimension d of the output vector can differ from that of the input vector, controlled by the weights matrix $\boldsymbol{W}$. Consequently, this allows embedding the inputs to expand hidden features and compress the features to constrained labels.

Convolution layers The convolution operation is introduced to neural networks because FC layers lead to expanding computational cost with high-dimensional inputs, for example, images. In convolution layers, computing is locally focused in much smaller windows with shared kernels, resulting in a huge decline in calculation and trainable parameters (Fig. 2.1). Notice that weight sharing and window operation are essential in computer vision tasks because they restrict the kernels to recurring features (e.g. textures) of images and make very deep neural networks possible. For instance, two-dimensional (2D) convolution layers are predominantly used in image processing networks because they preserve spatial information by operating directly on 2D images with channels. In this process, tensors transition from 1D vectors to 2D maps, embedding features within the third dimension, namely the channels. Concurrently, each convolution kernel generates a new channel in

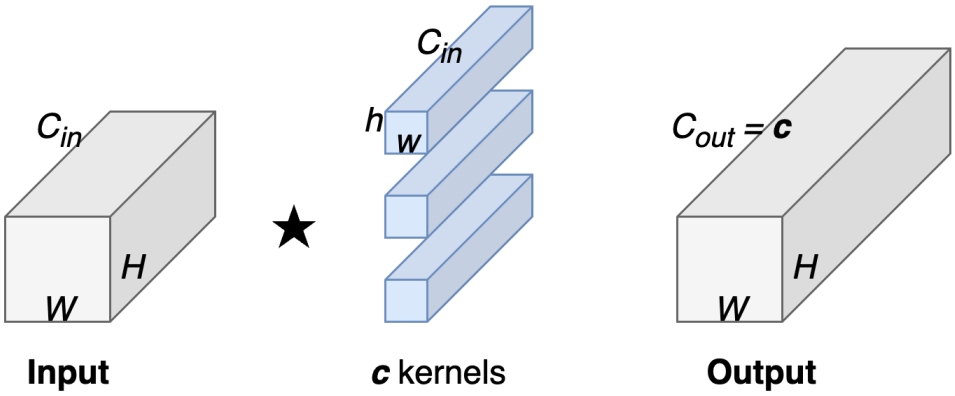


Figure 2.2: An example of convolution layers. H and W indicate the shape of input and output feature maps, while h and w represent the kernel size. Notice that each convolution kernel actually has an extra dimension C_{in} corresponding to the input feature maps for the matrix product. The number of kernels (i.e. layer width) decides the channel of output feature maps.

the output feature maps. Therefore, the number of channels in the output feature maps equates to the number of convolution kernels. In this context, the embedding dimension, also referred to as the width or channel of this layer, is dictated by this layer itself, much like how FC layers regulate the embedding dimensions of 1D features. This process can be illustrated as Fig. 2.2 and mathematically presented as:

$$\boldsymbol{F}^{out}_{H_{out},W_{out},c} = \boldsymbol{F}^{in}_{H_{in},W_{in},C_{in}} \star \boldsymbol{K}_{h,w,c} + \boldsymbol{b}_{H_{out},W_{out},c}, \tag{2.2}$$

where $\boldsymbol{b}$ indicates the bias and $\boldsymbol{K}_{h,w,c}$ indicates the convolution kernels with size $[h, w, c]$. Notice that $\star$ denotes the convolution operation, which modifies the shape and values of one function (e.g. an image or feature maps) by the kernels. The number of kernels, c, is called the convolution layer's width or the channels. H, W indicate the height and width of the feature maps $\boldsymbol{F}$, while the output shape depends on the input shape, padding, dilation [40], stride and kernel size:

$$H/W_{out} = \left\lfloor \frac{H/W_{in} + 2 \times \text{padding} - \text{dilation} \times (\text{kernel} - 1) - 1}{\text{stride}} + 1 \right\rfloor. \tag{2.3}$$

Thus, a convolution layer can remain or adjust the shape of feature maps with corresponding parameter settings, leading to an alternative spatial sub-sampling operation more than polling layers [41]. These parameters also decide the receptive field [42] of each convolution layer and the network with successive expansion along the layers. Fig. 2.3 illustrates an example of 2D convolution, which can be precisely described as:

$$\boldsymbol{F}_{out}(i,j) = \sum \boldsymbol{w}_c \times \boldsymbol{F}_{in}\left[i - \frac{h}{2} : i + \frac{h}{2}; \quad j - \frac{w}{2} : j + \frac{w}{2} \right], \tag{2.4}$$

where $\boldsymbol{w}_c$ is the convolution kernel. It conducts matrix multiplication on a patch of the

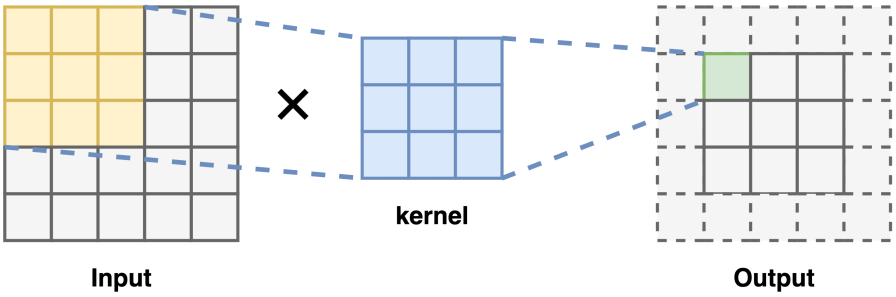


Figure 2.3: 2D convolution operation.

input feature maps.

Nonlinear activation: It is important to apply non-linear activation functions in multi-layer networks because consecutive linear operations are mathematically equivalent to one linear function. Thus, activation functions are often conducted after each linear layer (i.e. FC and convolution layer) in deep neural networks to boost the representation capability. It is reported that the choice of activation functions affects the training efficiency and overall performance [43–45]. In the training of deep neural networks, activation functions introduce non-linearity into the output of a layer and narrow the value of feature maps from $[-\infty, \infty]$ to limited ranges. It also becomes possible to update the trainable weights in backpropagation [46] because the calculation of gradients along with error is well-defined by the non-linearity. In contrast, the derivative of linear functions equals 0, resulting in no gradients for weight updating. Fig. 2.4 illustrates a comparison of common activation functions, while the most widespread ones are the Rectified Linear Unit (ReLU) [47] and its improvements. The ReLU activation is defined as:

$$y = \begin{cases} x & \text{if } x \geq 0; \\ 0 & \text{if } x < 0. \end{cases} \tag{2.5}$$

x and y are scalars.

It looks like a linear function but mathematically differs since its derivative equals 1 with any positive inputs. Meanwhile, it requires less computation than Sigmoid and Tanh and avoids the vanishing gradients caused by near-zero gradients [45]. However, ReLU has no responses to negative inputs because of the 0 gradients and leads to a perpetually inactive state of backpropagation, so-called the **"Dying ReLU"** problem. Thus, Leaky ReLU is

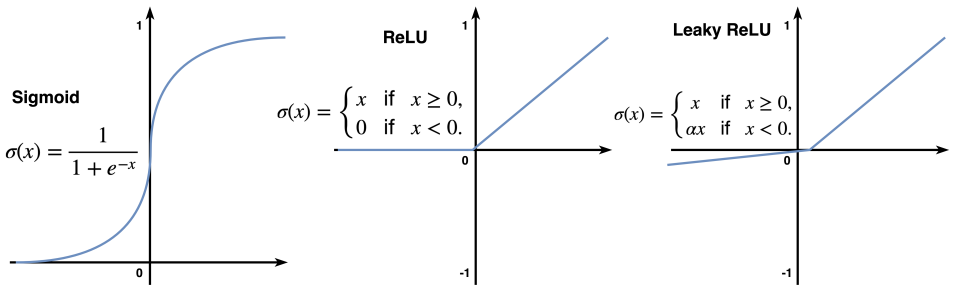


Figure 2.4: Activation functions.

proposed, which adds a closing-to-zero scalar α for the negative inputs of ReLU:

$$y = \begin{cases} x & \text{if } x \geq 0; \\ \alpha x & \text{if } x < 0. \end{cases} \tag{2.6}$$

The small constant α is usually set to 0.01, so it avoids the "Dying ReLU" problem with non-zero derivatives for negative inputs. There are various modifications of ReLU proposed for superior performance, such as Parametric Rectified Linear Unit (PReLU [48]), Randomized Leaky Rectified Linear Unit (RReLU [49]), Exponential Linear Unit (ELU [50]) and Gaussian Error Linear Unit (GELU [51]). Some other activation functions are also widely used, such as *softmax*, Maxout [52], Adaptive Piecewise [53] and Network-in-Network [54].

2.1.2 Neural network training

Gradient descent methods An end-to-end neural network G can be defined as:

$$\boldsymbol{y} = G(\boldsymbol{x}; \boldsymbol{\theta}_G), \tag{2.7}$$

where $\boldsymbol{\theta}$ is a set of trainable parameters, and $\boldsymbol{x}$ and $\boldsymbol{y}$ indicate the input and output of the network. In the data-driven supervised and non-supervised training, the aim is to find a group of parameters that minimise a pre-defined loss function $\mathcal{L}$ with optimisation methods:

$$\hat{\boldsymbol{\theta}_G} = arg \min_{\boldsymbol{\theta}_G} \mathcal{L}(G(\boldsymbol{x}), \bar{\boldsymbol{y}}); (\boldsymbol{x}, \bar{\boldsymbol{y}}) \in \mathbb{D}, \tag{2.8}$$

where $\mathbb{D}$ indicates the training dataset and $\bar{\boldsymbol{y}}$ is one ground truth label. However, there is no closed-form global optimum solution for this equation because of the nonlinearity in neural networks. Thus, backpropagation [46] is introduced to perform optimisation of deep neural networks. It updates the trainable parameters with gradient descent algorithms

[55] that calculate gradients depending on the loss of a batch of data:

$$\frac{\nabla \mathcal{L}}{\nabla \boldsymbol{\theta}} = \frac{1}{N} \sum^{N} \frac{\nabla \mathcal{L}(G_{\boldsymbol{\theta}}(\boldsymbol{x}), \bar{\boldsymbol{y}})}{\nabla \boldsymbol{\theta}}, \tag{2.9}$$

where N is the batch size. As one deep neural network consists of a stack of layers, the partial gradient of each layer can be computed in the opposite direction of data forward by the chain rule. In each training step, the parameters can be updated as:

$$\boldsymbol{\theta} = \boldsymbol{\theta} - \iota \frac{\nabla \mathcal{L}}{\nabla \boldsymbol{\theta}}, \tag{2.10}$$

where ι is the learning rate. SGD (Stochastic Gradient Descent [56]) is one of the most popular batch gradient descent methods because it mathematically guarantees global minimums without expensive computation. However, it works poorly for ravines where one dimension curves much more steeply than others [57], so optimisation methods with momentum [58] are implemented, such as NAG (Nesterov Accelerated Gradient [59]). On the other hand, the convergence of SGD tends to local minimums for non-convex surfaces but not the global minimums without a proper learning rate. Thus, several advanced methods with adaptive learning rates are proposed, such as Adadelta [60] and RMSprop[1]. Additionally, Adam (Adaptive Moment Estimation [61]) and Nadam (Nesterov-accelerated Adam [62]) combine momentum with adaptive learning rates. As an essential foundation of deep neural networks, gradient descent algorithms remain a hot research topic, and the discussion of which optimiser to use will always continue.

Training strategies Building a well-performing deep neural network highly depends on human expertise and time-consuming trial and error. To create high-quality networks, even experts need substantial resources and unproductive tuning of the model, including data preparation, network architecture, optimisation method and hyperparameters. Although AutoML (i.e. automated machine learning) techniques are processed and quickly developed on object detection, text processing, and image classification tasks [63], such pipelines rarely apply on low-level image processing tasks. They are limited by the requirement of onerous computational cost, the lack of objective evaluation and high coupling on the training dataset [64]. Thus, developing high-performance deep neural networks for super-resolution tasks still jointly relies on the prior knowledge of state-of-the-art works and manual tuning. Here I introduce common strategies, such as initialisation, normalisation and data preparation.

A deep neural network may not converge without careful initialisation because of gradients exploding or vanishing [65]. As a vital role of efficient backpropagation, weight initialisation

[1] `http://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf`

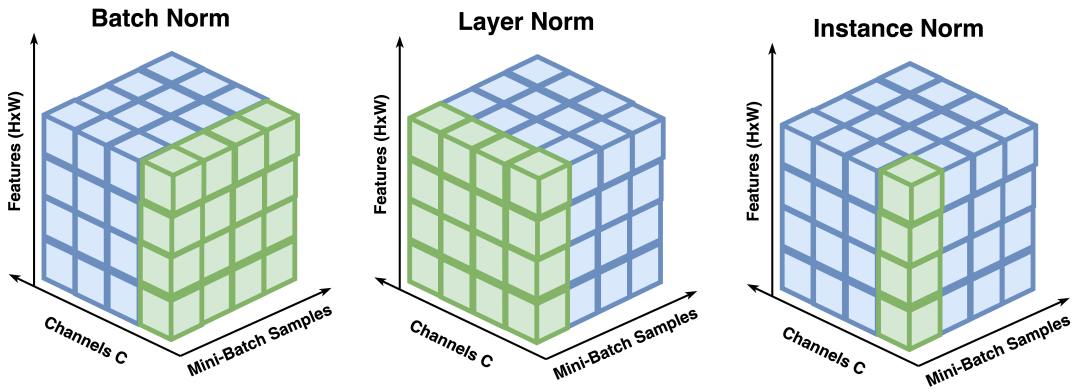


Figure 2.5: Normalisation methods in deep neural networks. Blue cubes indicate flatten feature maps for clear illustration, while green cubes represent where the normalisation works: a mini-batch, a layer, and an instance.

methods mainly includes two directions: with and without pre-training [66]. Initialising a network with pre-training means obtaining a start point from unsupervised or correlated tasks, so the learned meaningful representations may lead to better accuracy and faster convergence. This idea is close to transfer learning [67], in which some network layers are frozen, and only partial weights are fine-tuned. For example, powerful well-trained models, such as MAE (masked autoencoders [68]), successfully work on medical image processing tasks [69]. On the other hand, initialisation without pre-training consists of three main categories: data-driven, random and hybrid. In computer vision tasks, random initialisation methods are widely used because they meet the randomness requirement in gradient descent with proper-designed distributions, defined with network characteristics, such as activation function and layer width [70].

Normalisation methods [71] are also essential in training acceleration and stabilisation. The normalisation operation can conduct on various dimensions (Fig. 2.5), such as batch [72], layer [73], instance [74], weights [75], and group [76]. It transfers the input data $\mathbb{D} = \{\boldsymbol{x}_i\}_{i=1}^{N}$ to a new distribution $\hat{\mathbb{D}} = \{\hat{\boldsymbol{x}}_i\}_{i=1}^{N}$ with certain statistical properties. By removing the magnitude difference between different features, it benefits the training process. Interestingly, the learnable parameters in adaptive normalisation methods [77, 78] can represent the style of an image, which leads to high-fidelity image generation [79–81] and editing [82].

In addition to the above aspects, dropout [83] and data augmentation [84, 85] are also moderately applied in super-resolution tasks [86]. Dropout operation prevents overfitting with improved generalisation capability. It keeps each neuron (e.g. a convolution kernel) active with a probability p and inactive for the rest during training and behaves as usual in the testing stage. It mathematically equals a simple form of ensemble models, fusing numerous randomly sampled sub-networks. Data augmentation is a trick to increase the

diversity of the training dataset by applying random transformations. In super-resolution tasks, rotation and flip operations are used for training and inference.

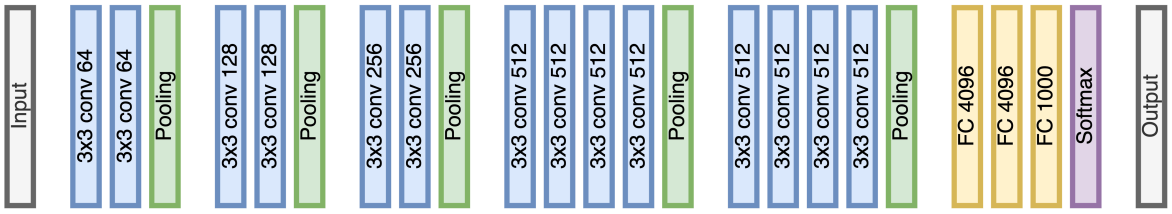


Figure 2.6: The VGG-19 [87] framework. It consists of successive convolution layers with small kernel size and pooling layers for feature sub-sampling. Fully-connected layers at the end for classification.

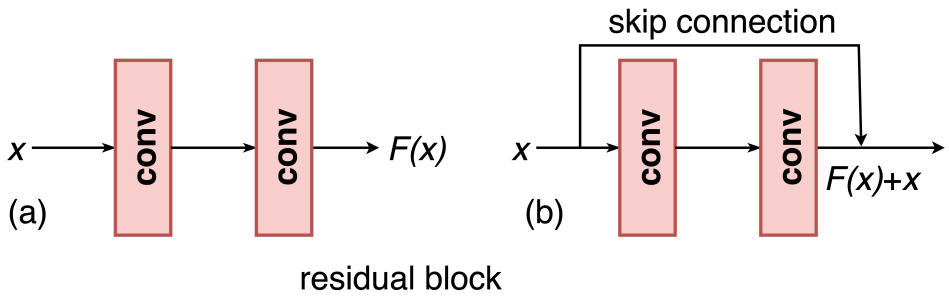


Figure 2.7: Comparison of (a) plain layer concatenation and (b) residual block. It avoids gradient vanishing with the skip connection and residual learning.

2.1.3 Network architecture

In this section, I will introduce several widespread deep neural network blocks and architectures that form the basis of many of my experiments.

VGG A landmark in the prosperity of deep learning is the dominant performance in large-scale computer vision tasks. AlexNet [88] first achieves dramatic improvements (10.8% lower top-5 error rate than previous methods) on the ImageNet large-scale visual recognition challenge (ILSVRC [89]) in 2012 by applying network initialisation, ReLU activation and dropout techniques to CNN. Following this direction, VGG [87] was proposed in 2014 with the superior performance of objection detection with the pioneering network design principles: smaller filters and more layers (Fig. 2.6). It claims that a stack of three 3×3 convolution layers has an equal receptive field to a 7×7 convolution layer (e.g. in AlexNet) but with much fewer parameters and less computation cost.

Residual networks The plain layer concatenation succeeds in AlexNet and VGG, but it results in a performance decline in very deep networks (e.g. more than 100 layers). ResNet [90] successfully solves this problem by proposing the residual block (Fig. 2.7), in which the identity skip connection passes the shallow features forward and the deep gradients

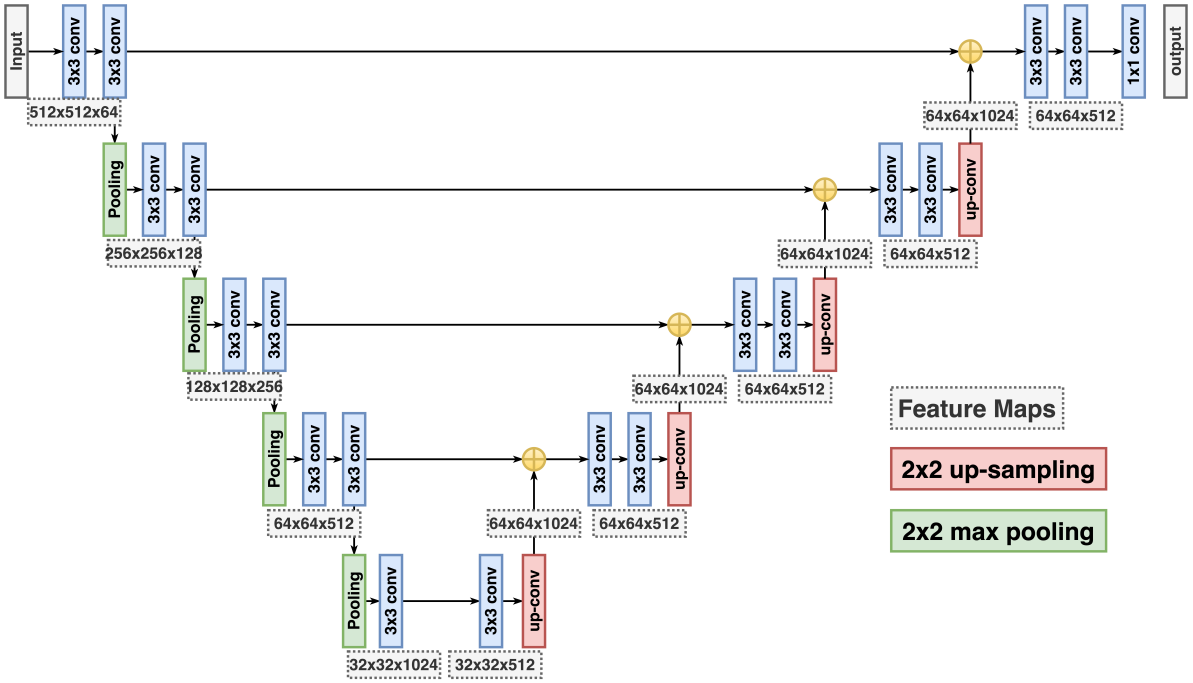


Figure 2.8: U-Net.

backwards directly in supplement to the main path. Since then, residual block variants, such as dense block [91], have become the fundamental unit of deep neural networks.

U-Net U-Net is primarily designed for medical image segmentation [92]. It consists of an encoder and a decoder that contract and expand the feature maps, respectively (Fig. 2.8). Each encoder block consists of several convolution layers and a pooling layer for $\times 2$ down-sampling. In contrast, each decoder block consists of convolution layers and one up-convolution layer for $\times 2$ up-sampling. Then, the multi-scale feature maps of the encoder pass and concatenate with corresponding feature maps in the decoder by skip connections. The U-shape framework ensures pixel-to-pixel outputs, which are different from downstream networks (e.g. AlexNet and VGG) with outputs of labels. Meanwhile, the encoder/decoder blocks and the down-/up-sample operations can easily change to alternatives, so U-Net has become one of the most popular frameworks in semantic segmentation [93–96] and image restoration [97–99] tasks, especially for medical images [100].

Generative adversarial networks Generative adversarial networks are probably the most exciting idea in machine learning in the past decade. Since 2014 [101], it has revolutionised a broad range of computer vision, natural language processing and signal processing tasks [102, 103]. Medical image analysis is also benefited [104, 105]. GANs consist of a generator and a discriminator, jointly trained by playing a game (Fig. 2.9). The generator aims to approximate a mapping from one distribution to a target distribution,

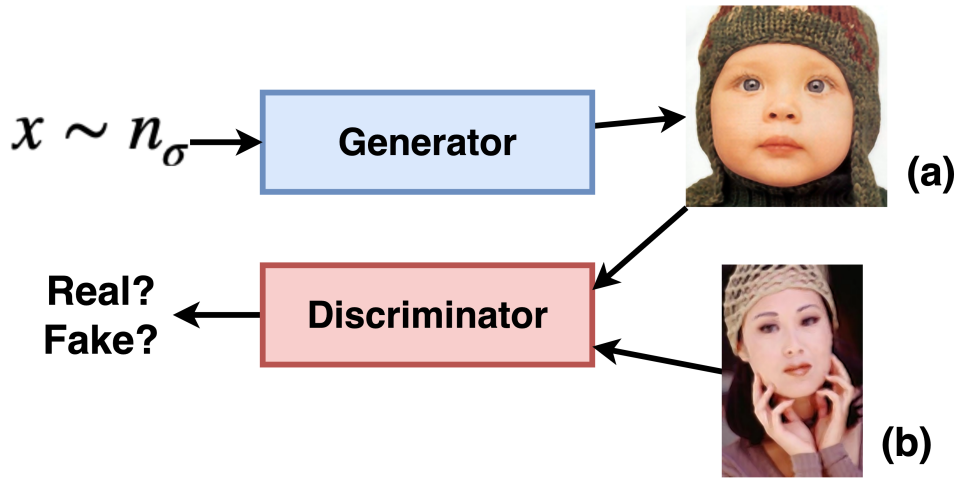


Figure 2.9: Generative adversarial networks: (a) fake image generated from noise; (b) real image from the training dataset.

while the discriminator aims to distinguish the estimated distribution from the real one. Take high-fidelity image generation as an example. The generator produces as authentic as possible images to fool the discriminator from a noise input (e.g. Gaussian noise $\boldsymbol{n}_\sigma$). Meanwhile, the discriminator learns to differentiate the generated images from real ones. Ideally when the optimisation of both networks terminates, the discriminator achieves the maximum classification accuracy while the generator maximally confuses the discriminator D:

$$\hat{\boldsymbol{\theta}_G}, \hat{\boldsymbol{\theta}_D} = arg \min_{\boldsymbol{\theta}_G} \max_{\boldsymbol{\theta}_D} \mathbb{E}_{p(\bar{\boldsymbol{y}})} \log D(\bar{\boldsymbol{y}}) + \mathbb{E}_{p(\boldsymbol{x})}(1 - \log D(G(\boldsymbol{x}))); \ \boldsymbol{x} \sim \boldsymbol{n}_\sigma, \bar{\boldsymbol{y}} \in \mathbb{Y}, \quad (2.11)$$

where $\mathbb{E}$ is expectation.

The fantastic advantage of GANs is that no training pairs and task-specific loss functions are required. The discriminator takes the place of the generator's loss function instead. However, the vanilla GAN struggles with unstable training, mode collapse and non-convergence. Thus, advanced works are proposed with improvements on representation [106, 107], loss distance [22, 108], training skills [109], network architectures [110, 111] etc. Meanwhile, GANs become a conditional model when specific data distributions replace the noise inputs [112], such as in image translation [113], super-resolution [114], and text-to-image synthesis [115].

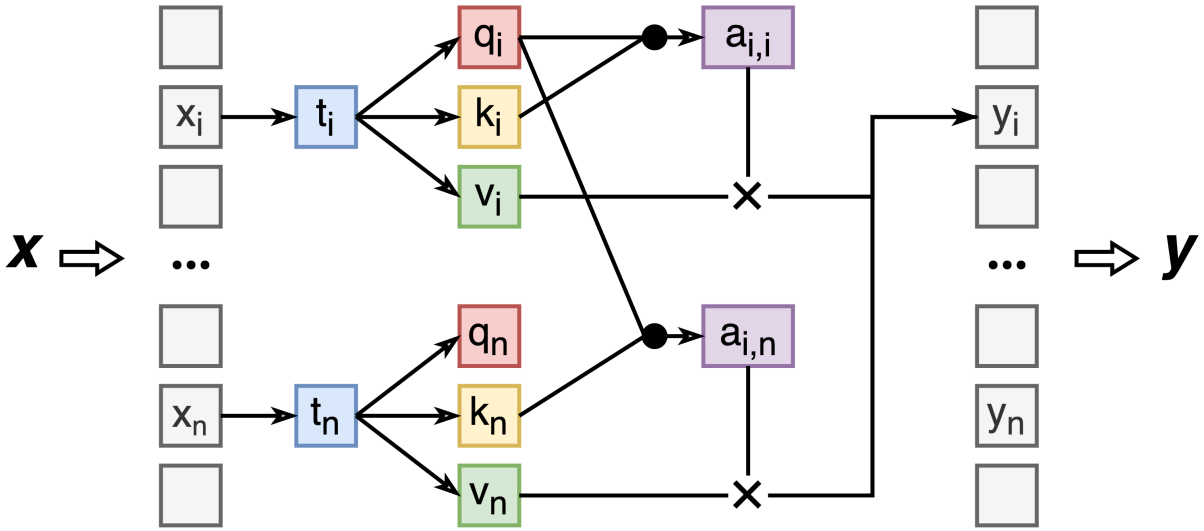


Figure 2.10: Self-attention module in transformers.

Vision transformers Transformers are first applied for natural language processing [116] and then introduced to vision tasks [117, 118] by embedding image and feature maps to tokens. Self-attention (Fig. 2.10) is the key component of transformers, which dynamically predicts weights relying on the input. After feature embedding (e.g. localisation or adaptive embedding), each token $\boldsymbol{t}_i$ is transformed into the query, key and value vectors $\boldsymbol{q}_i, \boldsymbol{k}_i, \boldsymbol{v}_i$ with the same dimension d:

$$\begin{aligned} \boldsymbol{t}_i &= \Psi(\boldsymbol{x}_i) \\ \boldsymbol{q}_i, \boldsymbol{k}_i, \boldsymbol{v}_i, &= \boldsymbol{w}_q \boldsymbol{t}_i, \boldsymbol{w}_k \boldsymbol{t}_i, \boldsymbol{w}_v \boldsymbol{t}_i, \end{aligned} \tag{2.12}$$

where Ψ is the embedding function and $\boldsymbol{w}_q, \boldsymbol{w}_k, \boldsymbol{w}_v$ are $d \times d$ matrices. Then, the attention vector $\boldsymbol{a}_{i,j}$ between two embedded tokens are calculated and normalised as:

$$\begin{aligned} \bar{\boldsymbol{a}}_{i,j} &= \boldsymbol{q}_i \cdot \boldsymbol{k}_i / \sqrt{d}, \\ \boldsymbol{a}_{i,j} &= exp(\bar{\boldsymbol{a}}_{i,j}) / \sum^{j} \bar{\boldsymbol{a}}_{i,j}, \end{aligned} \tag{2.13}$$

where *softmax* function translates the attention scores $\bar{\boldsymbol{a}}_{i,j}$ into probabilities $\boldsymbol{a}_{i,j}$. Finally, the output $\boldsymbol{y}_i$ is obtained by:

$$\boldsymbol{y}_i = \sum^{j} \boldsymbol{a}_{i,j} \cdot \boldsymbol{v}_i. \tag{2.14}$$

The above process of self-attention can be unified as:

$$\mathcal{A}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = softmax(\frac{\boldsymbol{Q} \cdot \boldsymbol{K}^\top}{\sqrt{d}}) \cdot \boldsymbol{V}, \tag{2.15}$$

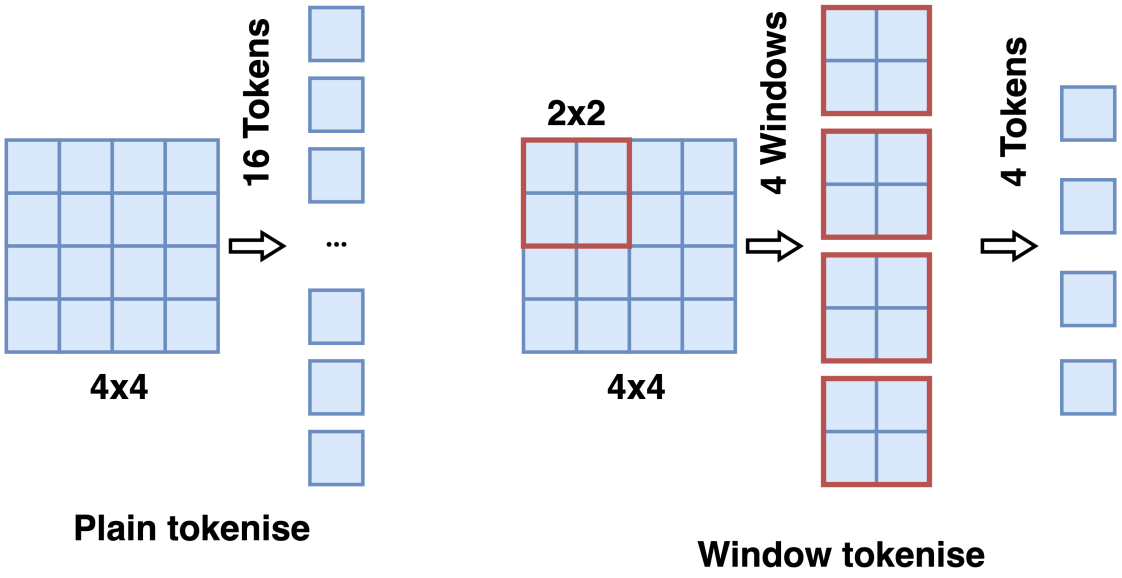


Figure 2.11: Self-attention in windows [119]. Compared to global attention with plain tokens (left), it conducts self-attention locally in each window with sharing weights, leading to a quadratic decline in computational cost.

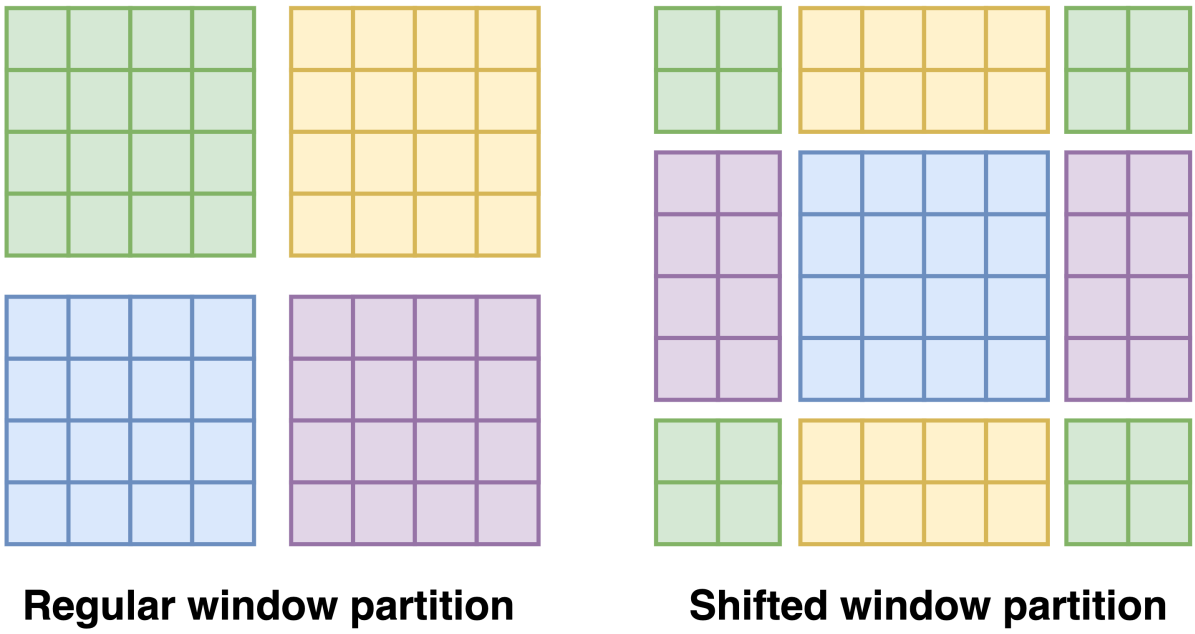


Figure 2.12: Shifted window partition in swin transformer [119]. Colours indicate the window partition w/wo shift operation. In practice, regular window partition (left) and shifted window partition (right) process alternately in successive layers for global information flow.

where $\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}$ are query, key and value matrices of all input tokens and $\mathcal{A}$ indicates the attention operation.

In transformers, the self-attention layer works in the encoder and the decoder. Notice that the key and value matrixes $\boldsymbol{K}$ and $\boldsymbol{V}$ are only calculated in the encoder and directly passed to the decoder. Additionally, multi-head attention boosts performance by giving tokens to various attention layers (i.e. heads) and fusing the final outputs. In recent years, advanced transformers have been proposed for computer vision tasks. ViT [120] first embeds image to 16×16 tokens and performs better than CNNs on image recognition tasks. The performance is further improved by DERT [121] and its variants with novel techniques such as 2D position encoding, deformable attention [122] and adaptive clustering [123]. Transformers also apply to high-quality image synthesis with GANs [124, 125] and from text captions [126]. Specifically for low-level image restoration tasks, IPT [127] involves CNN-based shallow feature embedding before self-attention layers and transfer learning of pre-trained model on ImageNet [128]. Furthermore, Swin transformer [119] dramatically improves the performance and efficiency with shifted window operation, drawing on localisation operation and weight sharing of CNNs. In contrast to previous vision transformers, the swin transformer splits the feature maps into small windows. It only applies self-attention to tokens in each window (Fig. 2.11), resulting in a quadratic decline of computation. On the other hand, the shifted window approach (Fig. 2.12) ensures connections of local information between the windows, leading to a global receptive field in successive layers.

2.2 Single image super-resolution

In this section, I introduce the basic concepts of single-image super-resolution, especially the current methods based on deep neural networks. In addition to a quick review of the super-resolution question definition, SR image evaluation and objective loss functions, widespread SISR networks are illustrated with an emphasis on the up-sampling modules and network frameworks.

2.2.1 The ill-posed super-resolution problem

Super-resolution techniques are proposed for high-resolution displays with low-resolution observation devices [129]. It aims to reconstruct a higher-resolution image from one or a sequence of observed low-resolution images. In practice, the observation is limited by various interferences, leading to the degradation of warping, blurring, down-sampling and noise in the acquired LR images. However, obtaining the correct model for each degradation factor for every LR image is hard, so super-resolution turns into an ill-posed inverse problem. One HR image has several corresponding LR images, while one LR image may be degraded from various potential HR images. Additionally, it is challenging to acquire multiple low-resolution images in practice. Thus, single-image super-resolution methods are proposed.

Single image super-resolution aims to restore a high-resolution image $\boldsymbol{I}_{hr} \in \mathbb{H}$ from one low-resolution observation $\boldsymbol{I}_{lr} \in \mathbb{L}$ of the same object, where $\mathbb{L}, \mathbb{H}$ are high and low dimensional spaces. Generally in the real world, the LR image is modelled as [8]:

$$\boldsymbol{I}_{lr} = (\boldsymbol{I}_{hr} \star \boldsymbol{\delta}) \downarrow_s + \boldsymbol{n}, \tag{2.16}$$

where $\star\boldsymbol{\delta}$ denotes a simplified degradation during the image-capturing process and $\downarrow_s$ indicates the down-sampling with scale s. When deep neural networks apply to the SISR problem, they approximate the inverse mapping of this Equation 2.16 to recover a super-resolved image $\boldsymbol{I}_{sr} \in \mathbb{H}$ from $\boldsymbol{I}_{lr}$ [9]:

$$\boldsymbol{I}_{sr} = G(\boldsymbol{I}_{lr}, s; \boldsymbol{\theta}_G), \tag{2.17}$$

where G is an SR image generator and s is the magnification scale. In each step of training, errors between the approximation $\boldsymbol{I}_{sr}$ and the HR ground truth $\boldsymbol{I}_{hr}$ are measured by a well-designed loss function $\mathcal{L}_{SR}$, and passed to the whole network in backpropagation

for gradients calculation and weights update:

$$\hat{\boldsymbol{\theta}_G} = arg \min_{\boldsymbol{\theta}_G} \mathcal{L}_{SR}(G(\boldsymbol{I}_{lr}), \boldsymbol{I}_{hr}), . \tag{2.18}$$

where $\{\boldsymbol{I}_{lr}, \boldsymbol{I}_{hr}\} \in \{\mathbb{L}, \mathbb{H}\}$ is a corresponding training pair.

2.2.2 Loss functions for super-resolution

Here I introduce the most commonly used loss functions in training SISR deep neural networks.

Pixel-wise L1 and L2 losses Since the nature of SISR is to predict the correct value of each pixel in the super-resolved images, loss functions that measure pixel-wise errors are always the fundamental term. L1 and L2 loss represent the pixel-wise mean absolute error (MAE) and mean square error (MSE), respectively:

$$\mathcal{L}_1(\boldsymbol{I}_{sr}, \boldsymbol{I}_{hr}) = \frac{1}{H * W} \sum_{(i,j) \in I} \left\| \boldsymbol{I}_{hr}[i,j] - \boldsymbol{I}_{sr}[i,j] \right\|, \tag{2.19}$$

$$\mathcal{L}_2(\boldsymbol{I}_{sr}, \boldsymbol{I}_{hr}) = \frac{1}{H * W} \sum_{(i,j) \in I} \left\| \boldsymbol{I}_{hr}[i,j] - \boldsymbol{I}_{sr}[i,j] \right\|^2, \tag{2.20}$$

where H and W are the height and width of the images, and $\boldsymbol{I}[i,j]$ denotes a pixel. Early researches on SISR, such as SRCNN [130], VDSR [131], and SRGAN [114], prefer L2 loss because it straightforwardly connects to the most popular evaluation metric PSNR. However, it is sensitive to outliers because it squares the differences. It tends to produce more stable solutions, leading to over-fitting and over-smoothed results. In contrast, L1 loss is more robust to outliers because of its linear errors. It can preserve edges and achieve more visually appealing results. Thus, L1 loss becomes more widely used in the following works [132–138] for improved performance. Notice that the pixel-wise loss functions do not present the global structures, so they cannot benefit the perceptual quality and realistic textures generation, resulting in unreal artefacts and fewer details.

Perceptual loss The above pixel-wise losses ensure good reconstruction fidelity of SR images, while perceptual losses are introduced for superior perceptual quality and realistic SR images. Instead of manually defining a distance between human experiences, the perceptual losses rely on the prior knowledge of pre-trained networks on related computer vision tasks because the hidden layers represent high-level image features. Take the widespread VGG-based perceptual loss [139] as an example. Instead of calculating the pixel-wise errors, it considers the distance of the distributions of generated SR images and

ground truth images in the feature domain. In practice, the images are first converted as features maps by a pre-trained VGG-19 [87] model $\mathcal{V}$, and then the L2 distance between the two distributions is calculated as:

$$\mathcal{L}_{perc}(\boldsymbol{I}_{sr}, \boldsymbol{I}_{hr}) = \mathbb{E}(\|\mathcal{V}_l(\boldsymbol{I}_{hr}) - \mathcal{V}_l(\boldsymbol{I}_{sr})\|^2), \tag{2.21}$$

where l denotes the specific layer to generate feature maps. The way of training $\mathcal{V}$ and the choice of l have inferences on the final performance. Following the success of SRGAN [114], most SISR works [140, 141] use the VGG-19 model, which is pre-trained with ImageNet [128] on classification tasks and use the feature maps of deeper layers to achieve more semantic information. Meanwhile, advanced perceptual losses are implemented for more realistic-looking images in various super-resolution tasks, such as perceptual loss of multiple feature maps [142], in the frequency domain [143] and in specific tasks [144]. In ESRGAN [134], feature maps before activation of shallow layers of a VGG model trained on material detection [145] are used for more texture information. Semantic information on segmentation labels also applies to single image super-resolution for boundary enhancement in SR images [146].

Adversarial loss As mentioned in Section 2.1.3, generative adversarial networks benefit high-fidelity image generation and editing. They are also widely used in single image super-resolution tasks for perceptually more realistic results [147]. As a specific task of conditional GANs, the generators take the low-resolution image as input, instead of the noise vector in Equation 2.11. During training, the discriminator considers the generated super-resolved images as fake images and the ground truth high-resolution images as real images. The vanilla adversarial loss function is defined as [101]:

$$\mathcal{L}_{GAN} = -\mathbb{E}_{\boldsymbol{I}_{hr}}\left[\log D(\boldsymbol{I}_{hr})\right] - \mathbb{E}_{\boldsymbol{I}_{lr}}\left[\log(1 - D(G(\boldsymbol{I}_{lr})))\right], \tag{2.22}$$

where $\mathbb{E}$ is the expectation of the whole dataset. This vanilla adversarial loss first applies to SISR tasks in SRGAN [114], where the same discriminator to DCGAN [110] is used. It has successfully achieved photo-realistic natural images with $\times 2$ and $\times 4$ magnifications but requires a time-consuming warm-up of the generator to stabilise the training of GANs. Advanced researches on GANs in super-resolution are mainly about applying novel adversarial loss functions, such as relativistic GAN [106], Wasserstein GAN variants [22, 108] and cycle GAN [148].

Additional task-specific regularisation terms and losses also work on super-resolution tasks. Total variation loss [149] can ensure smoothness across sharp edges in the generated SR images [150]. Texture similarity loss [151] improves the perceptual quality with a more

realistic reconstruction of image style [152]. Rank-content loss [153] learns the behaviour of perceptual metrics and combines the strengths of various SR methods for better visually pleasing results [154]. In practice, loss functions combine to improve reconstruction accuracy, perceptual quality and task-driven performance simultaneously. However, selecting appropriate loss components and weights seriously affects the final performance, relying on human experiences in trial and error [155].

2.2.3 Evaluation of SR images

Generally, image quality assessment metrics of SR images include subjective and objective methods. The former can ideally represent human perception but rely on time-consuming manual scoring with inter-/inner- variances. In contrast, objective methods are easy to compute and fair for comparison but usually focus on only one aspect of image quality evaluation. Thus, various metrics are used to comprehensively evaluate SR images, such as the reconstruction fidelity [156] and perceptual quality [157]. This section introduces the most popular objective metrics for reconstructed image quality evaluation.

Reconstruction fidelity Peak signal-to-noise ratio (PSNR) is the most widespread evaluation criteria for image restoration tasks (e.g. reconstruction, super-resolution and denoising). It involves the data range to measure the pixel-level mean squared error (MSE):

$$\mathrm{PSNR}(\boldsymbol{I}_{rec}, \boldsymbol{I}_{gt}) = 10 \cdot \log_{10}\left(\frac{L^2}{\frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{I}_{rec}(i) - \boldsymbol{I}_{gt}(i))^2}\right), \tag{2.23}$$

where L denotes the data range (generally $L = 1.0$ in medical image reconstruction tasks), and N is the number of all pixels in $\boldsymbol{I}_{rec}$ and $\boldsymbol{I}_{gt}$. PSNR represents the pixel-wise accuracy of the results without the impacts of image format variations.

Additionally, considering the importance of image structural information, such as luminance, contrast and structures, the structural similarity (SSIM) is proposed as [158]:

$$\mathrm{SSIM}(x, y) = \frac{2\mu_x\mu_y + \kappa_1}{\mu_x^2 + \mu_y^2 + \kappa_1} \cdot \frac{\sigma_{xy} + \kappa_2}{\sigma_x^2 + \sigma_y^2 + \kappa_2}, \tag{2.24}$$

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 is a typical top-down framework of full-reference image quality assessment (FR-IQA), which assumes that contrast and structural distortions are essential to the human visual system (HVS). Compared with PSNR, SSIM reflects global structural information. However, native SSIM and multi-scale SSIM [159] may struggle with noisy and distorted images [160], on which PSNR works well.

Perceptual quality assessment With the rapid development of generation models (e.g. GANs and diffusion models [161–164]), it becomes increasingly important to evaluate the perceptual quality of images in computer vision tasks [157], which represents how realistic an image looks. The most reliable perceptual quality assessment is the mean opinion score (MOS), which asks experienced raters to score the reconstructed images with criteria on characteristics (e.g. sharpness, artefacts, contrast, exposure) and average the scores. For example, in the domain-specific medical image restoration tasks [165, 166], experienced radiologists grade all images from 0 to 4 according to the quality (i.e., non-diagnostic, poor, fair, good, and excellent). Sometimes, the rater may also mark the low perceptual quality features such as low signal-to-noise ratio and motion artefacts. Although the MOS is faithful, it still has limitations such as inter-/inner-raters bias and variance of rating criterion and time consumption.

Thus, objective metrics are proposed for image perceptual quality evaluation, including learning-based and feature-based methods. The former ones learn the behaviour of the human visual system on specific IQA datasets. For example, DeepQA [167] trains a CNN for full-reference image quality assessment, taking both the distorted image and error map as input. DeepIQ [168], MEON [169] and NIMA [170] learn the blind opinion for no-reference IQA tasks. Additionally, RankIQA [171] learns from ranked images of known degradation and outperforms no-reference and full-reference SOTA IQA methods in specific applications. Notice that most learning-based methods are proposed for natural images and fine-tuned with specific IQA datasets and known distortions. Meanwhile, they mainly focus on the image quality assessment of one image w/wo a reference. In contrast, feature-based methods concentrate on the distribution of generated images. For example, inception score [172] relies on the pre-trained Inception [173] image classification network and evaluates images' variety and perceptual quality. However, it closely corresponds to the training dataset, limiting the application of out-of-domain tasks. Instead, the similarity of distributions in the feature domain of the target and generated images seems more reliable, such as LPIPS [174] and FID [175]. For instance, Frechet Inception Distance (FID) [175] assumes the distribution of images in the feature domain is a multidimensional Gaussian distribution (i.e. $\boldsymbol{\mathcal{N}}(\mu, \Sigma)$). Thus, the distance between the feature distributions can represent the perceptual similarity of two image sets, such as restored images $\mathbb{I}_{res}$ and ground truth images $\mathbb{I}_{gt}$. In practice, each image group is converted to a distribution of 2048 features in the latent space of a pre-trained image classification model Inception-V3 [173]. Then, the FID between these two distributions is:

$$FID(\mathbb{I}_{res}, \mathbb{I}_{gt}) = \|\mu_{gt} - \mu_{res}\|^2 + Tr(\Sigma_{gt} + \Sigma_{res} - 2(\Sigma_{gt}\Sigma_{res})^{1/2}), \tag{2.25}$$

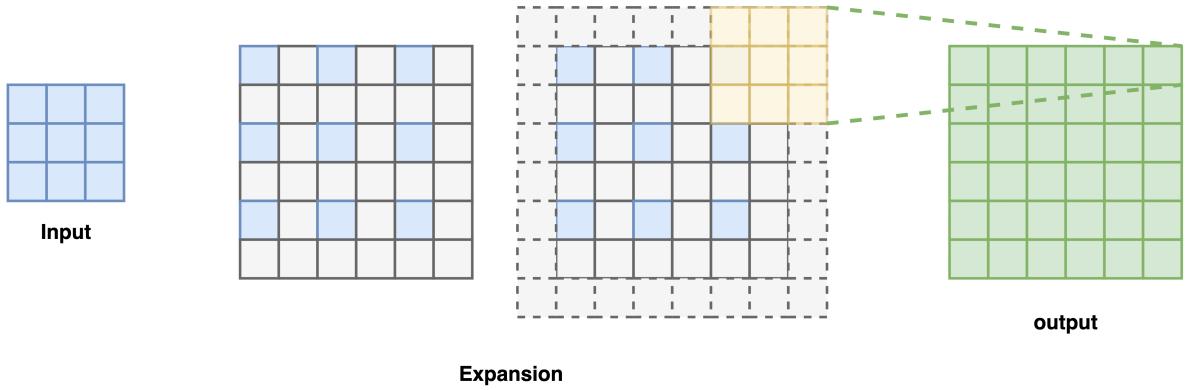


Figure 2.13: De-convolution layer for feature map $\times 2$ upsampling. It works for integer magnification scales with expansion and convolution operations. The low-resolution input is first expanded with additional zeros inserted between the pixels, and then a convolution operation applies on the expansion with zero-padding for the output feature maps with targeted shape.

where $\mathbb{I}$ indicates a group of images and Tr is the trace calculation (i.e. the sum of elements along the main diagonal of the square matrix). FID becomes popular for perceptual quality assessment in image generation tasks [81, 126, 161] because it is fully automatic and the features extracted from Inception-V3 are close to real-world object classifications which tend to mimic human perception similarity in images.

Other full-reference and non-reference IQA methods are also designed on natural scene statistics [176], prior knowledge of human judgements and distortions [177], spatially active regions [178] and a combination of local and global frequency and spatial features [179]. However, the image quality assessment of generated and enhanced images is still an open problem [156], especially on medical images. Although the diagnostic accuracy is not equal to image quality [19], the performance of SR images in downstream medical image analysis tasks occasionally represents the SR image quality [20].

The above metrics only focus on image quality assessment because it is crucial to explore the best super-resolution results regardless of computational cost [180]. Generally, larger models lead to advanced performance with more powerful representation capabilities of more parameters and layers. However, they also require higher memory consumption and longer runtime during inference, which limit their applicability on mobile devices and real-time tasks. Thus, researchers also discuss the model efficiency of SR networks with different metrics, such as runtime, number of parameters, computation complexity (e.g. number of multi-add calculations) and memory consumption [181]. The primary focus of this thesis is on enhancing the quality of SR images. Simultaneously, I strive to strike a balance between image quality and model efficiency.

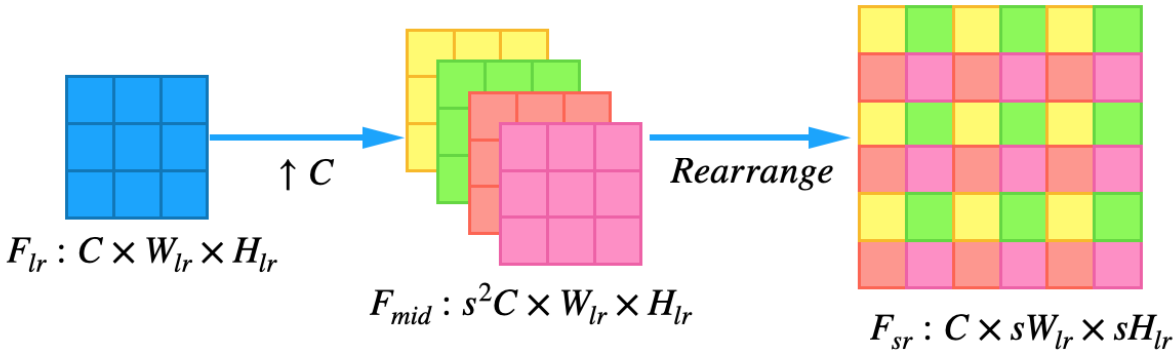


Figure 2.14: The Sub-Pixel layer [189] increases the width and height of the feature maps by two steps. First, it keeps the dimension of feature maps but increases the number of channels (i.e. C) by convolution operation. Second, it aggregates the low-resolution feature maps and builds the high-resolution SR feature maps by reshaping them. F_* represents the feature maps of each step, s is the magnification scale and $[W, H]$ are the width and height of the feature maps.

2.2.4 Feature map up-sampling

Before introducing the widespread network frameworks for super-resolution, I illustrate three popular up-scale modules in SR networks for feature map magnification. In super-resolution, it is the key component to increase the resolution of input LR images or feature maps. Although interpolation-based methods (e.g. bilinear and bicubic interpolation) and up-pooling operations are widely used in the decoders of U-Net architectures for segmentation and restoration tasks, they result in resolution increase with no representational capacity of the upsampling process. Thus, they lead to blocky and blurry results in super-resolution networks. In contrast, these three learning-based up-sampling modules can achieve superior SR results with the high-fidelity representation of magnifications learned from the training dataset.

Deconvolution layer Deconvolution layer [182] (also called the transposed convolution layer) is widely used in image super-resolution tasks [183–187]. However, it occasionally introduces chequerboard artefacts within the restored high-resolution images, resulting in poor SR performance [188]. The deconvolution layer is a converse of a convolution layer (Fig. 2.13). It predicts the potential input of high-resolution feature maps based on the available low-resolution feature maps. Like down-sampling with convolution layers, deconvolution layers implement the feature map up-sampling with proper parameters of *stride*, *padding*, *dilation* and *kernelsize*. Take $\times 2$ magnification as an example. The LR feature maps F_{lr} is first expanded with additional zeros inserted between the pixels, and then a convolution operation applies on the expansion with zero-padding.

Sub-pixel layer Instead of increasing the width and height of feature maps directly, the Sub-Pixel layer [189] introduces additional information by expanding the channels (Fig.

2.14). As mentioned in Section 2.1.1, suitable convolution layers can remain the width and height of the feature maps but increase the number of channels by s^2 times (where s is the aimed magnification scale). Then, it aggregates low-resolution feature maps and builds high-resolution feature maps by reshaping them. Compared to the deconvolution layer [182], the convolution operation within the sub-pixel layer applies on feature maps with low resolution, so it is more efficient. Meanwhile, it can avoid potential checkerboard artefacts caused by the overlapped window operations in deconvolution layers. Mathematically, this two-step process of increasing the dimensions of the LR feature maps $\boldsymbol{F}_{lr}$ of shape $[C \times W_{lr} \times H_{lr}]$ to feature maps $\boldsymbol{F}_{sr}$ of shape $[C \times sW_{lr} \times sH_{lr}]$ can be notated as:

$$\boldsymbol{f}_i = \boldsymbol{F}_{lr} \star \boldsymbol{k}_i; \quad \text{for } \boldsymbol{f}_i \in \boldsymbol{F}_{mid}[1 \cdots s^2], \tag{2.26}$$

where $\star\boldsymbol{\kappa}_i$ is a convolution operation with corresponding padding to remain the shape of the i-th feature map $\boldsymbol{f}_i$, while $\boldsymbol{F}_{mid}[1 \cdots s^2]$ denotes the hidden feature maps with expand channels. Then the feature maps $\boldsymbol{F}_{mid}$ of shape $[s^2C \times W_{lr} \times H_{lr}]$ are rearranged to the aimed shape by a periodic shuffling operation:

$$\boldsymbol{F}_{sr}(c, x, y) = \boldsymbol{F}_{mid}(c + C \cdot mod(x, s) + C \cdot s \cdot mod(y, s), \left\lfloor \frac{x}{s} \right\rfloor, \left\lfloor \frac{y}{s} \right\rfloor), \tag{2.27}$$

where $[c, x, y]$ are the location indexes. The sub-pixel layer can generate perceptually realistic textures since its wide respective field involves more contextual information in the up-sampling process. Thus, it becomes the most popular feature map up-sampling module in deep neural networks [133, 136, 138, 190, 191], although it only works with integer magnification scales. However, unreal artefacts still exist in some cases because of the uneven distribution of the respective field and random initialisation. Thus, advanced up-sampling operations are proposed to further improve the smoothness and perceptual quality of generated images without false artefacts [192, 193].

Meta up-scale module The above feature map up-sampling modules are all designed for specific integer scales, but arbitrary scale magnification is necessary for more practical applications. Meta up-scale module is proposed [135] for scale-free super-resolution tasks with introducing weight prediction strategy [194] of meta-learning [195] in existing SR networks. Compared to deconvolution [182] and sub-pixel [189] layers, it achieves promising performance on non-integer scale factors with no extra computational cost (less than 1% of the feature extraction module). Meanwhile, it dynamically predicts the weights of filters for each scale factor and requires no memory cost of storing groups of weights for different scales, such as in other multi-scale SR networks [133, 196]. The meta up-scale module consists of three components (Fig. 2.15: the location projection, the weight prediction and the feature mapping. In practice, the location projection first projects each pixel on

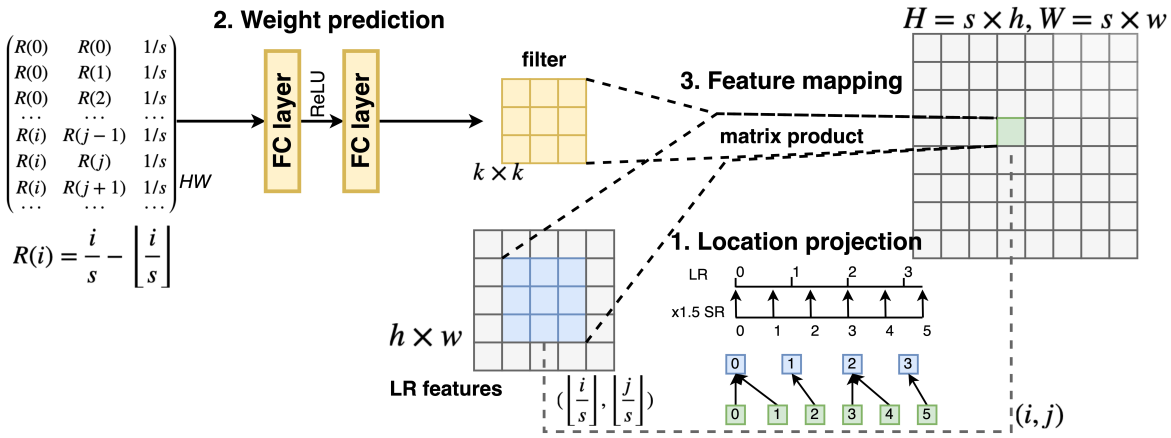


Figure 2.15: Meta upscale module [135] for feature map magnification with arbitrary scales. Value of each pixel in the up-sampled feature map are calculated in three steps: (1) local projection to the low-resolution feature maps; (2) weight prediction by a two-layer FC network; and (3) feature mapping with matrix product.

the SR feature maps (i, j) on the LR feature maps with floor function:

$$i', j' = \left\lfloor \frac{i}{s}, \frac{j}{s} \right\rfloor, \quad \text{for} \quad \boldsymbol{F}_{lr}(i', j') \sim \boldsymbol{F}_{sr}(i, j). \tag{2.28}$$

Then, the weights of the filters for each magnification scale are predicted with an input of the relative offsets of location projection and the magnification scale:

$$\boldsymbol{v}_{i,j} = (\frac{i}{s} - \left\lfloor \frac{i}{s} \right\rfloor, \frac{j}{s} - \left\lfloor \frac{j}{s} \right\rfloor, \frac{1}{s}). \tag{2.29}$$

Notice that the weights prediction applies on every pixel in the SR feature map respectively:

$$\boldsymbol{W}_{i,j} = P(\boldsymbol{v}_{i,j}), \tag{2.30}$$

where P is a two-layer fully-connected network. Finally in the feature mapping, the value of (i, j) in SR feature maps is calculated with the predicted weights and corresponding patch in $\boldsymbol{F}_{lr}$ by matrix product:

$$\boldsymbol{F}_{sr}(i, j) = \varrho(\boldsymbol{F}_{lr}(i', j')) \times \boldsymbol{W}_{i,j}, \tag{2.31}$$

where $\varrho(\cdot)$ is a cropping operation and the window size is controlled as the kernel size of convolution operations.

2.2.5 SR networks

Implementation of super-resolution networks includes CNNs [130, 133, 136], vision transformers [138, 190, 197], diffusion models [127, 161, 198] and hybrid methods [199]. De-

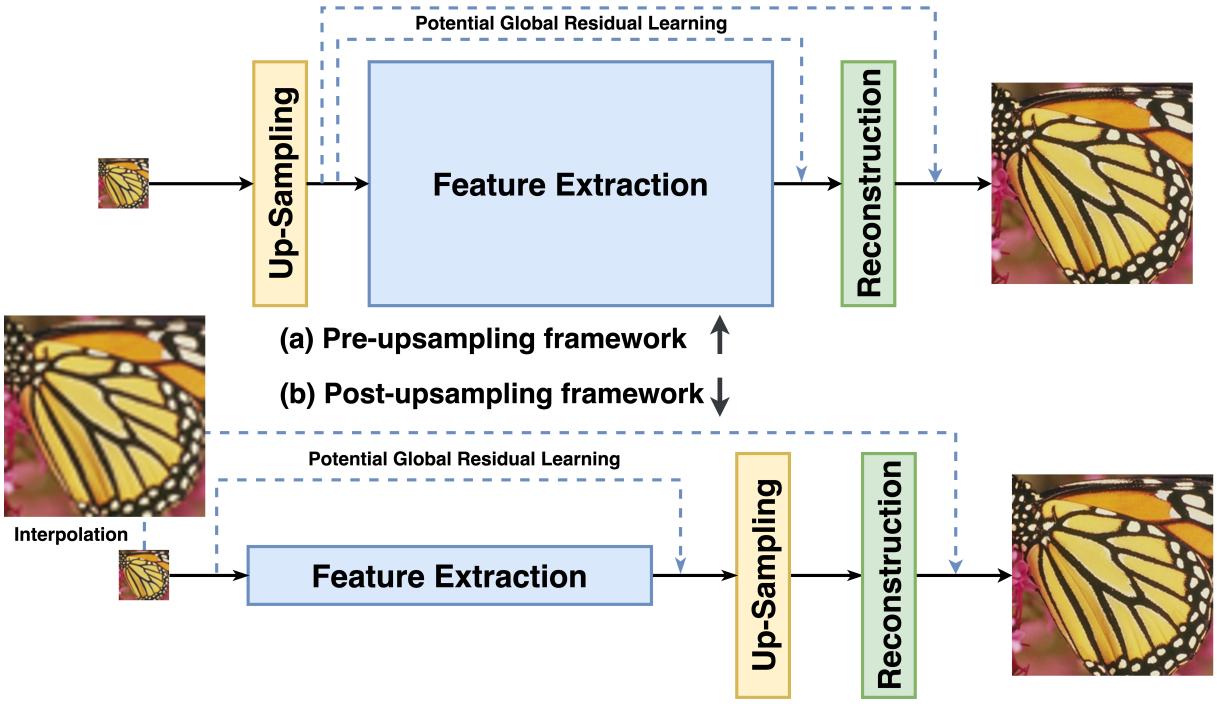


Figure 2.16: Pre-upsampling and post-upsampling super-resolution frameworks.

pending on where and how the magnification applies in the networks, the super-resolution frameworks mainly divide into pre- and post-upsampling frameworks (Fig. 2.16). These frameworks involve a wide range of deep learning techniques, such as recursive learning [199–202], local and global residual learning [131, 132, 203, 204], multi-path learning [133, 205–207], attention mechanisms [136, 137, 208] and U-Net architectures [97, 99, 209]. In this section, I restrict myself to the post-upsampling networks with an emphasis on SISR methods based on residual networks, attention CNNs and vision transformers. For a comprehensive review of SISR networks, I refer to the three citations [8–10].

Post-upsampling SR framework Early SISR networks follow the pre-upsampling architecture design, which first upsamples the LR image to the desired size and refines the magnified results with convolution layers, such as SRCNN [130], VDSR [131], MemNet[200] and DRCN [201]. Since this framework supports flexible magnification methods and scales, it can conveniently collaborate with non-network SR methods, such as degradation kernel estimation [210], learned data prior with flexible degradation control [211] and iterative back-projection [212–214]. However, the calculation of this pre-upsampling framework mainly conducts on high-resolution feature maps, leading to expensive computational costs and high memory requirements. Meanwhile, the predefined up-sampling may result in SR performance decline with noise amplification and image blur. Thus, the post-upsampling framework is introduced to SR networks in ESPCN [189] and FSRCNN [215], which first apply feature learning in low-resolution space and achieve feature maps magnification

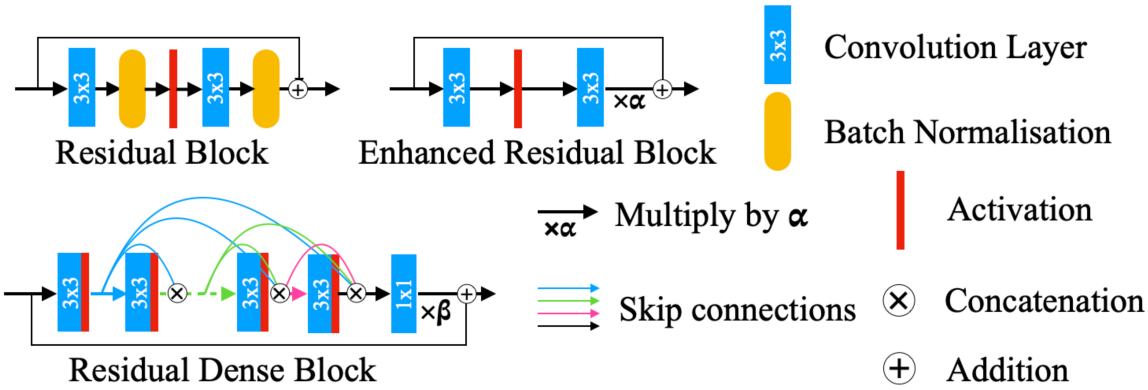


Figure 2.17: Widespread residual blocks in SISR networks. The residual block is introduced in SRGAN [114]; the enhanced residual block is introduced in EDSR [133]; and the residual dense block is introduced in RDN [132].

subsequently:

$$\boldsymbol{I}_{sr} = G(\boldsymbol{I}_{lr}) = \mathcal{M}(\mathcal{F}(\boldsymbol{I}_{lr})) \uparrow_s, \tag{2.32}$$

where $\mathcal{M}(\cdot) \uparrow_s$ is an up-sampling module introduced in Section 2.2.4 and $\mathcal{F}(\cdot)$ is the feature extraction module. This framework becomes the mainstream of SR networks with superior computational efficiency and improved super-resolution performance [114, 132, 133, 136, 138]. Meanwhile, progressive upsampling frameworks are used to reduce the learning difficulty of challenging SR tasks with big (e.g. $\times 4$ and $\times 8$) and multi magnification scales [24, 216–218], which perform intermediate supervision at multi scales in the network. Moreover, global residual learning further reduces the learning complexity because the target SR image highly corresponds to the input LR image. Networks can restrict the mapping to the high-frequency differences between interpolated LR images and HR images [131, 204] and to the residuals of shallow and deep feature maps before magnification [132, 138].

In the following part of this section, post-upsampling SISR networks based on CNNs, GANs and vision transformers will be introduced. They all have achieved SOTA performance on a broad range of SISR tasks with paper publishing.

Residual networks SRResNet [114] first introduces the residual block of ResNet [90] into deep super-resolution networks with GANs (i.e. SRGAN), resulting in photo-realistic high-resolution images. However, it is suboptimal to transplant this residual block from high-level recognition tasks to low-level image processing tasks with no modification. In EDSR [133], researchers claim that removing the unnecessary batch-normalisation layer [72] can lead to higher PSNR scores and more robust capability because this normalisation

disposes of the range flexibility from networks. Meanwhile, they experimentally show that training with L1 loss leads to better convergence and SR performance than L2, as reported in [219]. Moreover, SRDenseNet [220] employs the densely connected convolution block [91] to alleviate the gradient vanishing problem with improved information flow in super-resolution networks. In the dense block, feature maps are concatenated to the output of each convolution layer with skip connections, resulting in a growth rate of network width. Due to the extensive computation cost, the dense block struggles with a low growth rate that leads to relatively poor performance. Thus, the residual dense block (Fig. 2.17) is proposed in RDN [132], which introduces local feature fusion (LFF) in each dense block. Compared to a dense block, the LFF module fuses the feature maps of all convolution layers to the exact dimension of input feature maps with a 1×1 convolution layer. Meanwhile, the local and global residual learning and global feature fusion further improve the network representation ability of RDN. The feature fusion and residual learning mechanisms in RDN significantly reduce the computation cost of dense networks and encourage more stable information and gradient flows, leading to better performance than SRDenseNet with a bigger growth rate and more blocks. Additionally, the residual-in-residual dense block (RRDB) is introduced in ESRGAN [134], which combines successive RDBs with residual scaling [133, 221].

Attention CNNs The attention mechanism adaptively realises deep neural networks to the most informative regions of the input, leading to a more efficient and practical understanding of complex scenes in computer vision tasks [222]. It also demonstrates superior performance on CNN-based super-resolution networks with three categories: channel attention, spatial attention and non-local attention [223]. RCAN [136] first introduce the channel attention mechanism into residual SR networks. In contrast to non-attention methods (e.g. EDSR [133] and RDN [132]), RCAN considers the dependence of channel-wise features with global average pooling and *sigmoid* function, leading to flexible processing of low- and high- frequency information. The channel attention improves the representational ability of the network because SR tasks try to recover more high-frequency information than low-frequency information which remains in the LR image. DRLN [224] applies Laplacian pyramid attention to adaptively rescale features and model dependencies to learn the features at multiple sub-band frequencies. SAN [208] proposes second-order channel attention, which replaces the first-order global average pooling with second-order feature statistics to improve the discriminative ability further. Meanwhile, it captures long-distance spatial contextual information with region-level non-local attention, as in [225, 226]. On the other hand, SelNet [227] and RFANet [228] introduce spatial attention into SR networks. Moreover, PAN applies hybrid channel-spatial modules on efficient super-resolution while HAN [137] further implements a holistic attention network with

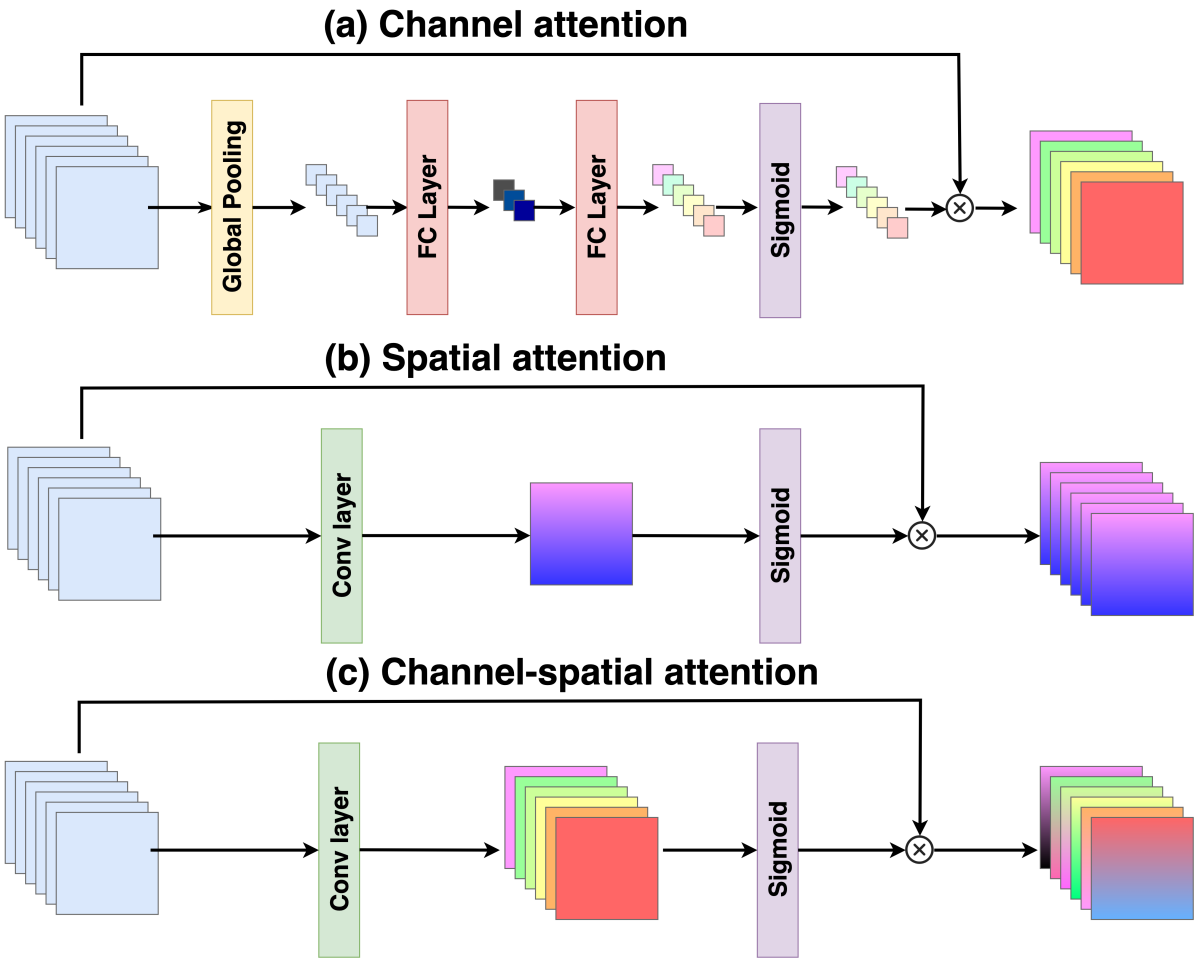


Figure 2.18: Attention modules in CNN for super-resolution. From top to the bottom: (a) channel attention considers the dependence of channel-wise features for flexible processing of low-/high- frequency information; (b) spatial attention captures long-distance spatial contextual information; and (c) hybrid channel-spatial attention holds both advantages.

the hybrid channel-spatial attention and the layer attention modules. Fig. 2.18 illustrates channel and spatial attention modules in these works.

Vision transformers With the boost success of vision transformers [117], IPT [127] first applies standard transformer blocks within a multi-task learning framework on low-level image restoration tasks, including super-resolution, deraining and denoising. Although it beats CNNs on a wide range of image processing tasks, the required computation resource and pre-training on a large-scale dataset [128] limits its applicability on domain-specific tasks (e.g. medical images). To avoid these limitations, SwinIR [138] proposes an efficient transformer block based on the shifted-window attention (Swin transformer [119] in Section 2.1.3 and Fig. 2.11 and 2.12). As a CNN-transformer hybrid method, it consists of a CNN-based shallow feature extractor and a group of high-quality image reconstruction modules (e.g. sub-pixel layers for super-resolution). Its main body is a stack of residual

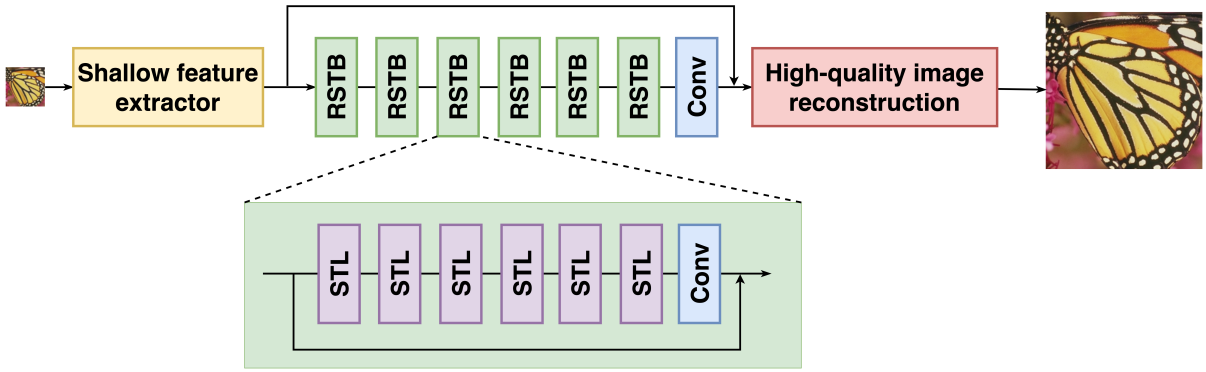


Figure 2.19: SwinIR [138] consists of three components: a CNN-based shallow feature extractor, various CNN-based tails for high-quality image reconstruction of multi-task learning, and a main body of shifted window attention transformer layers.

swin transformer blocks, each comprising six successive swin transformer layers (Fig. 2.19). Similarly, Uformer [229] implements a U-Net framework with locally-enhanced window transformer blocks and a novel multi-scale restoration modulator for multi-task learning on image restoration. Both transformers successfully avoid the expensive computational cost of global self-attention on high-resolution feature maps and the limitation of transformers in capturing local dependencies. Meanwhile, the multi-deconv transposed attention and gated-dconv feed-forward network are presented in Restormer [190] to aggregate local and non-local pixel interactions for controlled feature transformation. Additionally, hybrid CNN and transformer methods are proposed for efficient and lightweight super-resolution [197, 230, 231].

The progression of super-resolution approaches is not always linear, although the introduced methods successively push the SOTA performance forward. In practice, the choice of SR method depends on the exact requirements of the scenario, including the type and quality of available data, computational resources, desired output quality and image-specific characteristics of the images to be up-scaled. Classic methods (e.g. bicubic interpolation) are suitable when a quick and quality-regardless up-scaling result is required. They are fast, simple and friendly to any device. Deep learning-based methods can achieve SOTA SR image quality with substantial computational cost. They are suitable for scenarios when high-quality output is crucial, unlimited computational resources and an enormous amount of high-quality training data are available. Lightweight super-resolution methods balance output quality and model efficiency for using devices with finite computational resources (e.g. mobile devices) [232]. Additionally, super-resolution methods involve precise characteristics of domain-specific tasks, such as introducing identity preservation in face super-resolution tasks [233]. The following section will show how researchers explore the possibility of SR networks in the medical image domain.

2.3 Applications on medical images

In this section, I will briefly review super-resolution applications in the medical domain, emphasising CT and MR images. In contrast to natural image SISR tasks, medical image super-resolution often correlates with medical image analysis applications such as segmentation, classification and diagnosis, so it is required to preserve sensitive information and to enhance the structures of interest [234]. Meanwhile, modality-specific noises and artefacts may occur during the acquisition and further processing stages [16]. For example, the noise in MR images is generally formed as a stationary Rician distribution [235, 236], which performs as the additive square root of two independent Gaussian variables on the noiseless signal. The actual noise statistics in CT images are more challenging to determine, so it is assumed as a mixed Poisson-Gaussian distribution [237] or a Gaussian distribution [238, 239]. To simplify the impacts of noise models in super-resolution tasks (as in Equation 2.16), I use native Gaussian noises on the images in this thesis.

Before the explosion of deep neural networks, early applications of medical image super-resolution mainly relied on interpolation, reconstruction and example-learning methods [240]. Although these methods involve multi frames [241–243] and reference slices [244, 245] for HR image reconstruction, they achieve poor performance due to the limited representational capacity and lack of additional information of the training data. In contrast, deep learning-based methods significantly improve the performance of 2D (i.e. single-slice) and 3D (i.e. volumes) super-resolution of medical images. Generally, 2D methods have better applicability to a broad range of medical image modalities because they require fewer computation resources and adapt to most image formats with adjusted operations. On the other hand, 3D methods mainly focus on CT and MR images because of the limitation of data format. They may outperform 2D methods by involving more structural information along the slices (e.g. tumour boundaries), although heavier calculation and memory costs are required.

In [246], researchers implement 3D convolution in SRCNN [130] with global residual learning for brain MRI super-resolution. Meanwhile, SRCNN extends to multi-input image SR in [247] to reconstruct HR 3D images from multiple 2D slices of cardiac MR. In [248], researchers develop a context-sensitive SR algorithm with 3D SRCNN to learn organ-specific appearance for high-resolution image reconstruction with sharp edges and rich details. FSCWRN [249] presents a pre-upsampling SR framework with two main modifications in the residual block of EDSR [133] for 2D brain MR slices. First, it uses PReLU [48] to avoid the 'Dying ReLU'; second, it proposes a progressive wide residual block [250] with a fixed connection strategy to carry more high-frequency details to learn

local residuals effectively. The network in [251] is a deeper 3D SRCNN with local and global residual learning, applying on brain MR 3D super-resolution tasks with a comparison study to illustrate the superior performance than interpolation, non-local means [252] and sparse coding methods. DCSRN [253] applies 3D convolution in SRDenseNet [220] for 3D brain MR super-resolution. Due to the cumulative computation cost of 3D convolutions and dense connections, it consists of only one dense residual block of five layers. Meanwhile, Volumenet [254] implements a lightweight CNN module for fast and accurate 3D SR on brain MR and liver tumour CT images. It consists of a DenseNet [91] framework with group convolution and feature aggregation and a 3D sub-pixel module for feature map upsampling. In [255], researchers propose a channel splitting network with global feature fusion [132] and merge-and-run mapping [256], leading to a representational redundancy decline and hierarchical features integration. U-Net architectures are also widely used in medical image super-resolution tasks [97, 209, 257], especially for multi-task learning with segmentation [258, 259].

Due to the expensive computation cost, GAN-based medical image super-resolution methods rarely conduct 3D operations [260, 261]. MedSRGAN [262] implements a conditional GAN framework for 2D super-resolution on CT and MR images. The researchers use a modified RCAN [136] for SR image generation and a conditional discriminator, which takes the image pair of (LR, HR) and (LR, SR) as input. In addition to PSNR and SSIM, they also apply a 5-rank mean opinion score evaluation. FA-GAN [263] proposes a fused attentive GAN with CNN-based channel and non-local attentions for 2D MR super-resolution. FP-GANs [264] alleviates the detail-insensitive problem of CNN-based SR models by conducting SR in a divide-and-conquer manner with multiple GANs in the wavelet domain. Inspired by ESRGAN [134], a generator based on residual-in-residual blocks and the relativistic adversarial loss [106] is used for each sub-band of wavelet transformation. In [20], researchers propose a conditional GAN to synthesise high-resolution anatomically plausible 3D cardiac MR images with higher resolution. This method implements a multi-scale discriminator and a residual network generator with optical flow estimation to achieve through-plane super-resolution with transfer learning. Notice that they evaluate the SR result with reconstruction fidelity (i.e. PSNR and SSIM) and in segmentation tasks. CycleGAN [148] also benefits medical image super-resolution [184], such as with lesion-focused [265] and semi-supervised training [266]. Meanwhile, Wasserstein distances [22, 108] is widely used for robust and superior performance [267–270].

Vision transformers boost the performance of medical image super-resolution and related low-level image processing tasks such as reconstruction and denoising [271]. For example, ReconFormer [272] is proposed for high-resolution MR image reconstruction from

under-sampled $k-$space data. Compared to SwinIR [138], it incorporates the pyramid structure inside a transformer layer to perceive multi-scale representation and employs recurrent pyramid layers to exploit deep feature correlation. In [273], researchers present a SwinIR application on medical images in comparison with SRGAN [114], BSRGAN [274] and Real-ESRGAN [275], including chest x-ray, skin lesion and funds image. In [276], researchers implement 3D operation and self-supervised pre-training on a U-Net framework with swin transformer layers [119] for medical image analysis tasks such as segmentation. In contrast, in [277] the SwinIR [138] framework works in the frequency domain for fast MR reconstruction. In [278], McMRSR applies SwinIR for multi-contrast MR super-resolution with multi-scale contextual matching and aggregation schemes. It achieves SR images with rich details by capturing more long-range dependencies and transferring visual contexts from reference images to target LR MR images at different scales.

2.4 Chapter summary

In this chapter, I introduce the fundamental concepts of deep neural networks and widespread architectures of convolutional neural networks, generative adversarial networks and vision transformers. Then I explain the single image super-resolution problem with a comprehensive review of advanced deep learning-based methods, including problem definition, network frameworks, SR image quality assessment, loss function design and medical image applications. These SISR networks are the baseline for comparison in my research work, which I will present in the next Chapter 3, 4 and 5.

CHAPTER 3

Lesion-focused Multi-Scale GAN for Medical Image Super-Resolution

3.1 Introduction

Achieving perceptually realistic high-resolution results in medical image magnification tasks is challenging. While generative adversarial networks have shown impressive results in producing photo-realistic images for natural image super-resolution tasks, they face certain limitations when applied to medical images. Briefly speaking, there are two main research questions when applying GANs to medical image super-resolution tasks for robust performance with limited training data. First, how to stabilise the training of GANs? Second, how to avoid generating unreal textures? This chapter will present my research on developing a lesion-focused multi-scale GAN to solve these questions. By introducing several modifications to the original SRGAN [114], the proposed method has led to more stable and efficient training and significant improvements in pixel-wise reconstruction fidelity (e.g. PSNR scores) and perceptual quality.

SRGAN [114] is successful in generating photo-realistic natural images because of its powerful SR image generator and the well-designed combination loss. As mentioned in Section 2.2.5, it develops a deep generator of 16 residual blocks [90], which enhances the ability to reconstruct high-resolution details. Compared to earlier SR networks [130, 201, 215], the ResNet-based SR image generator (referred to as SRResNet) significantly improves the PSNR scores on $\times 2$ and $\times 4$ SR tasks for natural images, such as photos of animals and buildings. Meanwhile, SRGAN utilises adversarial learning to attain rich textures in high-resolution results. It incorporates pixel-wise L2 loss, vanilla GAN loss, and perceptual

loss to encourage generating reliable and perceptually realistic details that closely resemble natural images.

However, SRGAN experiences difficulty with unstable training and abnormal textures when applied to medical images. First, the unstable training of the vanilla GAN [101] means the issues of convergence during the training process [108, 110] (i.e. the loss value cannot decrease smoothly). When using deep architectures, GANs are susceptible to vanishing gradients and oscillation, leading to no updates of the networks, rapidly changing loss values or fluctuations in the quality of generated samples. Additionally, mode collapse and mode dropping may occur when the generator fails to capture the full diversity of the data distribution or the discriminator becomes too effective at distinguishing between real and generated images. In this case, the generator will restrict itself to certain modes of the training data distribution to avoid the heavy penalty of adversarial loss. Thus, network implementation and training settings are very important. Researchers work on training tricks, including hyper-parameters searching, discriminator architecture design [110] and generator pre-training [114]. For example, SRGAN applies warm-up training (i.e. to train the generator separately with L2 loss before training the GAN) to achieve a good start point for the GAN, because the training stability is sensitive to the network weights initialisation. However, these skills are either time-consuming or task-specific, resulting in limitations to apply in medical image super-resolution tasks generally.

In contrast, without GANs, CNN-based methods can avoid the unstable training and perform well in SR tasks with small magnification scales (e.g. $\times 2$ magnification) but lead to over-smoothing in SR tasks with larger magnification scales such as $\times 4$. These blurred images are unacceptable in clinics for following analysis tasks or for doctors because high-frequent textures with crucial information are missing.

Moreover, the distributions of textures in different regions in medical images are desperate. In practice, spatial resolution and image quality inside lesion regions capture more attention than in regular regions. For example, the boundaries and textures of tumours are more critical than other brain parts because they are the core information in tumour segmentation, reconstruction, and clinical processing [279]. LR-HR patches cropped from the noisy background or non-lesion areas increase convergence difficulty because they misdirect the training. As a result, SR networks learn useless noise transformations and tend to generate SR images with unrealistic textures.

Finally, it is also challenging to quantify the perceptual quality of generated high-resolution medical images. PSNR and SSIM perform well in measuring the reconstruction fidelity

but cannot evaluate how good the generated images are in clinical tasks. Thus, a reliable metric is also needed.

In this chapter, I restrict myself to the more challenging $\times 4$ single image super-resolution with medical images. To tackle the above problems in GAN-based methods, I take SRGAN [114] as a baseline and apply three improvements in the proposed method MS-GAN. First, the lesion-focused training strategy (Section. 3.2.2) is proposed to avoid the adverse effects of non-lesion regions in training SISR networks. Second, a multi-scale SR image generator (Section. 3.2.3) is designed to decompose the challenging $\times 4$ SR task into a series of simpler sub-problems and to improve the perceptual quality of generated images. Third, to stabilise the training, I implement Wasserstein distance with gradient penalty (WGAN-GP [22]) as the adversarial loss (Section. 3.2.4). Simulation experiments are conducted on three medical image datasets (one public dataset and two real clinical datasets) to compare the proposed method with state-of-the-art SISR methods. In addition to PSNR and SSIM, a subjective evaluation based on mean opinion scores (MOS) is designed and performed by experienced radiologists on the testing images to quantify the perceptual reality of generated SR images. Briefly speaking, the main contributions of this work are:

- A lesion-focused multi-scale GAN (MS-GAN) is proposed for medical image single-image super-resolution tasks with large magnification scales ($\times 4$ of each side). It dramatically improves the perceptual quality of generated SR images, leading to comparable scores to ground truth high-resolution images in the subjective evaluation of experienced radiologists. Compared with SRGAN [114] (SOTA method in 2018), it is more efficient and robust because the time-consuming warm-up training is no longer necessary.

- I implement a lesion-focused training strategy to stabilise the training of GAN-based SISR methods for medical images. In $\times 2$ and $\times 4$ SISR tasks (w/wo additive white Gaussian noise in the $k-$space) with brain tumour MR images, it results in a significant improvement of PSNR (+1.13 dB) and SSIM (+0.029) on average.

- I first introduce the Wasserstein GAN with gradient penalty into medical image super-resolution tasks. With a comprehensive comparison study of GAN variations, a combined SR loss function based on WGAN-GP is finally proposed. It has achieved the most perceptually realistic results in the challenging $\times 4$ magnification tasks. Compared with SRGAN in a 4-rank (1 for the worst and 4 for the best perceptual quality) MOS evaluation, it leads to significant improvements of +0.80 on cardiac MR images and +1.46 on brain MR images.

This chapter is organised as follows: Section 3.2 introduces the lesion-focused training strategy and the multi-scale GAN with the SR loss functions; Section 3.3 claims the

experimental settings, including data, evaluation and implementation details; Section 3.4 compares the SR results of the proposed MS-GAN with SOTA SISR methods and illustrates the impacts of each component in the ablation study; and finally Section 3.5 concludes the work of this chapter. All related publications and code are publicly realised on `https://github.com/GinZhu/MSGAN`.

3.2 Methodology

Single image super-resolution methods aim to obtain a high-resolution image from a low-resolution image. Generally, deep learning-based SISR methods are designed for magnification tasks with a specific scale such as $\times 2$, $\times 3$, and $\times 4$, which can be defined as:

$$\boldsymbol{I}_{sr} = G_s(\boldsymbol{I}_{lr}; \boldsymbol{\theta}_G), \tag{3.1}$$

where $\boldsymbol{I}_{lr}$ is the low-resolution image as the input, $\boldsymbol{I}_{sr}$ is the aimed super-resolved image, G_s is the SR image generator which is designed with the magnification scale s, and $\boldsymbol{\theta}_G$ is its trainable parameters. Most CNN-based frameworks are post-interpolation to avoid the bias of early interpolation and reduce computational costs. They first extract low-dimensional image features with a sub-network and increase the dimensions of the feature maps to achieve a high-resolution result at the end, which can be defined as:

$$\boldsymbol{I}_{sr} = G(\boldsymbol{I}_{lr}) = \mathcal{M}_s(\mathcal{F}(\boldsymbol{I}_{lr})), \tag{3.2}$$

Where $\mathcal{F}$ is the low dimensional feature extraction network and $\mathcal{M}_s$ is an up-sampler which increases both dimensions of the feature maps by s times.

Here, I introduce a popular single image super-resolution generator SRResNet [114] in Fig. 3.1, which is the baseline of my proposed method.

3.2.1 SRResNet

SRResNet is an end-to-end high-resolution image generator, which consists of a residual neural network [90] for low dimensional feature extraction and an upsampling module (Fig. 3.1).

ResNet for LR feature extraction The feature extractor, $\mathcal{F}$, is constructed by stacking up n residual blocks $\mathcal{B}_R$. Each residual block consists of two convolution layers $\mathcal{C}_i$, non-linear activation layers $\varphi(\cdot)$, batch normalisation layers $\mathcal{N}_i$ and a residual connection:

$$\boldsymbol{F}_{out} = \mathcal{B}_R(\boldsymbol{F}_{in}) = \boldsymbol{F}_{in} + \varphi(\mathcal{N}_2(\mathcal{C}_2(\varphi(\mathcal{N}_1(\mathcal{C}_1(\boldsymbol{F}_{in}; \phi_{\mathcal{C}_1}); \phi_{\mathcal{N}_1})); \phi_{\mathcal{C}_2}); \phi_{\mathcal{N}_2})), \tag{3.3}$$

where $\phi_{\mathcal{C}_1}, \phi_{\mathcal{C}_2}, \phi_{\mathcal{N}_1}, \phi_{\mathcal{N}_2} \in \boldsymbol{\theta}_{G_s}$ are trainable parameters of each layer in the block. $\boldsymbol{F}_{in}$ and $\boldsymbol{F}_{out}$ are the input and output feature maps, respectively. The dimensions of the input and output feature maps remain because zero-padding operations are applied correspondingly in these convolutional layers. Finally, the LR feature maps $\boldsymbol{F}_{lr}$ are extracted from the

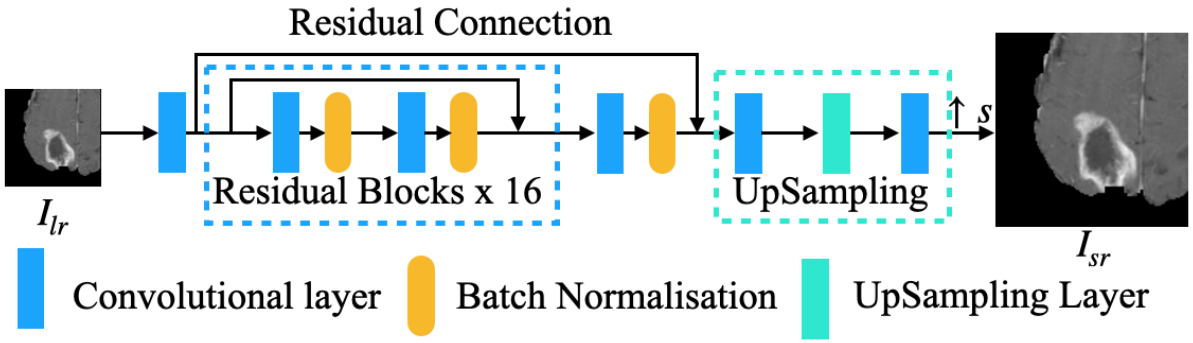


Figure 3.1: SRResNet [114] consists of a residual neural network-based low-dimension image feature extractor and an upsampling layer. The feature extractor contains 16 residual blocks, each of which has two 3×3 convolutional layers followed by non-linear activation and batch normalisation. The upsampling layer is designed for one specific magnification scale s. The network takes one low-resolution image $\boldsymbol{I}_{lr}$ as the input and generates its super-resolved version $\boldsymbol{I}_{sr}$ with a higher resolution.

input LR image:

$$\boldsymbol{F}_{lr} = \mathcal{B}_R^n(\boldsymbol{I}_{lr}; \phi_{\mathcal{B}}), \tag{3.4}$$

where $\phi_{\mathcal{B}} \in \boldsymbol{\theta}_G$ denotes the trainable parameters of each block, and n is the number of residual blocks.

Sub-Pixel for upsampling In the original SRResNet, a transposed convolutional layer (i.e. de-convolution layer [182]) is used for the feature maps magnification. Here a more efficient up-sampler layer, the sub-pixel layer [189], is used to reduce the calculation cost by s^2 times where s is the upscale factor. The sub-pixel layer increases the dimensions of the LR feature maps $\boldsymbol{F}_{lr}$ of shape $[C \times W_{lr} \times H_{lr}]$ to feature maps $\boldsymbol{F}_{sr}$ of shape $[C \times sW_{lr} \times sH_{lr}]$ in two steps. First, the width and height remain, but the number of channels is increased by s^2 times to achieve the hidden feature maps $\boldsymbol{F}_{mid}$. Then the feature maps $\boldsymbol{F}_{mid}$ of shape $[s^2C \times W_{lr} \times H_{lr}]$ are rearranged to the aimed shape. Take $\times 2$ magnification as an example:

$$\begin{aligned} \boldsymbol{F}_{mid}(i) &= \boldsymbol{F}_{lr} \star \boldsymbol{k}_i, \quad \text{for } i \in (1, 2, 3, 4) \\ \boldsymbol{F}_{sr} &= \boldsymbol{F}_{mid}(1) \oplus \boldsymbol{F}_{mid}(2) \oplus \boldsymbol{F}_{mid}(3) \oplus \boldsymbol{F}_{mid}(4), \end{aligned} \tag{3.5}$$

where $\boldsymbol{k}_i$ is a convolution kernel and $\oplus$ denotes the periodic shuffling operation. Notice that the sub-pixel layer can only work with integer scales such as 2, 3, and 4.

Training of SRResNet The SISR image generator can be trained end-to-end with paired

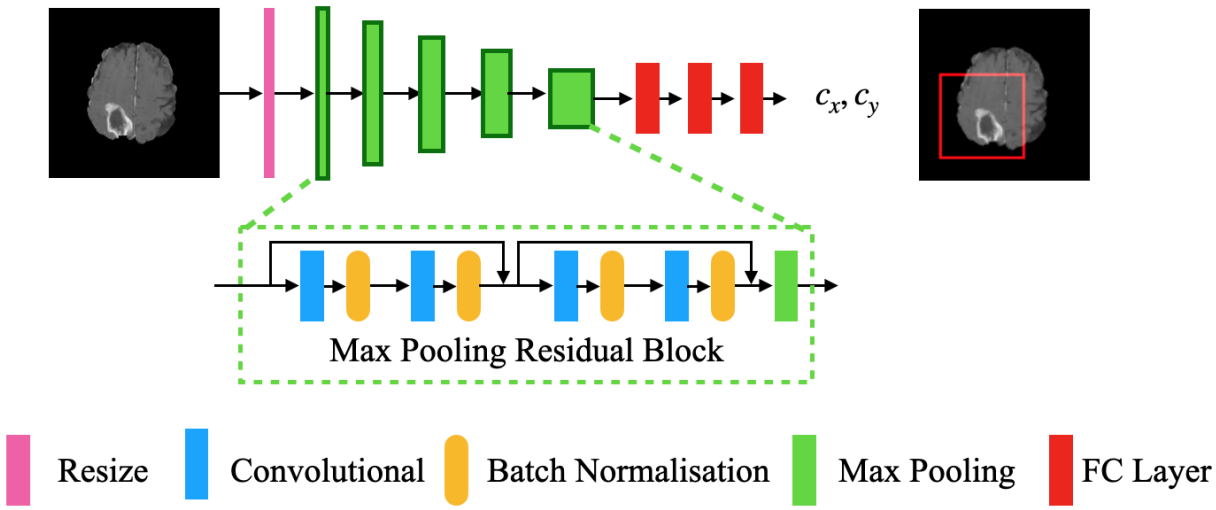


Figure 3.2: The lesion detection network consists of a resizing layer as the head, a 5-level encoder as the body and three fully-connected layers as the tail. Each level of the encoder has a max pooling residual block, which consists of two residual blocks and a max pooling layer. Similar to the U-Net [92] encoder, the size of the feature maps is halved at the end of each level while the number of convolutional kernels is doubled. Finally, ROI is cropped based on the predicted centre (c_x, c_y).

LR and HR images:

$$\hat{\boldsymbol{\theta}_G} = arg \min_{\boldsymbol{\theta}_G} \mathcal{L}(G(\boldsymbol{I}_{lr}; \boldsymbol{I}_{lr} \in \mathbb{I}_{lr}), \boldsymbol{I}_{hr}; \boldsymbol{I}_{hr} \in \mathbb{I}_{hr}), \tag{3.6}$$

where $\mathbb{I}_{hr}$ and $\mathbb{I}_{lr}$ are the distributions of HR and LR images in high dimensional and low dimensional spaces, and for each $\boldsymbol{I}_{hr} \in \mathbb{I}_{hr}$, there is a degraded version $\boldsymbol{I}_{lr} \in \mathbb{I}_{lr}$. $\mathcal{L}_{SR}$ is a well-designed loss function for super-resolution as mentioned in Section 2.2.2.

3.2.2 lesion-focused training

However, in practice, solving Equation (3.6) for medical image SISR tasks with large magnification scales is very challenging. In practice, the images are normally cropped to smaller patches before feeding to the network because of the limitation of GPU memory. However, the texture distribution of different regions, such as tumours and brains, are divergent. This requires the network to learn various magnification transformations synchronously, increasing convergence difficulty. Meanwhile, not all regions are equally important in the clinic. Diagnosis and follow-up image analysis tasks such as segmentation and detection mainly concentrate on the lesion.

Thus, I propose a lesion-focused training strategy for medical image SISR tasks to avoid the impacts of incidental textures. The region of interest (ROI) of the lesions or abnormalities (e.g. tumours in brain MR images) are first cropped by a lesion detection network LD.

As a result, the original training LR-HR image pairs $(\mathbb{I}_{lr}, \mathbb{I}_{hr})$ are reduced to $(\mathbb{I}'_{lr}, \mathbb{I}'_{hr})$. For each pair of $(\boldsymbol{I}_{lr}, \boldsymbol{I}_{hr})$, there is:

$$\boldsymbol{I}^{'}_{lr}, \boldsymbol{I}^{'}_{hr} = LD(\boldsymbol{I}_{lr}, \boldsymbol{I}_{hr}; \boldsymbol{\theta}_{LD}), \tag{3.7}$$

where $\boldsymbol{\theta}_{LD}$ denotes the trainable parameters of the lesion detection network. Notice that $\boldsymbol{I}^{'}_{lr}$ and $\boldsymbol{I}^{'}_{hr}$ must be corresponded, which means each new image pair must be in the same region. The lesion detection network only takes one image as input. It predicts the centre (c_x, c_y) of the ROI, then projects the coordinates to the image pair and crops the ROIs of a preset size on both images.

The lesion detection network (Fig. 3.2) is input-scale free because a resizing layer is applied as the head to adjust the shape of input images. In practice, HR images are used for training, but LR images are used to predict the ROI for inference. The main body of LD is a 5-level encoder. Similar to the U-Net encoder [92], each level consists of two residual blocks and halves the size of feature maps by max pooling. Meanwhile, the width of each level (i.e. the number of convolutional kernels) is doubled. In the end, three fully-connected layers flatten the feature maps and predict the ROI centre (c_x, c_y). In the lesion-focused SISR tasks, LD is first trained separately with the same dataset:

$$\hat{\boldsymbol{\theta}_{LD}} = arg \min_{\boldsymbol{\theta}_{\mathcal{LD}}} \mathcal{L}_2(LD(\boldsymbol{I}_{hr}), (\bar{c_x}, \bar{c_y}); \boldsymbol{I}_{hr} \in \mathbb{I}_{hr}), \tag{3.8}$$

where $(\bar{c_x}, \bar{c_y})$ is the centre of ROI generated from manual labels, and the training aims to minimise the L2 distance (i.e. $\mathcal{L}_2$) between the predicted and ground-truth centres.

Then it runs on $\mathbb{I}_{lr}$ and updates both LR and HR images in the training dataset:

$$\mathbb{I}'_{lr}, \mathbb{I}'_{hr} = LD(\mathbb{I}_{lr}, \mathbb{I}_{hr}). \tag{3.9}$$

Thus, the training processing of the SR image generator is updated as:

$$\hat{\boldsymbol{\theta}_{G}} = arg \min_{\boldsymbol{\theta}_{G}} \mathcal{L}(G(\boldsymbol{I}^{'}_{lr}; \boldsymbol{I}^{'}_{lr} \in LD(\mathbb{I}_{lr})), \boldsymbol{I}^{'}_{hr}; \boldsymbol{I}^{'}_{hr} \in LD(\mathbb{I}_{hr})). \tag{3.10}$$

3.2.3 Multi scale SR image generator

GAN-based adversarial learning is introduced to SISR tasks [114] to generate perceptually realistic images. However, it tends to introduce non-realistic textures and unstable training in SR tasks with large magnification scales (e.g. $\times 4$ super-resolution). To stabilise the training process and avoid these textures, I propose a multi-scale GAN (MS-GAN)

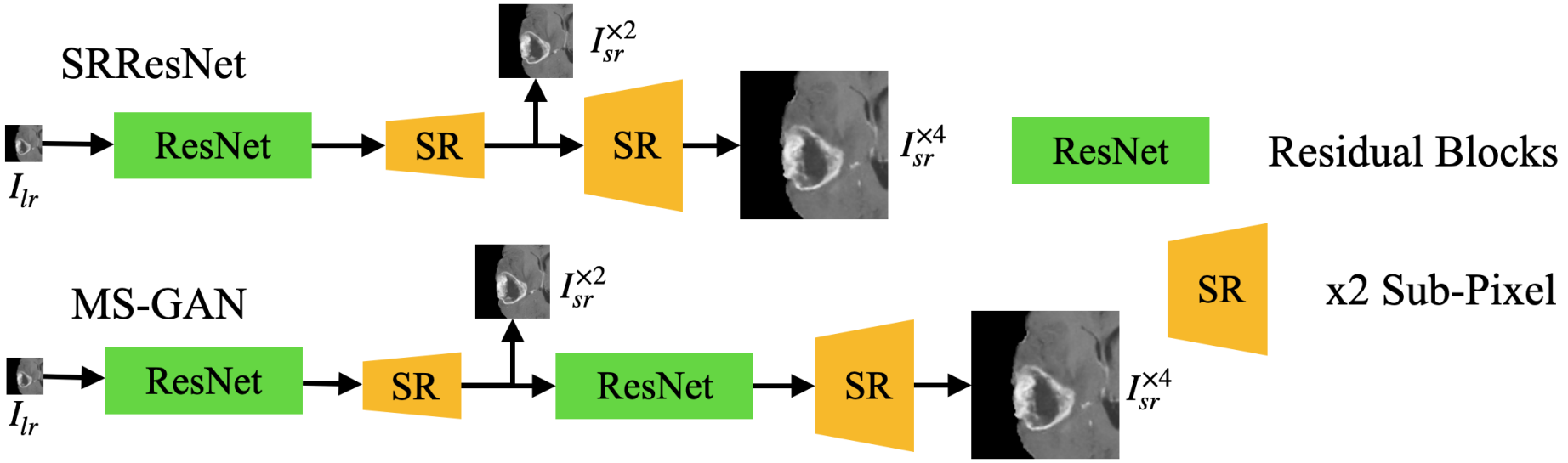


Figure 3.3: Two multi-scale image generators for $\times 4$ SISR tasks. The top one is based on SRResNet [114], in which the $\times 4$ sub-pixel layer is replaced by two $\times 2$ ones and a convolutional layer is inserted to generate the intermediate $\times 2$ SR image. The bottom one, MS-GAN, decomposes the $\times 4$ SR task into two sequential $\times 2$ SR tasks. It first generates the $\times 2$ SR image by one SRResNet and then passes the enlarged feature maps to another $\times 2$ SRResNet to achieve the final $\times 4$ SR image $\boldsymbol{I}_{sr}^{\times 4}$.

architecture to decompose this problem into simpler sub-problems.

Take the $\times 4$ SR task as an example. The MS-GAN not only generates the final SR image $\boldsymbol{I}_{sr}^{\times 4}$ but also generates a $\times 2$ SR image $\boldsymbol{I}_{sr}^{\times 2}$ as an intermediate result:

$$\boldsymbol{I}_{sr}^{\times 2}, \boldsymbol{I}_{sr}^{\times 4} = G(\boldsymbol{I}_{lr}; s = 4, \boldsymbol{\theta}_G). \tag{3.11}$$

Correspondingly, the $\times 2$ ground truth images need to be generated from the $\times 4$ high-resolution images by down-sampling, and the loss function also needs to be updated:

$$\hat{\boldsymbol{\theta}_G} = arg \min_{\boldsymbol{\theta}_G} \mathcal{L}_{SR}(G(\boldsymbol{I}_{lr}), \boldsymbol{I}_{hr}, \downarrow_2 \boldsymbol{I}_{hr}). \tag{3.12}$$

Two multi-scale SR image generators (Fig. 3.3) are proposed and tested for $\times 4$ medical image SR tasks. Both generators are based on SRResNet. The first one, the so-called MS-SRResNet, remains the LR feature extractor but replaces the $\times 4$ up-sampling layer with two $\times 2$ sub-pixel layers. It also adds a convolutional layer in between to reconstruct the intermediate $\times 2$ SR image. The second one, MS-GAN, consists of two sequential $\times 2$ SRResNets. The first network generates $\times 2$ SR image and passes the enlarged feature maps to the second network to achieve the final $\times 4$ result. Although I implement both models for $\times 4$ SR tasks, the same architectures can be easily extended to other magnification scales. Depending on the superior performance in the following experiments in Section 3.3, the MS-GAN framework is suggested. It can be mathematically represented as:

$$\begin{aligned} \boldsymbol{I}_{sr}^{\times 2} &= \mathcal{C}_1(\mathcal{M}_{\times 2}^1(\mathcal{F}_1(\boldsymbol{I}_{lr}))), \\ \boldsymbol{I}_{sr}^{\times 4} &= \mathcal{C}_2(\mathcal{M}_{\times 2}^2(\mathcal{F}_2(\mathcal{M}_{\times 2}^1(\mathcal{F}_1(\boldsymbol{I}_{lr}))))), \end{aligned} \tag{3.13}$$

where $\mathcal{C}_i$ represents a convolutional layer which converts feature maps to super-resolved images.

3.2.4 SR loss functions with WGAN-GP

Following the success of SRGAN [114] on natural image super-resolution tasks, a combined loss is used in this work. It consists of the pixel-wise mean-square-error (MSE), adversarial loss and VGG-based perceptual loss:

$$\mathcal{L}_{SR} = \mathcal{L}_{MSE} + \lambda \mathcal{L}_{adv} + \eta \mathcal{L}_{perc}. \tag{3.14}$$

Briefly speaking, the MSE loss represents the reconstruction fidelity of generated images. In contrast, the perceptual and adversarial losses help make the generated images percep-

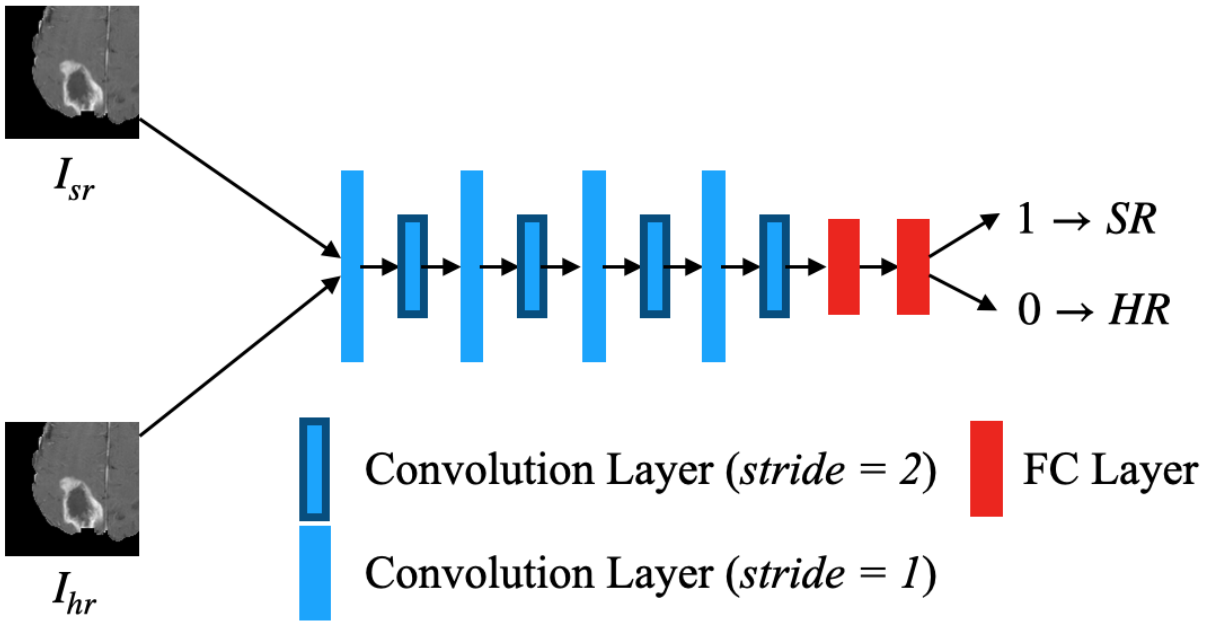


Figure 3.4: The discriminator of Wasserstein GAN consists of 4 dual-convolution blocks and two fully-connected layers. In the dual-convolution block, the first convolutional layer has $stride = 1$ and remains the size of feature maps; the second layer doubles the number of convolution kernels and halves the size of the feature maps by setting $stride = 2$. After that, the fully-connected layers flat the features maps and predict a flag in $[0, 1]$, where 0 means the input image is a generated SR image while 1 means the input image is an HR ground truth image.

tually more realistic, so the scale factors λ and η balance their impacts.

Mean square error (L2 loss) Mean-square-error represents the pixel-wise L2 distance between ground truth HR images and generated SR images and is prevalent in training SISR models. It is defined as:

$$\mathcal{L}_2(\boldsymbol{I}_{sr}, \boldsymbol{I}_{hr}) = \frac{1}{H * W} \sum_{(i,j) \in \boldsymbol{I}} \|\boldsymbol{I}_{sr}[i,j] - \boldsymbol{I}_{hr}[i,j]\|^2 \tag{3.15}$$

where H and W are the height and width of the images. In this work, both MSEs in between the $\times 4$ and $\times 2$ images are considered:

$$\mathcal{L}_{MSE} = \alpha \mathcal{L}_2(\boldsymbol{I}_{sr}^{\times 2}, \boldsymbol{I}_{hr} \downarrow_2) + \beta \mathcal{L}_2(\boldsymbol{I}_{sr}^{\times 4}, \boldsymbol{I}_{hr}), \tag{3.16}$$

where α and β are scale factors.

The VGG-based perceptual loss The perceptual loss [139] is based on a VGG-19 network [87], which has been well-trained on natural image classification tasks. It measures the average L2 distance of two image groups in the feature domain:

$$\mathcal{L}_{perc}(\boldsymbol{I}_{sr}, \boldsymbol{I}_{hr}) = \mathbb{E}(\|\mathcal{V}_l(\boldsymbol{I}_{hr}) - \mathcal{V}_l(\boldsymbol{I}_{sr})\|^2), \tag{3.17}$$

where l denotes the layer index of feature maps in the pre-trained VGG-19 network $\mathcal{V}$. Technically, these feature maps represent hidden visual information of images. In the feature domain, early layers are more about local and structural information, such as edges and corners, while deep layers may denote global and semantic information. Similar to SRGAN [114], the feature maps after activation of deep layers are used in this work.

Wasserstein GAN with gradient penalty Unlike SRGAN, I use Wasserstein GAN with gradient penalty (WGAN-GP) in the training of MS-GAN, to avoid unstable training and mode collapse.

GANs can help to generate more perceptually realistic images. During training, the discriminator aims to learn a metric to distinguish the fake samples (e.g. generated SR images) from the real ones (e.g. HR ground truth images). In contrast, the generator aims to fool the discriminator. Both networks are trained jointly, so the capability of the discriminator for recognising generated SR images and the generator's ability to generate realistic SR images ideally increase synchronously. This learning process can be defined as:

$$\hat{\boldsymbol{\theta}_G}, \hat{\boldsymbol{\theta}_D} = arg \min_{\boldsymbol{\theta}_G, \boldsymbol{\theta}_D} \mathcal{L}_{adv}(G(\mathbb{I}_{lr}), D(\mathbb{I}_{hr}, G(\mathbb{I}_{lr}))). \tag{3.18}$$

The WGAN discriminator (Fig. 3.4) predicts a value as the real-or-fake flag for each image. It consists of 4 dual-convolution blocks and two fully-connected layers. Each dual-convolution block reduces the size of feature maps by a convolution operation with $stride = 2$ and doubles the number of convolution kernels. The FC layers flat the feature maps and predict the flag as a float number. Notice that this is not a binary classification problem, although fake and real samples are respectively labelled as 0 and 1. Instead, the Wasserstein distance between the predicted flags and the true labels is used to measure the error of D:

$$\begin{aligned} \mathcal{L}_{WGAN} &= \mathcal{L}_{real} + \mathcal{L}_{fake} \\ &= \mathbb{E}_{\boldsymbol{I}_{hr}}(|1 - D(\boldsymbol{I}_{hr})|) + \mathbb{E}_{\boldsymbol{I}_{lr}}(|0 - D(G(\boldsymbol{I}_{lr}))|) \\ &= \mathbb{E}_{\boldsymbol{I}_{hr}}(|1 - D(\boldsymbol{I}_{hr})|) + \mathbb{E}_{\boldsymbol{I}_{lr}}(|D(G(\boldsymbol{I}_{lr}))|). \end{aligned} \tag{3.19}$$

The gradient penalty must be added as a restriction term of the gradients to ensure that the weights of D will not change rapidly for the condition of derivable Wasserstein distance. Thus, the advanced adversarial loss is defined as:

$$\mathcal{L}_{adv} = \mathcal{L}_{WGAN} + \mathbb{E}_{\boldsymbol{I}} \left[\left\| \bigtriangledown_{\boldsymbol{I}} D(\boldsymbol{I}) \right\|_p - 1 \right]^2, \tag{3.20}$$

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

Notice that the adversarial and perceptual losses are only applied to $\times 4$ SR images to avoid introducing non-realistic textures in the early image reconstruction stage. In summary, the loss I used to train MS-GAN is defined as:

$$\mathcal{L}_{SR} = \alpha\mathcal{L}_2(\boldsymbol{I}_{sr}^{\times 2}, \boldsymbol{I}_{hr}\downarrow_2) + \beta\mathcal{L}_2(\boldsymbol{I}_{sr}^{\times 4}, \boldsymbol{I}_{hr}) + \lambda\mathcal{L}_{adv}(\boldsymbol{I}_{sr}^{\times 4}, \boldsymbol{I}_{hr}) + \eta\mathcal{L}_{perc}(\boldsymbol{I}_{sr}^{\times 4}, \boldsymbol{I}_{hr}). \quad (3.21)$$

In summary, the lesion-focused training strategy (Equation. 3.10) and the multi-scale loss (Equation. 3.21) are jointly applied in $\times 4$ medical image SISR tasks.

3.3 Experiments

The experiments are conducted on one public dataset and two clinical ones from my collaborator. Section 3.3.1 will introduce more details about the data format and acquisition. Two objective metrics and one subjective metric are used to evaluate the quality of generated images (Section 3.3.2). Section 3.3.3 presents the implementation details. Notice that all experiments are conducted on 2D images, but the proposed method can be extended to 3D constructs by modifying 2D operations to 3D.

3.3.1 Data and pre-processing

This section will introduce the details of data extraction and pre-processing of the three MR datasets. Due to the limitation of acquiring LR-HR medical image pairs in the clinic, experiments are conducted on simulated samples. Low-resolution images are generated by down-sampling the original high-resolution slices, as follows.

Training/testing data generation In the training dataset, the original images, considered as the HR ground truth, are down-sampled with scale 2 as the intermediate ground truth images and with scale 4 as the LR images:

$$\begin{aligned} \forall I_{ori} \in \mathbb{I}_{train}, \quad & \boldsymbol{I}_{hr} = I_{ori}; \\ & \boldsymbol{I}_{hr}\downarrow_2 = I_{ori}\downarrow_2; \\ & \boldsymbol{I}_{lr} = I_{ori}\downarrow_4 . \end{aligned} \quad (3.22)$$

In the testing dataset, the HR images are only down-sampled by 4 to generate LR images:

$$\begin{aligned} \forall I_{ori} \in \mathbb{I}_{test}, \quad \boldsymbol{I}_{hr} &= I_{ori}; \\ \boldsymbol{I}_{lr} &= I_{ori} \downarrow_4 . \end{aligned} \tag{3.23}$$

Notice that no random patch cropping is used in this work. During the training of SR networks, the undivided ROIs of HR GT and generated LR samples are fed to the networks. Thus, the input and output shapes of training patches depend on the ROI size of each dataset. The HR patch size (i.e. the output shape) is the same as the ROI size, while the LR patch size (i.e. the input shape) is scaled down. The original pixel value and data format varies depending on the MR sequences. In the pre-processing, I convert all samples to 32-bit floats in the $[0, 1]$ range by dividing each dataset's maximum value, leading to consistent network implementation and simplified comparison. Additionally, lesions and sizes are defined separately for each dataset. The lesion centre coordinates are calculated based on manual labels by averaging the location of all masked pixels. More details will be introduced with each dataset soon.

BraTS 2018 As one of the biggest and most widespread open-access medical image datasets, the brain tumour segmentation dataset (BraTS) [280–282] provides MRI scans acquired from patients with various types of brain tumours, including glioblastoma and lower grade glioma. It consists of multiple MR modalities, including T1-weighted (T1), T1-weighted contrast-enhanced (T1ce), T2-weighted (T2) and fluid-attenuated inversion recovery (FLAIR) images. Although these scans are acquired from different sources and institutions, they have been pre-processed to the same original size $[155 \times 240 \times 240]$ before releasing. Because this chapter focuses on single-image super-resolution, only one sequence is used in this simulation experiment. I randomly select the tumour slices from 163 T1ce MR scans on the axial plane (i.e. horizontal plane). Afterwards, these $N = 11927$ images are divided into training ($N = 9559$) and testing ($N = 2368$) datasets. This dataset also provides manually segmented labels (approved by experienced neuron radiologists) of the enhancing tumour (ET), the peritumoral edema (ED), and the necrotic and non-enhancing tumour core (NCR/ET). In this work, I consider the whole tumour as the region of interest (i.e. the lesion) by fusing all the labels. All slices have the original shape of $(240, 240)$, and the ROI size is set as $(120, 120)$ to cover tumours as much as possible.

Late Gadolinium Enhancement CMR Late gadolinium-enhanced (LGE) cardiovascular MR (CMR) in patients with atrial fibrillation (AF) can show native and post-ablation treatment scar within the left atrium (LA) [283]. Although many studies have shown promising results, there are still ongoing concerns regarding the accuracy of identifying

scars using this technique [284, 285]. This is partially because the LA wall is very thin, and the limited spatial resolution of the LGE CMR can lower its diagnostic value. The acquisition durations for 3D LGE imaging are long (typically 5–10 minutes) and increasing the acquired spatial resolution is not usually practical. Instead, super-resolution (SR) based post-processing has the potential to provide an inexpensive yet effective way to increase the spatial resolution of the LGE data.

This is a clinic dataset provided by my collaborator from Imperial College. With ethical approval, CMR data were collected from 20 patients presenting with longstanding persistent AF on a Siemens Magnetom Avanto 1.5T scanner. Transverse navigator-gated 3D LGE CMR [286, 287] was performed using an inversion prepared segmented gradient echo sequence ($(1.4\text{–}1.5) \times (1.4\text{–}1.5) \times 4\text{mm}^3$ reconstructed into $(0.7\text{–}0.75) \times (0.7\text{–}0.75) \times 2\text{mm}^3$) 15 minutes after gadolinium administration (Gadovist—gadobutrol, 0.1mmol/kg body weight) [288]. A dynamic inversion time (TI) was designed to null the signal from normal myocardium [289]. The 3D LGE data were acquired during free-breathing using CLAWS respiratory motion control to increase respiratory efficiency [284]. Navigator artefact in the LA was reduced by introducing a navigator-restore delay of 100 ms [286]. In this work, I select the 2D slices ($N = 743$) on the axial plane from these 3D LGE scans and randomly divide them into training ($N = 615$) and independent testing ($N = 128$) groups. The original shape of each slice is $[512 \times 512]$, while the ROI (i.e. lesion) is defined as a $[160 \times 120]$ box with left-atrium inside. Experienced radiologists manually annotate the labels of the left atrium.

Diffusion Tensor CMR Diffusion tensor cardiovascular MR (DT-CMR) is an emerging contrast-free non-invasive technique providing rich information on myocardial microstructure [290, 291]. Despite great efforts to drive DT-CMR towards a clinical utility, it is still limited by the low spatial resolution [292, 293]. Here, I use super-resolution-based post-processing to provide a low-cost but effective way to boost the spatial resolution of DT-CMR data retrospectively.

My collaborator from Imperial College also provides this DT-CMR dataset. With ethical approval, short-axis DT-CMR data were collected on a Siemens Skyra 3T scanner. All DT-CMR data were acquired at peak-systole ($N = 133$) or in diastasis ($N = 115$) in healthy volunteers, using a breath hold STEAM-EPI sequence with diffusion encoded over one complete cardiac cycle [292]. The acquired spatial resolution was $2.8 \times 2.8 \times \text{mm}^2$, $1.4 \times 1.4 \times \text{mm}^2$ reconstructed, with 8mm slice thickness, repetition time 2 cardiac cycles, echo time 23–25ms, with SENSE factor of 2. During acquisition, slices of 6 diffusion directions at one coordinate are obtained with various diffusion weightings ranging from 150

to $600s/\text{mm}^2$. These slices are one case for clinic diagnosis and parameter map calculation. Each case may contain 8 to 10 images because of duplicate acquisitions on the same diffusion direction. In this super-resolution research, I randomly select $N = 208$ cases for training and another $N = 40$ for testing (including 20 diastole cases and 20 systole cases). The original shapes of these slices vary from 160 to 256, but the ROI is set as $[80 \times 80]$ with the vessel inside. Experienced radiologists manually annotate labels of vessel walls.

3.3.2 Evaluation protocol

The proposed method MS-GAN is compared with bilinear interpolation, SRResNet and SRGAN [114]. Conventional Peak signal-to-noise ratio (PSNR) and Structural SIMilarity (SSIM) index are used to measure the pixel-wise and image-wise similarity between generated SR results and ground truth HR images. A mean opinion score (MOS) evaluation is designed and performed quantify the perceptual reality of generated SR images. Ground truth slices and SR images generated by different methods are shuffled for blind scoring by an experienced MR physicist. The evaluation is based on a Likert-type scale—0 (non-diagnostic), 1 (poor), 2 (fair), 3 (good), and 4 (excellent)—depending on the image qualities [165, 166]: over-**S**mooth; motion and other kinds of **A**rtefacts; **U**nrealistic textures; and too **N**oisy or low SNR. The MOS is then derived by calculating each method's mean and standard deviation.

3.3.3 Implementation details

All the implementation, including MS-GAN and the comparison methods, has been done using Python 3.5, with TensorFlow [36] library. OpenCV-python [294] has been used for image pre-processing, such as resize and blur operations. All experiments are performed on a Linux workstation with a single NVIDIA TITAN X Pascal GPU. The lesion detection network has 5 max pooling residual blocks, each of which has 4 convolutional layers (width $= [32, 64, 128, 128, 128]$ and kernel size $= 3 \times 3$). The three fully connected layers have $[1024, 128, 2]$ nodes, respectively. ReLU activation applies after each batch normalisation layer and the FC layer. This network is trained for 50 epochs with Adam optimiser [61] (momentum $= 0.9$, betas$= (0.9, 0.999)$, learning rate $= 0.001$). The multi-scale SR image generator and the WGAN-GP discriminator are trained jointly from scratch. The generator (Fig. 3.3) consists of two $\times 2$ SRResNet, each of which has 16 residual blocks, and each block has 64 convolution kernels (kernel size $= 3 \times 3$) in every convolutional layer. ReLU [47, 295] applies as the non-linear activation after each layer. Notice that the warm-up training of the SR image generator is not applied, although it is required in the previous GAN-based method SRGAN [114]. The networks are trained with Adam optimiser ((momentum $= 0.9$, betas$= (0.9, 0.999)$) for 300 epochs. The learning rate starts

as $\iota = 10^{-4}$ and decays to $\iota = 10^{-5}$ at the midpoint of the training.

3.4 Results and Discussion

In this section, I first present the results of MS-GAN on two private medical image datasets in Section 3.4.1. In addition to PSNR and SSIM representing the reconstruction fidelity, SR images are evaluated with the mean opinion score and clinical analysis tasks to illustrate the real perceptual quality in the clinic for doctors. In the following Section 3.4.2 and 3.4.3 simulation experiments are executed on the BraTS dataset for an ablation study of the lesion-focused training and the multi-scale GAN. As a clean and public dataset with numerous brain tumour slices, it can support the representative discussion of the impacts of each component in the proposed method, and the conclusions can be reliably extended to a wide range of medical image modalities.

3.4.1 MS-GAN performance

Performance on LGE-CMR First, in the $\times 4$ SISR experiments with the LGE-CMR dataset, the proposed method is compared with bilinear interpolation, SRResNet and SRGAN. The boxplots in Fig. 3.5 summarise the PSNR and SSIM results. On average, it achieves significantly superior scores compared with SRGAN and bilinear interpolation. SRResNet, trained without perceptual and GAN-based loss, has the best PSNR and SSIM scores. Regarding the perceptual quality of the SR results (Table. 3.1), the ground truth images and generated SR images of 30 random samples (150 images in total) are shuffled and blind-scored by an experienced CMR image processing physicist. The ground truth images naturally achieve the highest score, while Wilcoxon rank-sum test shows that the proposed MS-GAN method achieves comparable performance with the ground truth images. In contrast, the other three SR methods can only generate low perceptual quality SR images with multiple artefacts, such as over-smoothing. Fig. 3.6 shows an example, while Fig. 3.7 illustrates the error maps. The SR results of bilinear interpolation are conspicuous because of their poor perceptual quality. SRResNet aims to reduce the mean pixel-wise error but leads to over-smoothing because of the lack of high-frequency textures. SRGAN, which introduces perceptual loss and GAN-based adversarial loss to SRResNet, successfully generates rich textures to avoid blurring. However, the textures are either unreal or too noisy. In contrast, MS-GAN has successfully generated realistic textures, significantly improving the perceptual quality of SR images.

Notice that as in previous works [114, 166], this study also demonstrates that conventional quantitative metrics such as PSNR and SSIM have limitations on evaluating SR results alone. PSNR and SSIM primarily depend on global information and pixel-wise accuracy, with no attention to local structures (e.g. textures). Following the same idea, pixel-wise loss functions such as $\mathcal{L}_{MSE}$ are designed. They can lead to high PSNR/SSIM scores but

also limit the network to generate high-frequent information to avoid increasing pixel-wise errors. As a result, the trained networks, such as SRResNet, only generate over-smoothed SR images. In Table 3.1, 28 of 30 SR images of SRResNet are marked as over-smoothed. Introducing perceptual and adversarial losses in training SR image generators can solve this problem. However, it may cause new issues because incorrect textures may make the SR images unreal and too noisy, as SRGAN has performed. In the 30 SR results of SRGAN, 23, 16 and 17 images are marked with artefacts, unrealistic textures and too much noise, respectively. In contrast, the proposed MS-GAN method has tackled both issues and generated SR images with high perceptual quality, few artefacts, and good PSNR/SSIM scores.

Table 3.1: Perceptual quality is measured by the MOS metric [166]. Thirty samples are randomly selected from the testing dataset of LGE-CMR images. Each sample has one ground truth image and 4 SR results corresponding to the 4 SR methods. A CMR image processing physicist (>3 years' experience in LGE CMR) has performed blinded scoring of the image quality of these 150 images on a Likert-type scale: 1 (Poor), 2(Fair), 3(Good), and 4(Excellent)[165, 166]. Wilcoxon rank-sum test shows no significant difference between the proposed MS-GAN (grey row) and the ground truth. The best two MOS scores are in bold. S, A, U and N are MOS remarks defined as over-**S**mooth, motion and other kinds of **A**rtefacts, **U**nrealistic textures, and too **N**oisy or low SNR.

	MOS	Poor	Fair	Good	Excellent	S	A	U	N	p-value
Bilinear	1 ± 0	30	0	0	0	30	22	22	0	< 0.05
SRResNet[114]	1.93 ± 0.36	3	26	1	0	28	2	1	0	< 0.05
SRGAN[114]	1.93 ± 0.63	6	21	2	1	1	23	16	17	< 0.05
MS-GAN	$\mathbf{2.73} \pm 0.73$	0	13	12	5	1	8	7	9	$> \mathbf{0.05}$
GT	$\mathbf{2.97} \pm 0.80$	0	10	11	9	1	7	4	5	

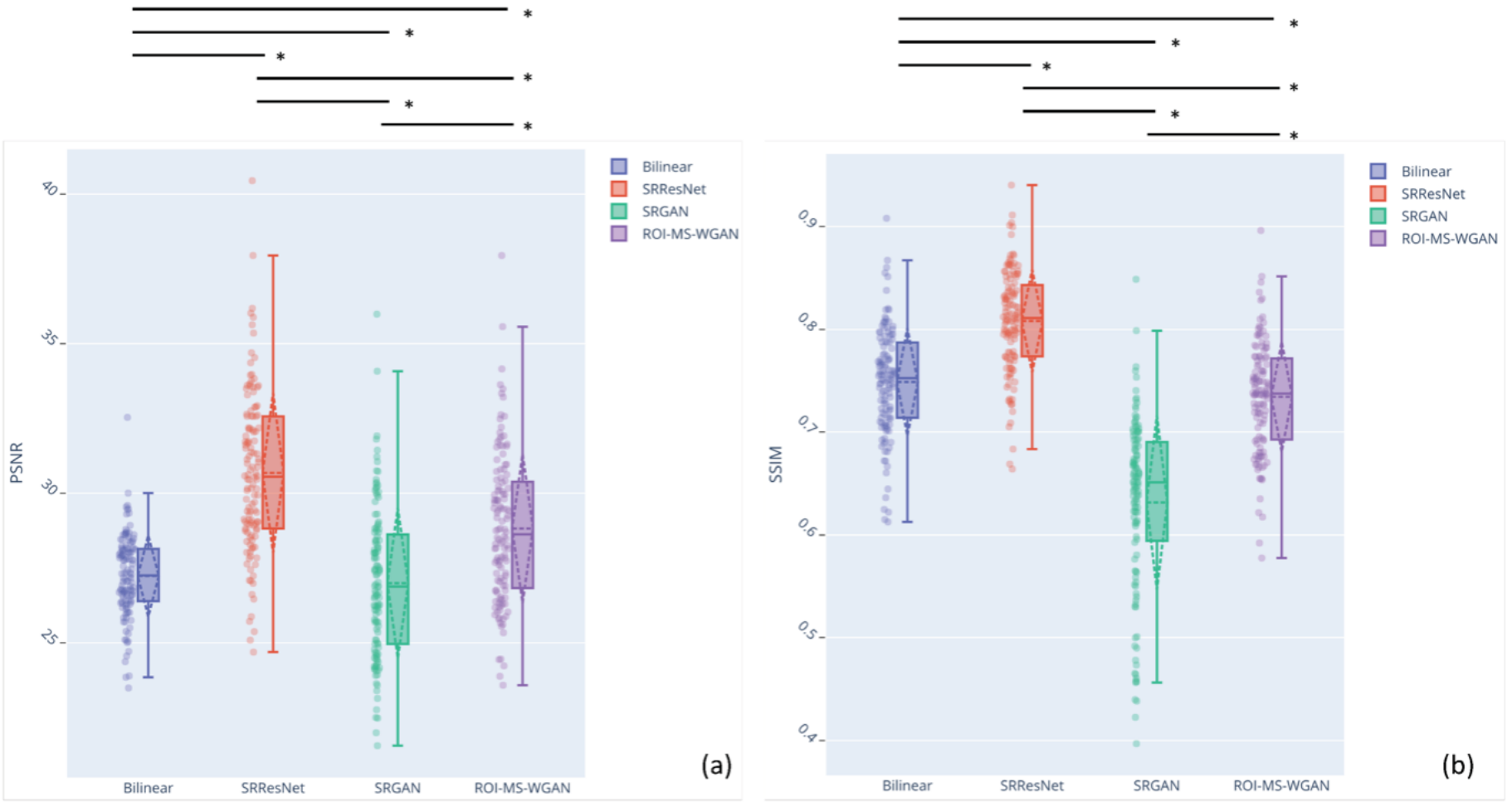


Figure 3.5: Comparison of SR methods (×4 magnification) on the LGE-CMR dataset. Boxplots comparing the calculated PSNR (a) and SSIM (b) of the lesion-focused MS-WGAN method with the bilinear interpolation, SRResNet and SRGAN. Wilcoxon rank-sum test showed that all the methods performed significantly differently. As previous studies [114, 166], this study also demonstrates that conventional quantitative metrics (PSNR/SSIM) may not be the best measure of SR results alone but need to be considered with qualitative visual inspections. Wilcoxon rank-sum shows significant statistical differences between the tested methods, e.g. MS-WGAN vs. bilinear interpolation (indicated by the *).

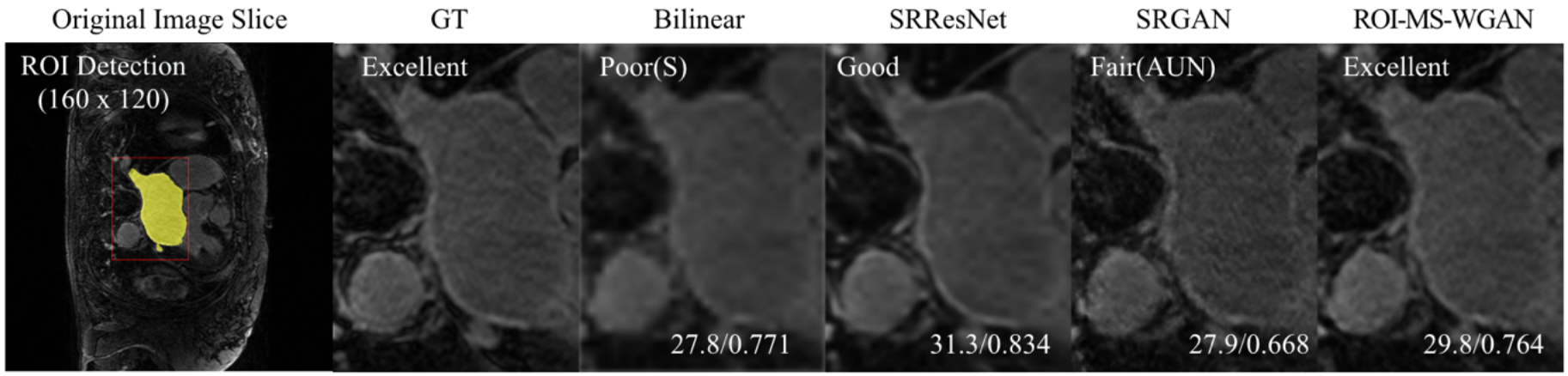


Figure 3.6: A random sample selected from LGE-CMR testing dataset. SR images generated by bilinear interpolation, SRResNet, SRGAN and the proposed ROI-MS-GAN are compared with the GT image by **PSNR** and **SSIM**. The MOS and remarks were based on blinded scoring of the images by a CMR image processing physicist: over-**S**mooth; motion and other kinds of **A**rtefacts; **U**nrealistic textures; and too **N**oisy or low SNR. The left atrium is labelled, and the lesion detection result is also presented.

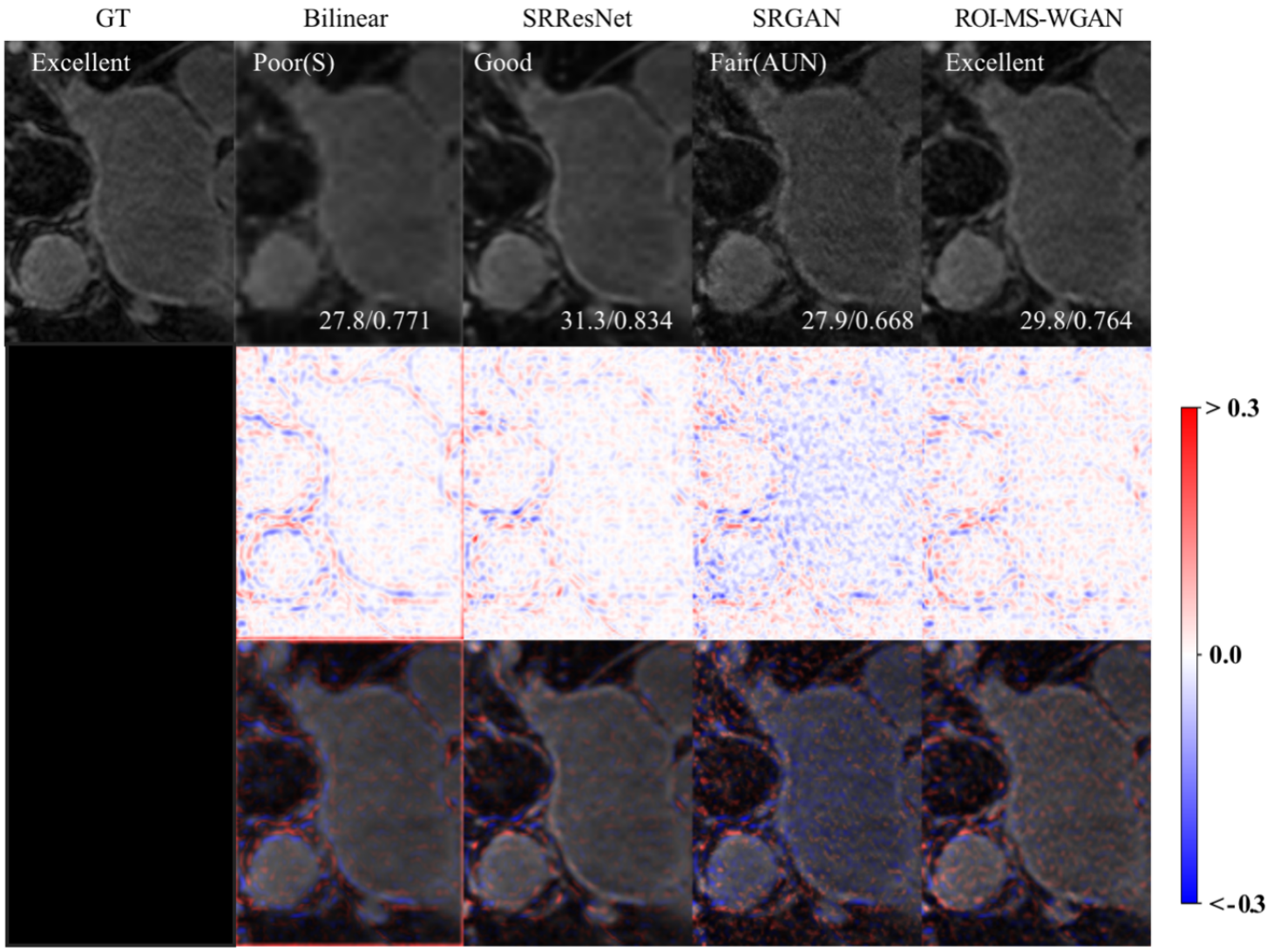


Figure 3.7: Comparison of SR methods in $\times 4$ tasks. This sample is randomly selected from the testing dataset of LGE-CMR images. All images (the top row) are converted into $[-1.0, 1.0]$. The pixel-wise error maps (the middle row) between the generated SR images and the GT image are clipped into a range of $[-0.3, 0.3]$ and projected on SR images (the bottom row) for visualisation. The MOS is presented with remarks: over-**S**mooth; motion and other kinds of **A**rtefacts; **U**nrealistic textures; and too **N**oisy or low SNR. PSNR and SSIM are also displayed.

Performance on DT-CMR The proposed method obtains promising SR results in the experiments with DT-CMR images (Fig. 3.9). Bilinear interpolation and SRResNet show over-smoothed results as expected (Fig. 3.8). SRGAN achieves perceptually acceptable results, but the difference in image and PSNR/SSIM demonstrate significant errors. The mean PSNR and SSIM scores of all slices are calculated for each case in the testing dataset. On average, SRResNet achieves the best PSNR and SSIM scores for diastole and systole cases, closely followed by the lesion-focused MS-GAN. Furthermore, the SR results are measured in the following clinical tasks. Four DT-CMR parameter maps are calculated and compared with ground truth with pixel-wise root-mean-square-error (RMSE): mean diffusivity (MD), fractional anisotropy (FA), helix angle (HA) and secondary eigenvector angulation (E2A). Although for MD, FA and E2A, bilinear interpolation achieves relatively low RMSE, the inter-subject mean transmural HA line profile extracted shows that the results from SRResNet, SRGAN and MS-WGAN are more realistic (Fig. 3.10). Furthermore, the DT-CMR parameter maps suggest that the results obtained by SRResNet are over-smoothed (Fig. 3.11).

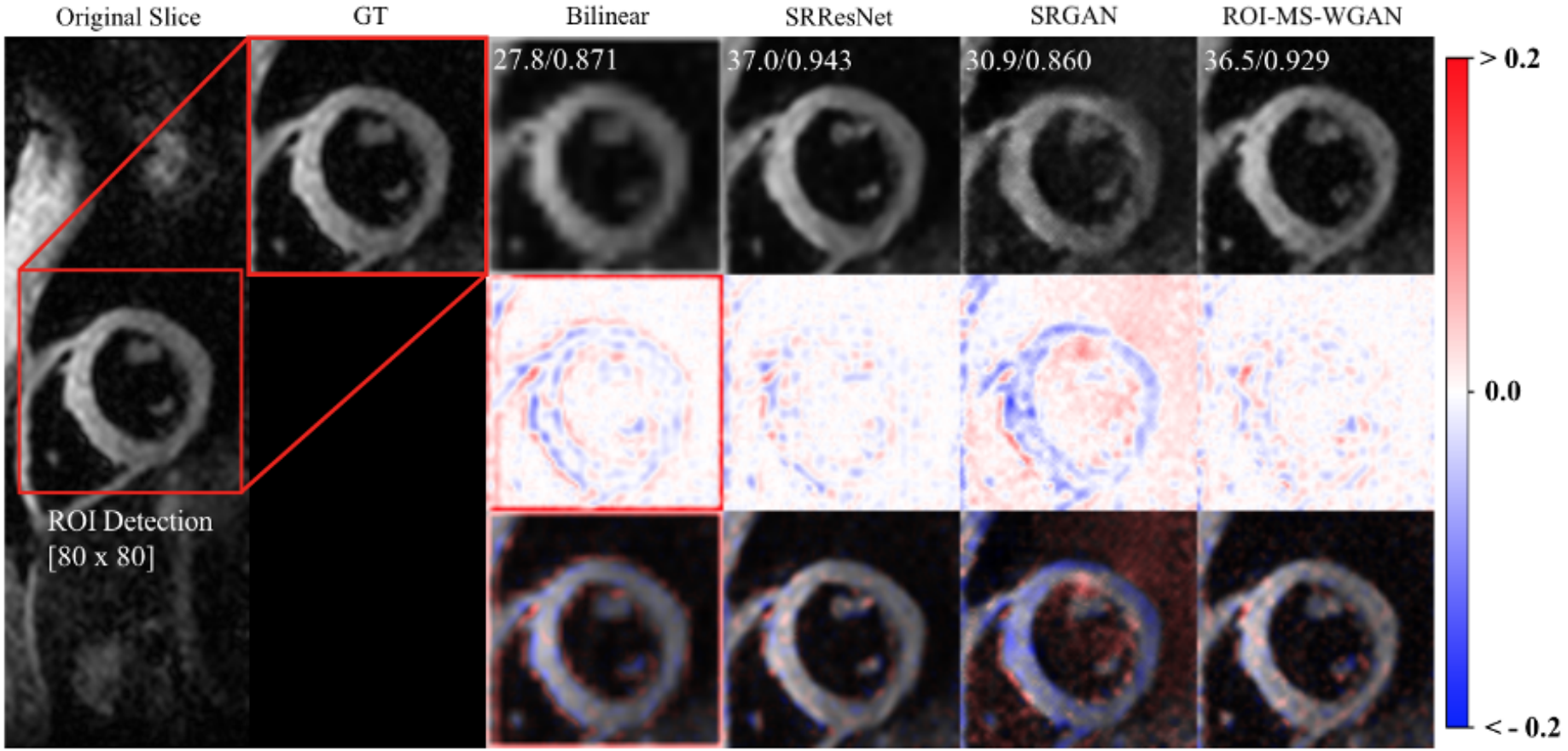


Figure 3.8: Comparison of SR methods in $\times 4$ magnification with a random sample of DT-CMR testing images. Two left columns show the results of ROI detection. The SR results (top row) and the error maps (middle and bottom row) are displayed on the right. All images are normalised to $[-1., 1.]$, and the image differences are scaled into a range of $[-0.2, 0.2]$ for visualisation. PSNR/SSIM are also displayed.

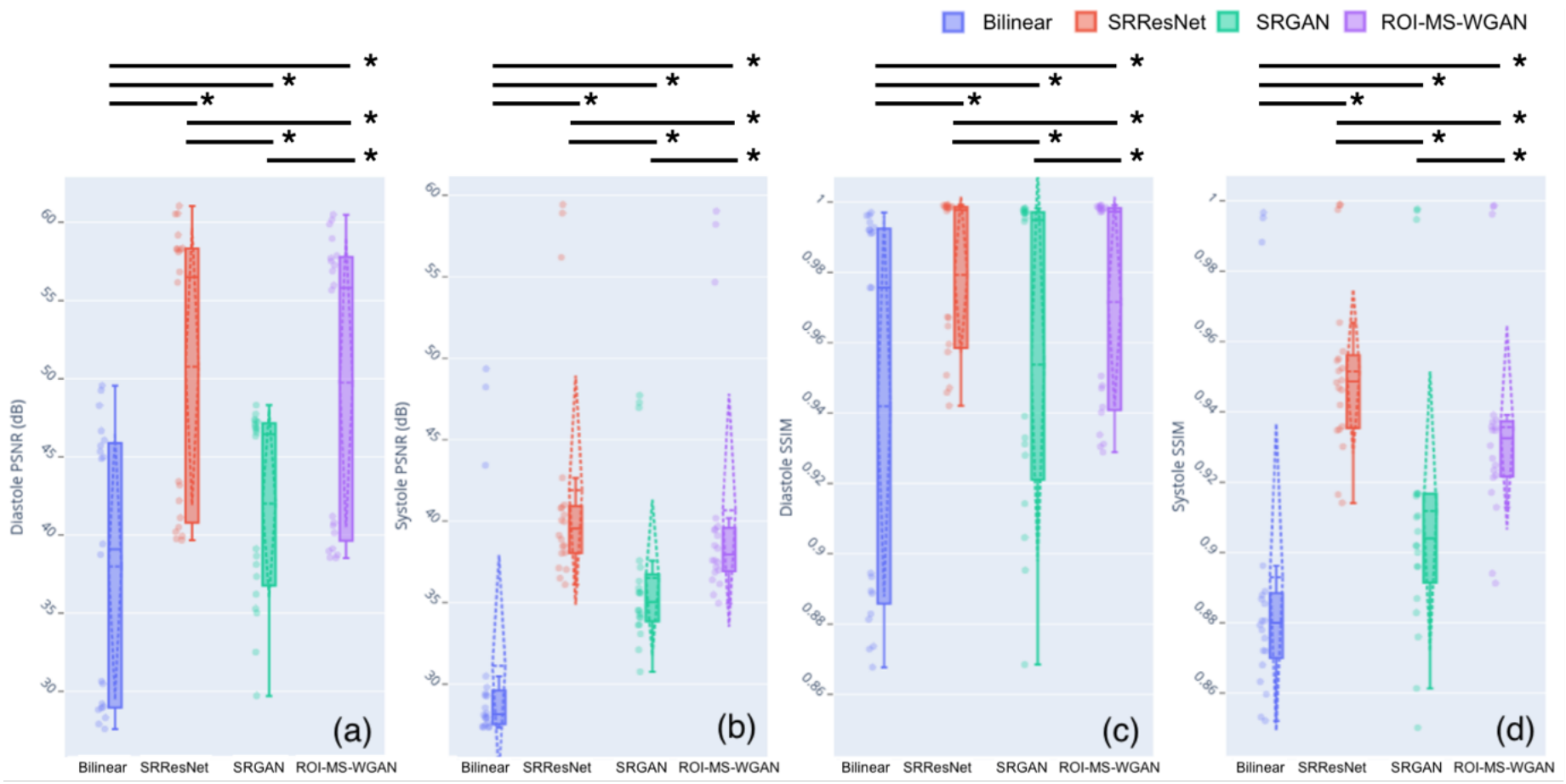


Figure 3.9: Case-wise mean PSNR and SSIM of the proposed lesion-focused MS-WGAN, comparing with bilinear interpolation, SRResNet and SRGAN: (a) PSNRs of 20 diastole independent testing cases; (b) PSNRs of 20 systole indepent testing cases; (c) SSIMs of 20 diastole independent testing cases; (d) SSIMs of 20 systole independent testing cases. Wicoxon rank-sum test shows significant statistical differences between the tested methods (indicated by the *).

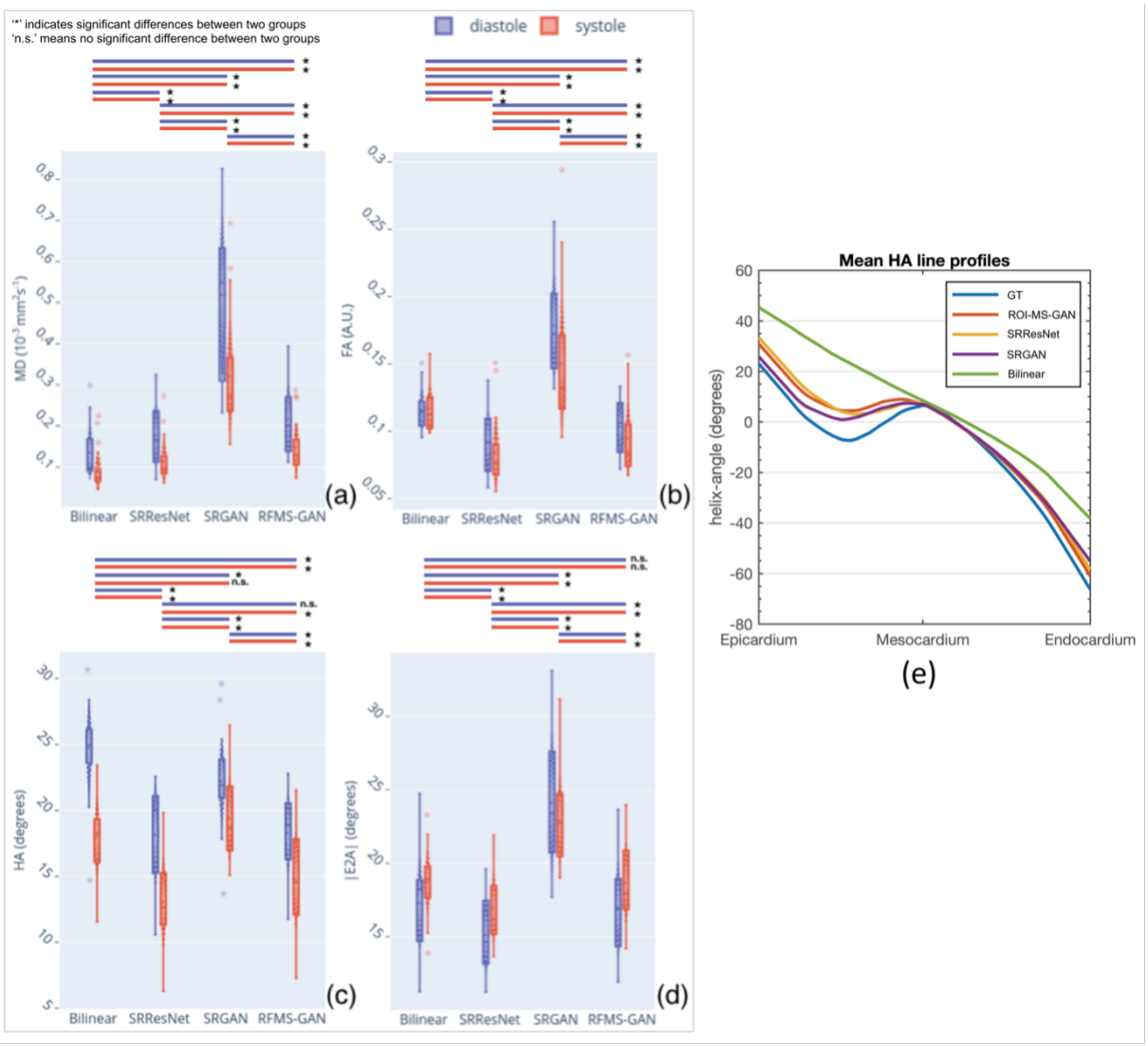


Figure 3.10: Pixel-wise root-mean-square errors of the DT-CMR parameter maps are on the left: (a) mean diffusivity (MD), (b) fractional anisotropy (FA), (c) helix angle (HA) and (d) secondary eigenvector angulation (E2A). On the right is (e) the inter-subject septal mean HA line profiles for each SR method and the ground truth (GT). Notice that the dip close to the epicardium is due to warpping of the helix angles as it approaches the right ventricle for some of the subjects. All methods, except bilinear interpolation, keep this structural information. The statistical differences are given by Wilcoxon rank-sum test (* indicates significant differences while n.s. means no significant difference.

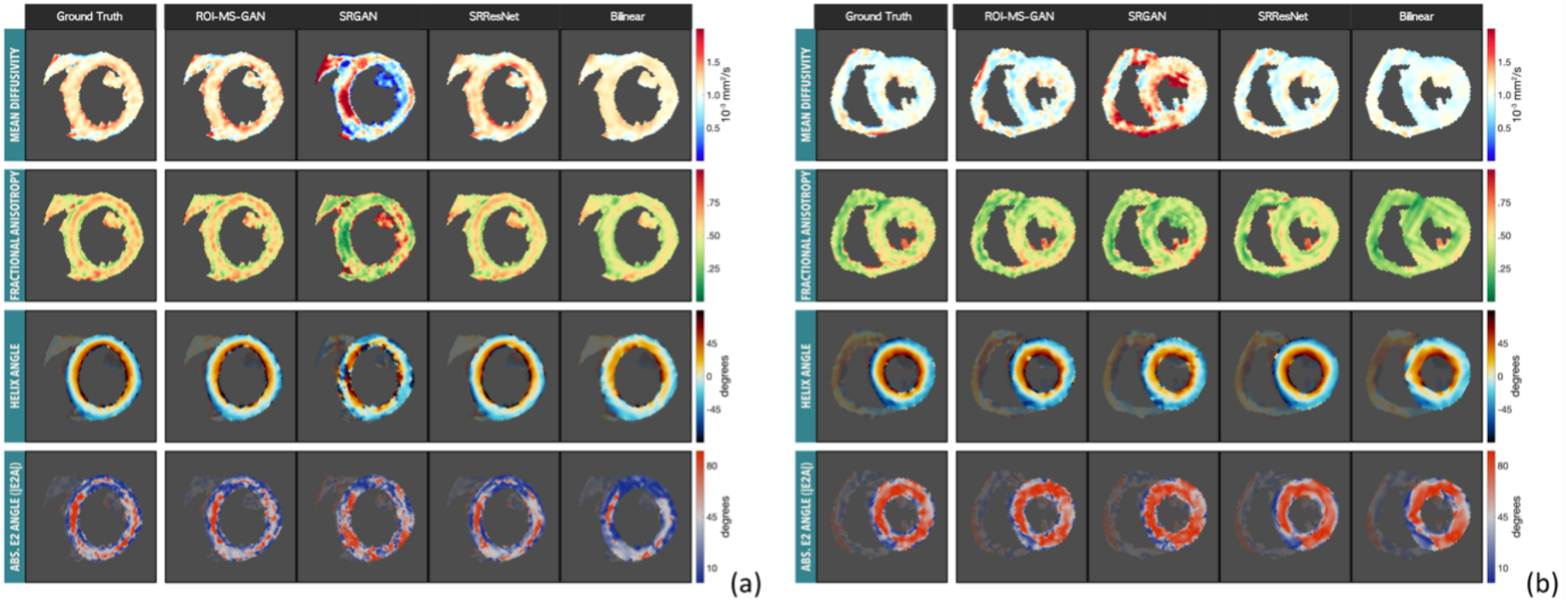


Figure 3.11: Calculated DT-CMR parameter maps for each tested method and compared them to the ground truth: (a) an example diastole case and (b) an example systole case. From top to bottom: mean diffusivity (MD), fractional anisotropy (FA), helix angle (HA) and secondary eigenvector angulation (E2A).

3.4.2 Impacts of the lesion-focused training

Experiments are designed to research how the lesion-focused training strategy affects the final SR performance. For a simple and fair comparison, I implement lesion-focused super-resolution (LFSR) by adding lesion detection and ROI-focused training to the original SRGAN. Methods are compared in both $\times 2$ and $\times 4$ SR tasks with the clean dataset BraTS of brain MR images for a general discussion of the impacts. Meanwhile, to simulate the noise in the acquisition process of MR scans, various levels of additive white Gaussian noise (AWGN) are added in $\times 2$ SR tasks. Bilinear interpolation (with non-local mean denoising [296] if necessary) and SRResNet are also considered baseline SR methods.

Lesion detection accuracy Regarding the lesion detection performance, three levels of ROI detection accuracy are defined depending on the percentage of tumours covered by the predicted bounding box (Fig. 3.12). First, perfect detection means the tumour is 100% covered by the detected ROI. Second, the acceptable detection denotes that more than 95% of the tumour is covered. Finally, the other cases are called bad detection. In the brain tumour MR images experiments, the lesion detection network LD is trained with HR images from the training dataset but works on LR images from the testing dataset. It has achieved high accuracy on average of 2368 testing images in both $\times 2$ and $\times 4$ SR tasks (Table. 3.2). For $\times 2$ down-sampled LR images, it has perfectly predicted the ROIs of 2218 (93.7%) slices. For $\times 4$ down-sampled LR images, it has perfectly predicted the ROIs of 2109 (89.1%) slices. For both cases, the acceptable predictions are 111 (4.7%) and 119 (5.0%) cases, respectively, while the bad predictions are 39 (1.6%) and 140 (5.9%) cases, respectively. Notice that lesion detection is not the core goal of this project, so the occasional errors are acceptable. Correct lesion detection results help address the most informative patches during the SR network training, improving the final performance.

Table 3.2: ROI detection accuracy on LR images. Depending on the percentage of tumours covered by the predicted ROI bounding box, three accuracy levels are defined: **perfect** (100%), **acceptable** ($> 95\%$) and **bad** ($< 95\%$). The testing dataset includes 2368 brain tumour MR images in total.

	Perfect Detection	Acceptable Detection	Bad Detection
$\times 2$	2218 (93.7%)	111 (4.7%)	39 (1.6%)
$\times 4$	2109 (89.1%)	119 (5.0%)	140 (5.9%)

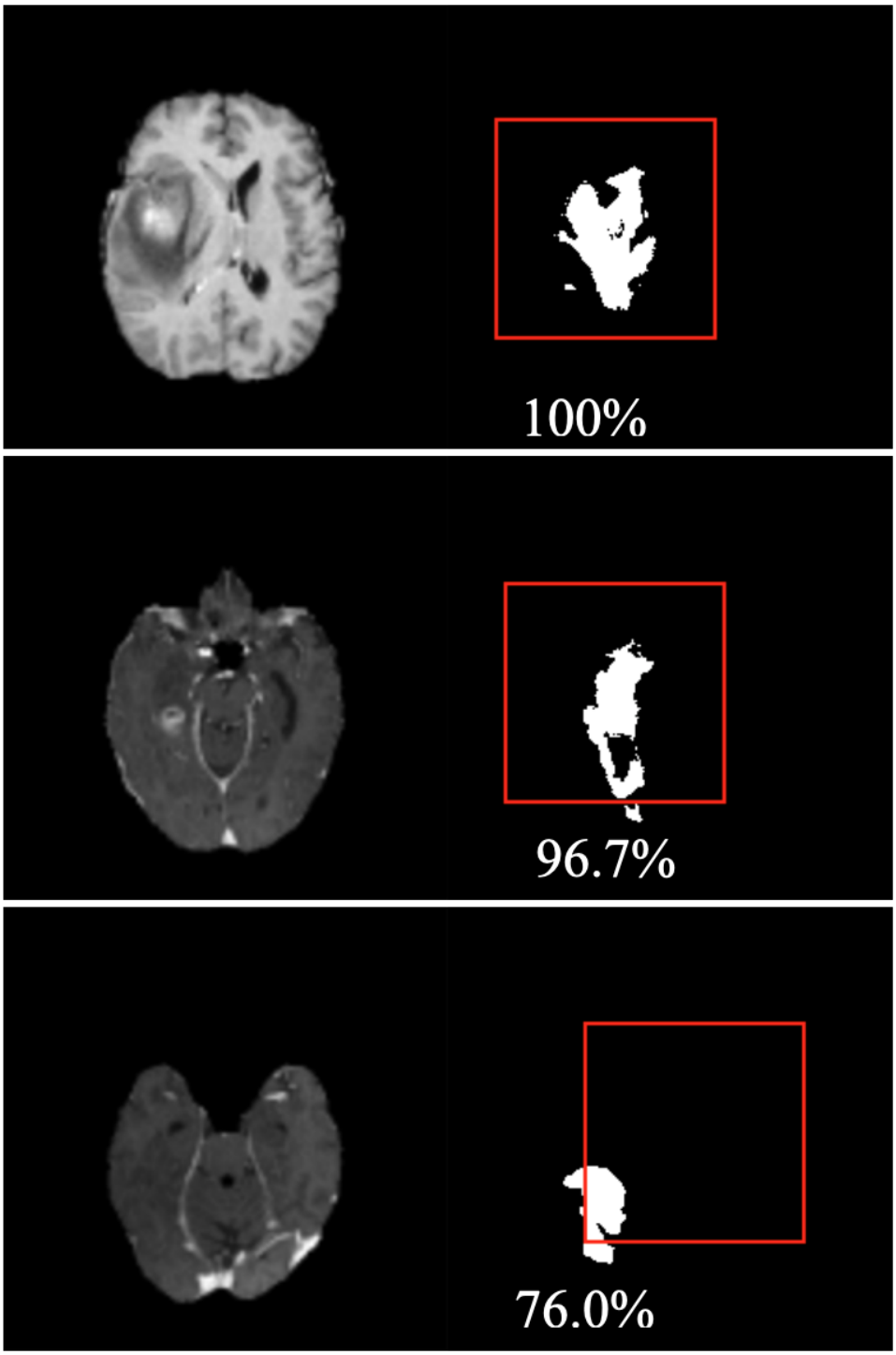


Figure 3.12: Examples of ROI detection results by *LD* with brain tumour MR images. From top to bottom: a perfect detection is defined as the tumour is 100% covered by the detected bounding box; an acceptable detection is defined as more than 95% of the tumour is covered; and other cases are defined as bad detection.

Improvements with lesion-focused training In $\times 2$ and $\times 4$ SR tasks, LFSR generates more perceptual realistic textures than the original SRGAN (Fig. 3.15). Both GAN-based methods avoid the over-smoothing issue in bilinear interpolation and SRResNet. Compared with SRGAN, the lesion-focused method achieves superior (of the $\times 2$ cases) and equivalent (of the $\times 4$ cases) PSNR and SSIM performance on average (Fig. 3.13). Although SRResNet obtains the highest scores, it fails to generate authentic texture-rich SR images. In $\times 2$ SR with denoising tasks, the lesion-focused strategy significantly increases the PSNR and SSIM scores compared with the original SRGAN (Fig. 3.14) because the noise in the background is not involved in training anymore. In $\times 2$ SR w/wo noise tasks, lesion-focused training results in improvements of $[+2.5dB, +1.4dB]$ PSNR and $[+0.049, +0.043]$ SSIM scores. Although lesion-focused training slightly decreases performance ($-0.6dB$ PSNR and -0.018 SSIM) in $\times 4$ SR, it improves the results ($+0.8dB$ PSNR and $+0.041$ SSIM) in more challenging cases with additive noise. Fig. 3.16 shows the noisy LR image, ground truth image and SR images of a random example. Perceptually the results of bilinear-interpolation with non-local mean denoising and SRResNet have lost almost all textures inside the brain and the tumour. In summary, SRResNet achieves the best PSNR and SSIM performance in all tasks (Table. 3.3) but fails to generate realistic textures in SR images. In contrast, GAN-based methods improve the perceptual quality of SR images. Moreover, the lesion-focused training strategy performs high perceptual quality and results in much better reconstruction fidelity than the original SRGAN.

Table 3.3: An ablation study of the lesion-focused training strategy in simulation $\times 2$ and $\times 4$ SR experiments w/wo additive Gaussian noise ($\sigma = 20, 40$). SRGAN with lesion-focused training strategy (LFSR, in grey) is compared with (a) bilinear interpolation plus non-local means [296] method (B+NLD), SRResNet, and original SRGAN [114].

	$\times 2$		$\times 2(\sigma_g = 20)$		$\times 4$		$\times 4(\sigma_g = 40)$	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
B+NLD [296]	29.1	0.900	20.7	0.623	25.2	0.761	17.1	0.483
SRResNet[114]	35.6	0.962	32.8	0.895	27.9	0.832	30.6	0.840
SRGAN[114]	29.6	0.865	27.6	0.789	25.7	0.741	26.2	0.731
LFSR	32.1	0.914	29.0	0.832	25.1	0.723	27.4	0.772

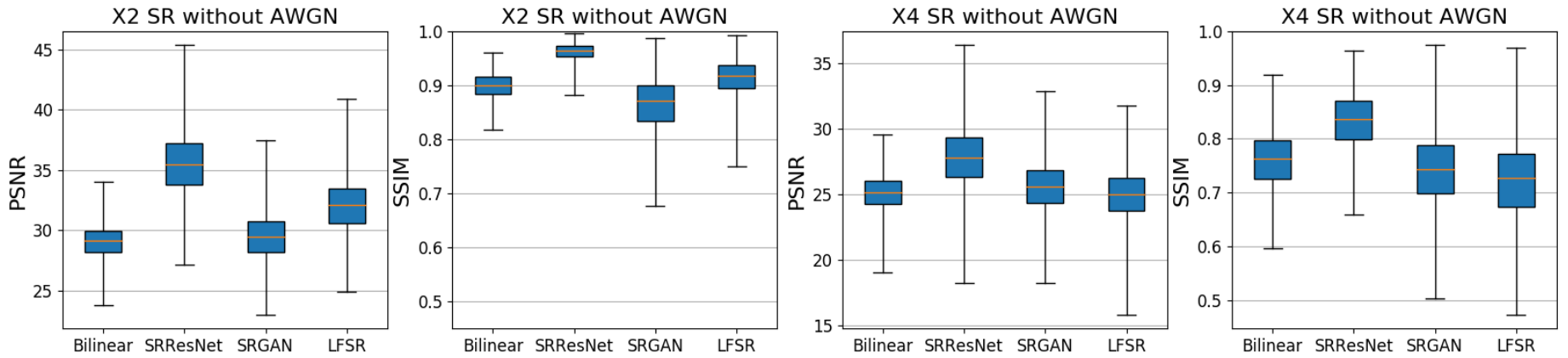


Figure 3.13: An ablation study of the lesion-focused training strategy in the $\times 2$ and $\times 4$ SR experiments on the BraTS dataset. Slice-wise PSNR and SSIM scores are illustrated in boxplots.

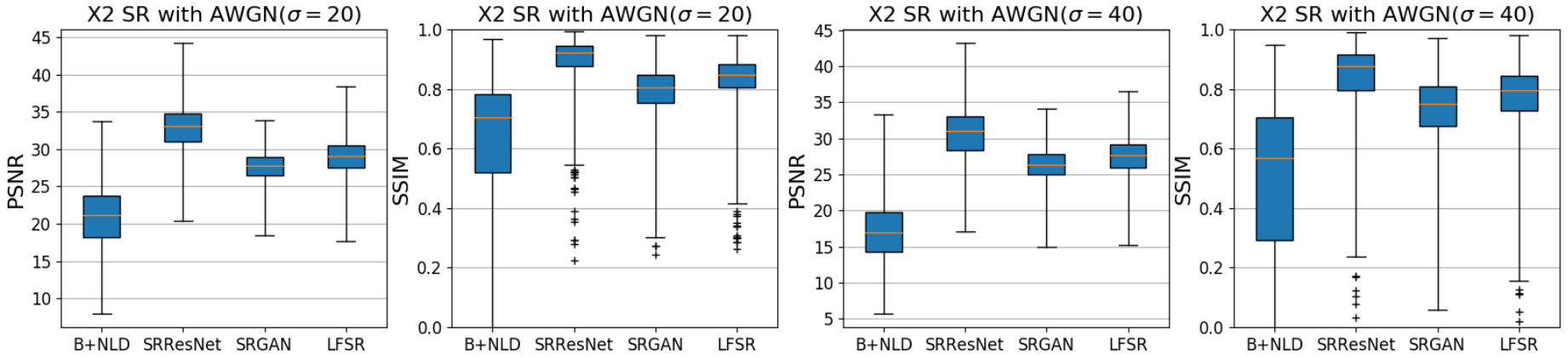


Figure 3.14: An ablation study of the lesion-focused training strategy in the $\times 2$ SR and denoising experiments on the BraTS dataset. Slice-wise PSNR and SSIM scores are illustrated in boxplots. Additive white Gaussian noises ($\sigma = 20, 40$) are applied in $k-$space of LR images.

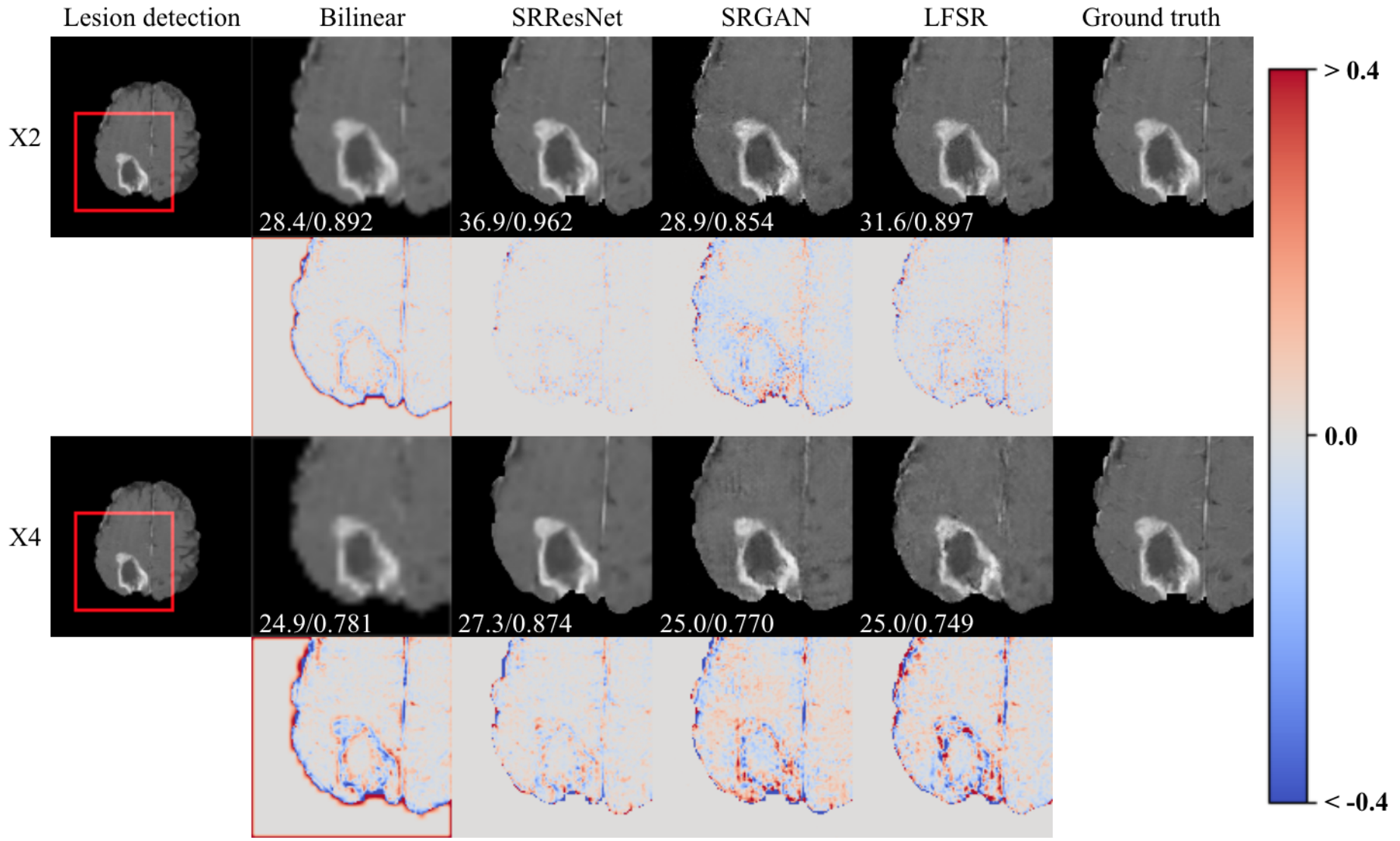


Figure 3.15: Compare lesion-focused SRGAN (LFSR) with SOTA SR methods in $\times 2$ and $\times 4$ SR tasks. All images are converted into $[-1., 1.]$, and the error maps are clipped into $[-0.4, 0.4]$ for visualisation purposes. On the left, the lesion detection results are drawn on the original slices, and PSNR and SSIM are also displayed.

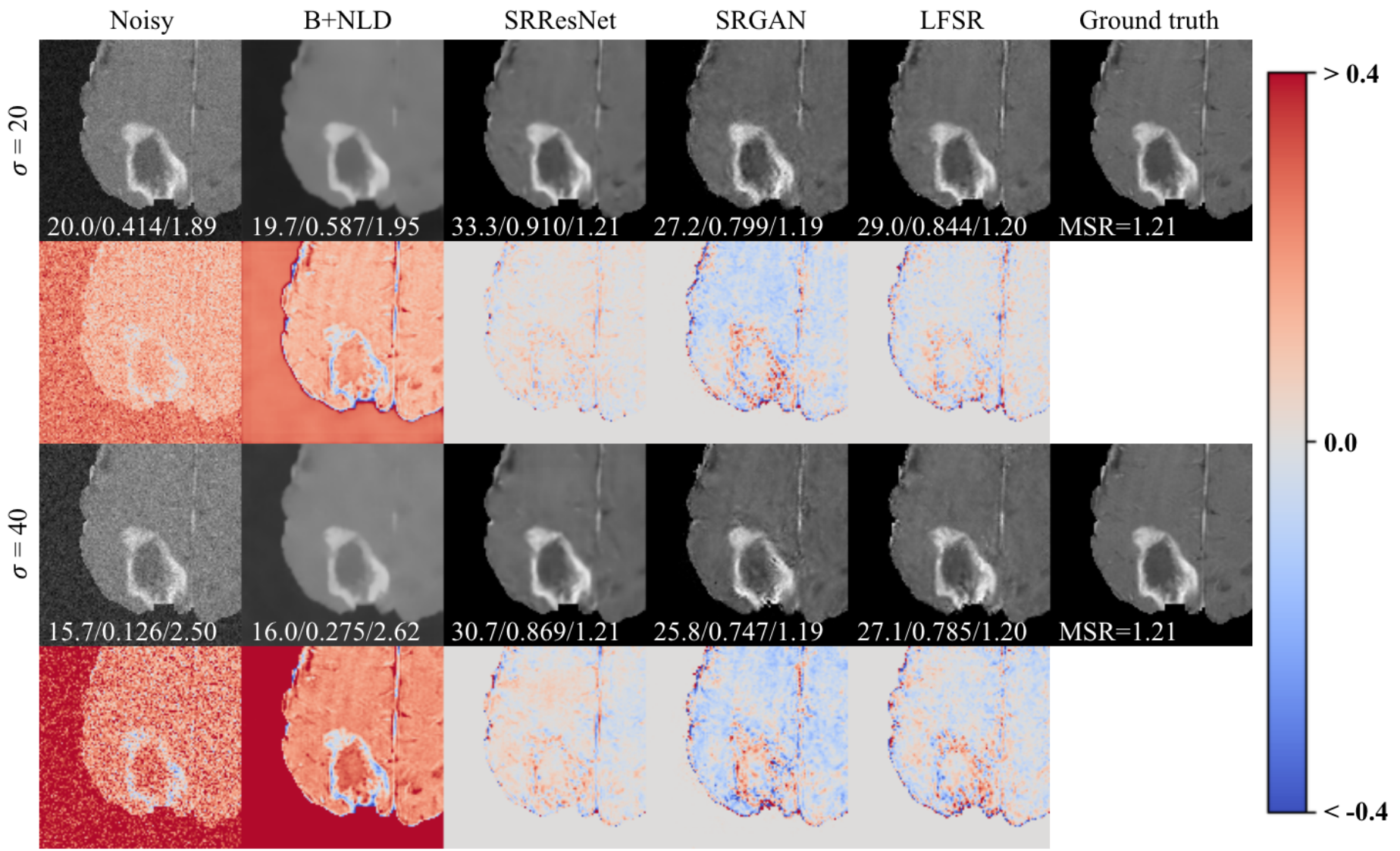


Figure 3.16: Comparing lesion-focused SRGAN (LFSR) with SOTA methods in $\times 2$ SR and denoising tasks. Additive white Gaussian noises ($\sigma = 20, 40$) are applied on LR images (the left column). All images are converted into $[-1., 1]$ and the image differences between SR results and GT images are clipped into $[-0.4, 0.4]$ for visualisation purposes. PSNR, SSIM and mean-square-regression (MSR) are also displayed.

3.4.3 Impacts of GAN variations

For a comparison study, I have implemented and tested 6 GAN-based variations, which differed in SR image generator, discriminator, loss function and training strategy. Regarding the SR image generator, I have tested SRResNet [114], the multi-scale and the stacked SRResNet in Fig. 3.11. The vanilla GAN [101], Wasserstein GAN [108] and WGAN-GP [22], with and without the pre-training of the generator, are compared. Loss functions, especially the adversarial loss and the multi-/sinlg-scale MSE loss, are chosen correspondingly with the architectures of the generator and the discriminator. The implementation details of all methods are introduced in Table. 3.4, in which they are named as: (1) GAN+Pre_train; (2) WGAN+Pre_train; (3) WGAN; (4) WGAN-GP; (5) WGAN-GP $\mathcal{L}_{MSE}^{\times 2}$; and (6) MS-GAN. The SR image generators of GAN+Pre_train and WGAN+Pre_train are trained for 50 epochs before the experiments, but the other generators are not. The training of all GAN variations follows the same settings. A lesion detection network LD is trained independently and used for lesion-focused training of all GANs. Although methods based on WGAN(-GP) might converge faster than others, all tested methods are trained for the same epochs to establish a fair comparison. In addition, I involve the bilinear interpolation, SRResNet and SRGAN for a comprehensive study.

In addition to PSNR and SSIM, the mean opinion score (MOS) evaluation is used to quantify the perceptual reality of generated SR images (Table. 3.5). In this study, 100 testing slices are randomly selected for MOS evaluation. The scored group of each slice has 1 HR ground truth and 6 SR results corresponding to the 6 GAN variations. Then, I randomly shuffle these 700 images (including 100 HR ground truths). An MR physicist (>6 years experience on brain tumour MRI images) has performed blinded scoring for these shuffled images.

The vanilla GAN produced relatively poor PSNR/SSIM, but other GAN variations have resulted in similar high PSNR/SSIM. The MS-GAN method obtains the highest MOS. Figs. 3.17 and 3.18 show the qualitative visualisation of an example slice. The MS-GAN achieves high PSNR/SSIM with lesion edge and textural information preserved well. The vanilla GAN produces noisier SR results than the ground truth images. All WGAN-based models achieve similar results but are slightly smoother than the results produced by the MS-GAN. Compared with the MS-GAN, although SRResNet yielded higher PSNR/SSIM, the results are more blurry. SRGAN achieves lower PSNR/SSIM mainly due to the synthesised stripy artefacts in the SR results with less SNR. Notice that both SRResNet and SRGAN conduct on the whole slice, but only the ROIs are evaluated (Fig. 3.17). All the learning-based SR methods show significant improvement over the bilinear interpolation. Both WGAN and WGAN-GP can provide perceptually more realistic SR than the vanilla GAN, resulting

in better PSNR/SSIM and significant improvement of the MOS. The proposed MS-GAN achieves the most realistic SR with the highest MOS close to the ground truth images.

Similar to SRGAN [114], the study demonstrates the limitations of using PSNR/SSIM as evaluation metrics for medical image SR tasks. Although blurry images are not perceptually realistic enough, they can still achieve relatively high PSNR/SSIM. Comparing all the methods, SRResNet achieves the highest PSNR/SSIM. However, it smoothes out the edge and textural information of the lesion, which are valuable and crucial for clinical diagnosis.

I have also evaluated the training and inference efficiency of all methods. The generators impact training and inference costs, while the discriminators only affect the training cost. The GAN+Pre_train costs 229.6s/epoch for training and 4.04s to generate SR images for the whole testing dataset (2368 slices). According to the additional calculation of weight clipping in WGAN [108], the training time increases to 233.8s/epoch. WGAN-GP [22], because of the calculation of gradient penalty, further increases the training time to 305.7s/epoch. Moreover, generating the intermediate $\times 2$ SR result slightly slows the training process (314.3s/epoch using WGAN-GP). Finally, because the stacked SRResNet generator has the most layers, it synchronously increases the training (422.2s/epoch) and inference costs (7.75s for the whole testing dataset). In contrast, using WGAN and WGAN-GP can reduce training costs. The time-consuming 'warm-up' plays a critical role in SRGAN and GAN+Pre_train but is no longer necessary with WGAN(-GP) because they can stabilise the training much better than the vanilla GAN. Furthermore, WGAN/WGAN-GP require fewer training epochs because they converge much faster than the vanilla GAN.

Interestingly, the proposed lesion-focused MS-GAN method shows image quality improvement and signal restoration along with the SR. In Fig.3.19, it can be observed that for these two example slices, the ground truth images are with lower SNR and noticeable aliasing artefacts (thus, relatively lower MOS). The MS-GAN method can improve the image quality by boosting the SNR and reducing the artefacts resulting in better lesion characteristics (cyan arrows in Fig. 3.19). The benefits of the proposed SISR method can be envisaged for the following clinical image analysis, segmentation and bio-marker extraction and characterisation tasks.

Table 3.4: Implementation details of the GAN variations in the comparison study. The SR image generators include the SRResNet (feature maps up-sampling by two $\times 2$ sub-pixel modules), the multi-scale SRResNet (an intermediate $\times 2$ SR result $\boldsymbol{I}_{sr}^{\times 2}$ generated after the first up-sample module), and the stacked SRResNet (achieving $\times 4$ SR by two $\times 2$ SRResNet). The architectures of the discriminators are the same as in SRGAN [114] and DCGAN [110]. However, the activation layers are modified following the instructions in WGAN [108] and WGAN-GP [22] correspondingly. Pre-train denotes the 'warm up' training of the SR image generators, which helps to stabilise the training of GAN. Loss function terms $\mathcal{L}_{WGAN}$, $\mathcal{L}_{adv}$, $\mathcal{L}_{perc}$, $\mathcal{L}_2^{\times 2}$, and $\mathcal{L}_2^{\times 4}$ are defined in Section 3.2.4. $\mathcal{L}_{GAN}$ is the vanilla GAN loss and defined as [101]: $\mathcal{L}_{GAN} = -\mathbb{E}_{\boldsymbol{I}_{hr}}[log(D(\boldsymbol{I}_{hr}))] - \mathbb{E}_{\boldsymbol{I}_{lr}}[log(1 - D(G(\boldsymbol{I}_{lr})))]$.

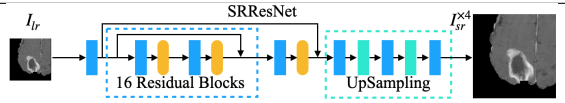
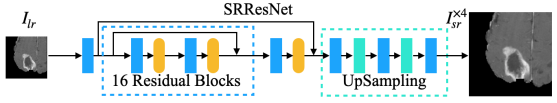
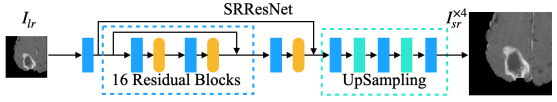
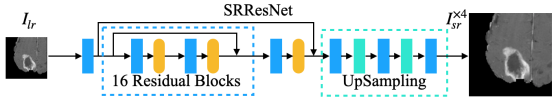
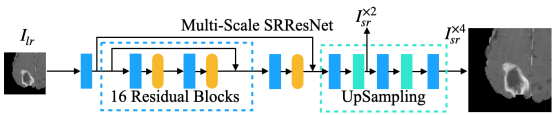
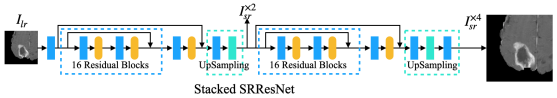
	Generator	Discriminator	Pre-Training	Loss Function
GAN+Pre_train		GAN	$\mathcal{L}_2^{\times 2}$	$\mathcal{L}_{GAN} + \mathcal{L}_{perc} + \mathcal{L}_2^{\times 4}$
WGAN+Pre_train		WGAN	$\mathcal{L}_2^{\times 2}$	$\mathcal{L}_{WGAN} + \mathcal{L}_{perc} + \mathcal{L}_2^{\times 4}$
WGAN		WGAN	X	$\mathcal{L}_{WGAN} + \mathcal{L}_{perc} + \mathcal{L}_2^{\times 4}$
WGAN-GP		WGAN-GP	X	$\mathcal{L}_{adv} + \mathcal{L}_{perc} + \mathcal{L}_2^{\times 4}$
WGAN-GP $\mathcal{L}_{MSE}^{\times 2}$		WGAN-GP	X	$\mathcal{L}_{adv} + \mathcal{L}_{perc} + \mathcal{L}_2^{\times 2} + \mathcal{L}_2^{\times 4}$
MS-GAN		WGAN-GP	X	$\mathcal{L}_{adv} + \mathcal{L}_{perc} + \mathcal{L}_2^{\times 2} + \mathcal{L}_2^{\times 4}$

Table 3.5: Quantification comparison of GAN variations using PSNR, SSIM and MOS. Pre_train means the generator has been trained for 50 epochs as a 'warm-up' before the generator and discriminator's joint training. $\mathcal{L}_{MSE}^{\times 2}$ denotes the additional intermediate $\times 2$ SR result. lesion-focused training strategy is applied for all SR methods, so only the ROIs are measured. The MOS remarks are defined as: over-**S**mooth; motion and other kinds of **A**rtefacts; **U**nrealistic textures; and too **N**oisy or low SNR. The grey row denotes the proposed method, and the best performances (except ground truth) are in bold.

	MOS	Poor	Fair	Good	Excellent	S	A	U	N	PSNR	SSIM
GAN+Pre_train	1.80±0.529	25	71	3	1	2	5	92	91	25.2±1.88	0.715±0.0670
WGAN+Pre_train	3.15±0.669	1	13	56	30	19	0	4	3	27.0±1.94	0.804±0.0498
WGAN	3.09±0.694	1	17	54	28	18	3	7	3	**27.2±1.96**	**0.806±0.0499**
WGAN-GP	3.10±0.686	2	13	58	27	3	0	14	14	26.7±1.96	0.788±0.0558
WGAN-GP+$\mathcal{L}_{MSE}^{\times 2}$	3.15±0.698	1	15	52	32	3	0	17	17	26.2±2.00	0.786±0.0577
MS-GAN	**3.26±0.626**	1	7	57	35	5	1	11	8	26.7±1.97	0.789±0.0554
Ground Truth	3.28±0.838	3	16	31	50	1	20	14	9		

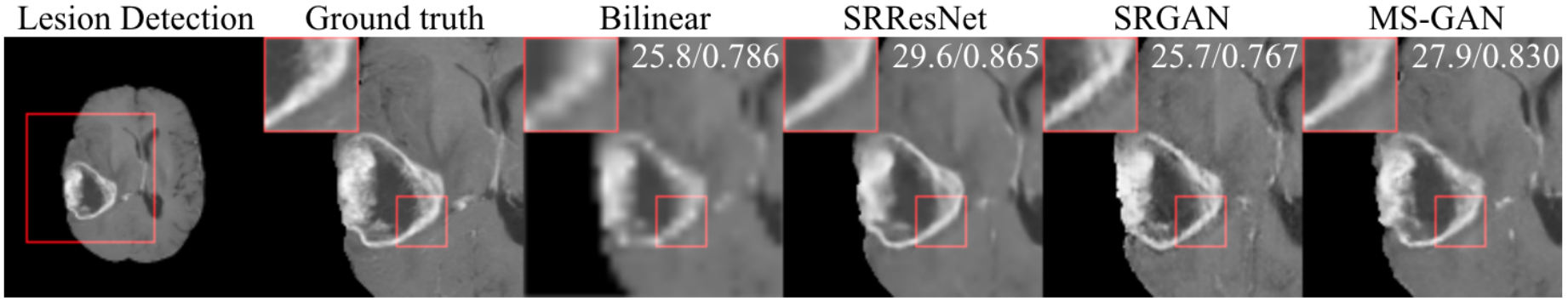


Figure 3.17: Comparison of SR methods in $\times 4$ SR task with brain tumour MR image. The ROI is drawn on the left. Part of the boundary of the tumour is zoomed in to show more details in the ground truth image and SR results. PSNR and SSIM are also displayed.

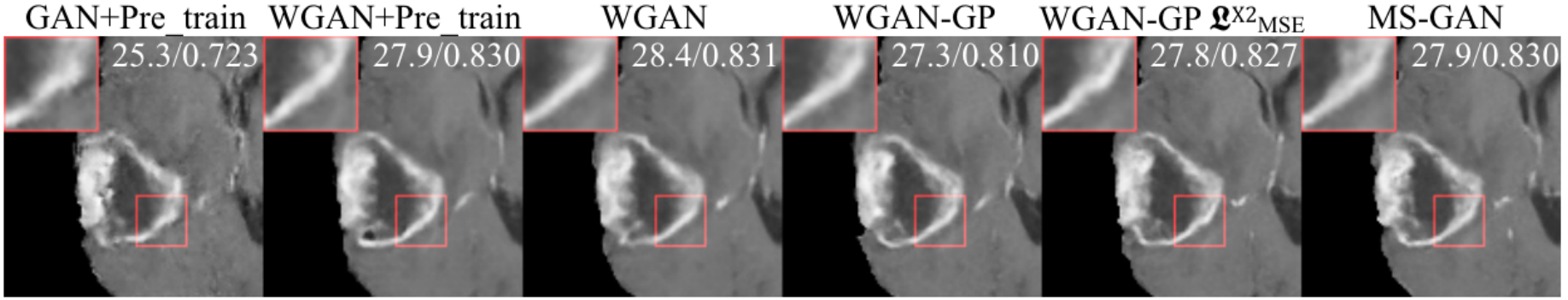


Figure 3.18: Comparison of GAN variations in $\times 4$ SR task with brain tumour MR images. Pre_train denotes that the SR image generator has been pre-trained for 'warm-up' before the generator and discriminator's joint training. $\mathcal{L}^{\times 2}_{MSE}$ means that extra output layer for intermediate $\times 2$ SR images and the corresponding $\mathcal{L}_2$ loss are applied. The ROI is drawn on the left. Parts of the images are zoomed in to show the tumour boundary with more detail. PSNR and SSIM are also displayed.

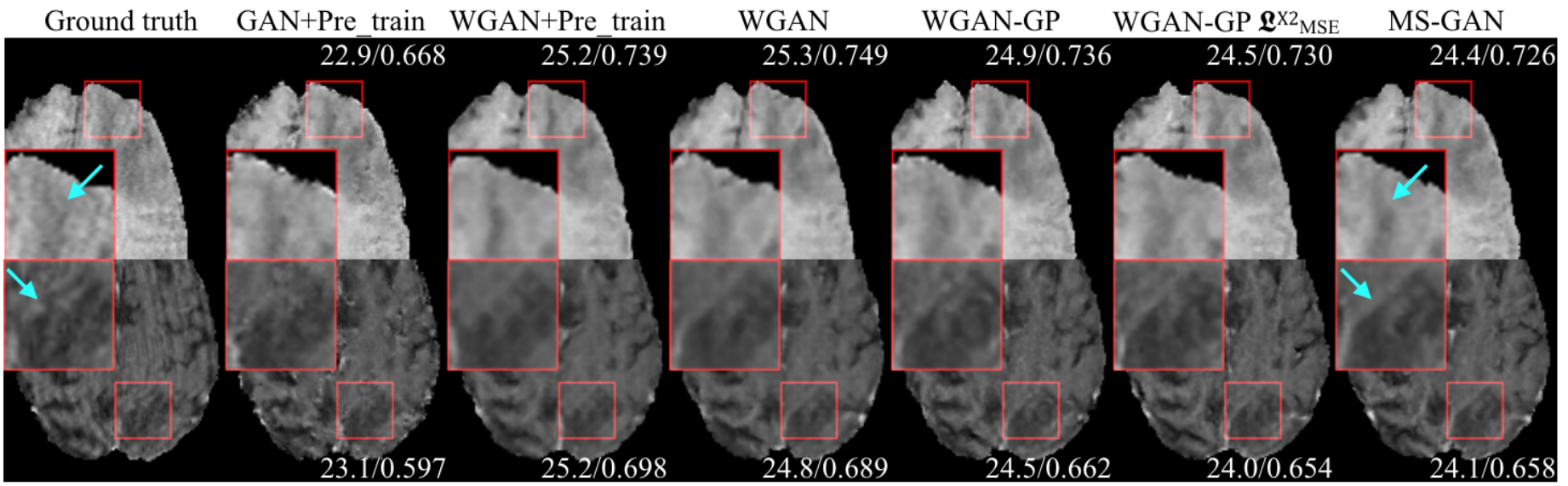


Figure 3.19: GAN-based methods can remove the artefacts in ground truth images with poor image quality. Furthermore, the method MS-GAN can enhance the edges and textures of tumour regions. PSNR and SSIM of each result are also displayed.

3.4.4 Limitations

Although the proposed method has achieved good performance on various MR image datasets, there are certain limitations to be resolved in the future.

There is a gap between simulation experiments and clinical needs when processing medical images. The current method involves processing images slice by slice, which can lead to potential risks as scans are obtained and processed patient-wise in the clinic. Furthermore, dividing the slides of the same scans into training and testing groups may cause network over-fitting, as the similarities between the slides will push the networks only to remember specific images rather than learning the LR-to-HR reconstruction. Thus, the slice-wise experiment cannot accurately represent the SR performance for real applications. Therefore, all experiments will be conducted patient-wise in my upcoming research to address these challenges.

Additionally, lesion detection may limit the general applicability of the proposed method because the training of SR networks requires extra labels of lesions. Although the lesion-focused training strategy has improved the final performance, it cannot be applied to datasets without lesion labels. Meanwhile, the lesion areas may be relatively small or hard to distinguish from other regions, such as in microbleed and stroke detection scans. Thus, I explore the more robust and flexible way of supervising SR networks with the pre-trained segmentation models of public datasets in Chapter 5.

Furthermore, assessing the perceptual quality of generated images remains an ongoing challenge. While the mean-opinion-score is a dependable metric, it is also time-consuming and not widely applicable. During the experiments in DT-CMR images, I seek to evaluate the SR outcomes using downstream medical image analysis tasks. Nevertheless, these parameter maps are specific to this MR modality, and the error maps are unsuitable for quantitative measurement. Thus, I will present my efforts towards automated, robust and quantitative perceptual quality assessment in the upcoming chapters.

Finally, Understanding why deep neural networks perform well on medical image SR tasks is consequential. On the one hand, conducting more ablation studies can help compare various network implementations and identify the impact of small changes in detailed modules. For instance, removing batch normalisation [133] or adding dropout operations [86] can significantly help in general or specific tasks. Discussing other fundamental components, such as padding operations and activation layers, can be worthwhile. On the other hand, network interpretability is essential in healthcare research and applications. However, it is challenging due to the difficulty in explaining and controlling the impact

of each operation in the layers and nodes. In the following chapters, I will present more precise ablation and comparison studies on network architecture and loss functions to uncover the critical factors behind advanced SR networks.

3.5 Chapter summary

This chapter presents a GAN-based SISR method with large magnification scales for medical images. I propose a multi-scale WGAN-GP method with a lesion-focused training strategy, which successfully enhances spatial resolution without introducing unrealistic textures in the simulation experiments with one open access and two clinical medical image datasets. The merits of this work are three-fold: (1) the lesion-focused super-resolution is developed to constrain the deep network to focus on the lesion ROIs, which does not only imitate the clinicians' scrutinising procedure, e.g., enlarge the ROIs, but also dramatically reduce the possible synthesised artefacts from the organs beyond the lesion areas; (2) a comparison study is carried out to test vanilla GAN with WGAN variants to seek possible better GAN-based solutions for a more stabilised and efficient training that can yield an improved perceptual quality for the super-resolved results; (3) based on the promises of LFSR and more advanced GAN architectures, a novel MS-GAN model is developed to tackle the challenges of SISR for medical images, especially for the more tricky cases with X4 magnification. In addition to the widely used quantitative metrics (PSNR/SSIM), the MOS is designed to incorporate experts' domain knowledge to evaluate the medical image SR results. Results have shown that the proposed lesion-focused multi-scale SISR method (MS-GAN) can achieve efficient SISR for various MR scans and potentially benefit other medical image modalities. Such models can be envisaged in a broader range of clinical applications.

CHAPTER 4

GANs with Meta-learning for Arbitrary Scale Super-Resolution

4.1 Introduction

CNN-based and GAN-based super-resolution methods have performed remarkably on various medical image modalities [23, 24, 97, 184]. However, most of these works are designed for specific magnification scales and treat SR with different scales as independent tasks. Thus, several models must be trained and stored for different magnification tasks. Furthermore, collecting large clinical datasets of high- and low-resolution image pairs is challenging to train these SR methods for new applications. As a result, the high cost of training and implementation leads to poor clinical applicability and limits their applications.

Magnification with arbitrary scales is necessary for super-resolution tasks because the zoom-in/-out process is continuous in practice. Preliminary deep learning super-resolution methods such as SRCNN [130] and DRCN [201] support this characteristic because they upsample the LR image to the targeted size with interpolation methods. However, this pre-upsampling architecture in these networks requires lots of calculations on feature maps with high resolutions and is time-consuming. Thus, state-of-the-art SISR methods such as RDN [132] are mainly developed with a post-upsampling framework. The feature maps are first learned with low-resolution and finally magnified by up-sampling modules (e.g. sub-pixel [189] and deconvolutional layers [182]). This framework leads to computation reduction and results in superior performance with more efficient training. Either the sub-pixel or the deconvolution module is designed with a fixed magnification scale, so the weights of multiple upscale modules must be trained and stored respectively for SR tasks with multi-scales (e.g. $\times 2$, $\times 3$ and $\times 4$) as in MDSR [133]. For magnification with arbitrary scales, the input LR or output HR images must be up-sampled or down-sampled with in-

terpolation methods. However, this pre-/post-processing may lead to image quality decline.

There are two main limitations of scale-free super-resolution. First, existing learning-based up-scale modules (e.g. sub-pixel and deconvolution layer) can not meet the requirements of continuous magnification scales. Second, it is impossible to acquire HR-LR pairs of all potential scales. Thus, it is necessary to introduce meta-learning to this task. Meta-learning [194, 195] is well-known as learning-to-learn, which focuses on leveraging prior experiences to learn a model or strategy that can rapidly adapt to new tasks with limited data or resources. This idea has been widely used in computer vision tasks, including model-agnostic for fast adaptation [297], few-shot and zero-shot learning [298, 299], cross-domain model adaptation [300] and weight prediction [301, 302]. Primarily, the weight prediction strategy aims to train an extra network to predict the weights for the main task. For example, in the scale-free super-resolution task [135], a weight prediction network is applied to generate the convolution kernels for arbitrary magnification scales. Additionally, this meta-upscale module operates location projection and feature mapping based on matrix product and the predicted weights for feature map up-scaling to any shape. 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.

Considering the success of the meta-upscale module on natural image SISR tasks, it is worth seeking to apply meta-learning to tackle scale-free super-resolution in medical images. Specifically, what is the most efficient network architecture for feature extraction cooperating with the meta-upscale module? How can the GANs result in robust visual quality improvement on all scales? In this chapter, I implement an end-to-end medical image SR network, which takes one LR image as input and generates corresponding SR images of an arbitrary magnification scale. Additionally, the model is trained with pixel-wise error, perceptual loss and GAN-based adversarial loss to improve generated images' perceptual and fidelity quality synchronously. Furthermore, I overcome the cost of modifying well-trained models to new medical modalities using transfer learning. In particular, I focus on SR tasks with arbitrary scales in $(1, 4]$ to meet the most common needs in clinical practice. Meanwhile, the method can tackle larger scales with proper training settings. Briefly speaking, the main contributions of this work are:

- I first introduce meta-learning to medical image SR tasks and propose the first scale-free super-resolution method for medical images. Compared with SOTA SISR methods with a specific magnification scale (i.e. EDSR) and with arbitrary scales (i.e. MetaRDN), the proposed MIASSR has much fewer parameters (1% of EDSR and 26% of MetaRDN). It achieves comparable PSNR scores (-0.30 dB on average) and

superior FID scores (-8.43 improvement on average) in the simulation experiments with four public medical image datasets.

- A comprehensive comparison study of SR image generators and SR loss functions is conducted with the scale-free medical image SISR task. These comparison results show that a lite version EDSR network is proper for model size reduction with no performance decline in this scale-free method. Meanwhile, I introduce Wasserstein GAN with gradient penalty for advanced adversarial learning, significantly improving the perceptual quality of the enhanced images. The proposed method achieves the best FID scores on all seven medical image modalities of the four datasets.

- The proposed method successfully applies to various medical image modalities, including single-/multi-modal brain MR scans, cardiac MR scans and chest CT scans of COVID-19 patients. Additionally, with transfer learning, the pre-trained model of brain MR scans is robustly extended to new medical images (i.e. cardiac MR and chest CT images) with only one-fifth training steps.

The rest of this chapter is structured as follows: in Section 4.2, I introduce the proposed MIASSR method with details of the model architecture and the training loss; in Section 4.3, I demonstrate the experimental settings, including data, evaluation metrics and project implementation; in Section 4.4, I illustrate the comparison study of MIASSR with SOTA SISR methods on four medical image datasets and the impacts of each component in the model; and finally in Section 4.5 I conclude the work of this chapter. All related publications and code are publicly realised on `https://github.com/GinZhu/MIASSR`.

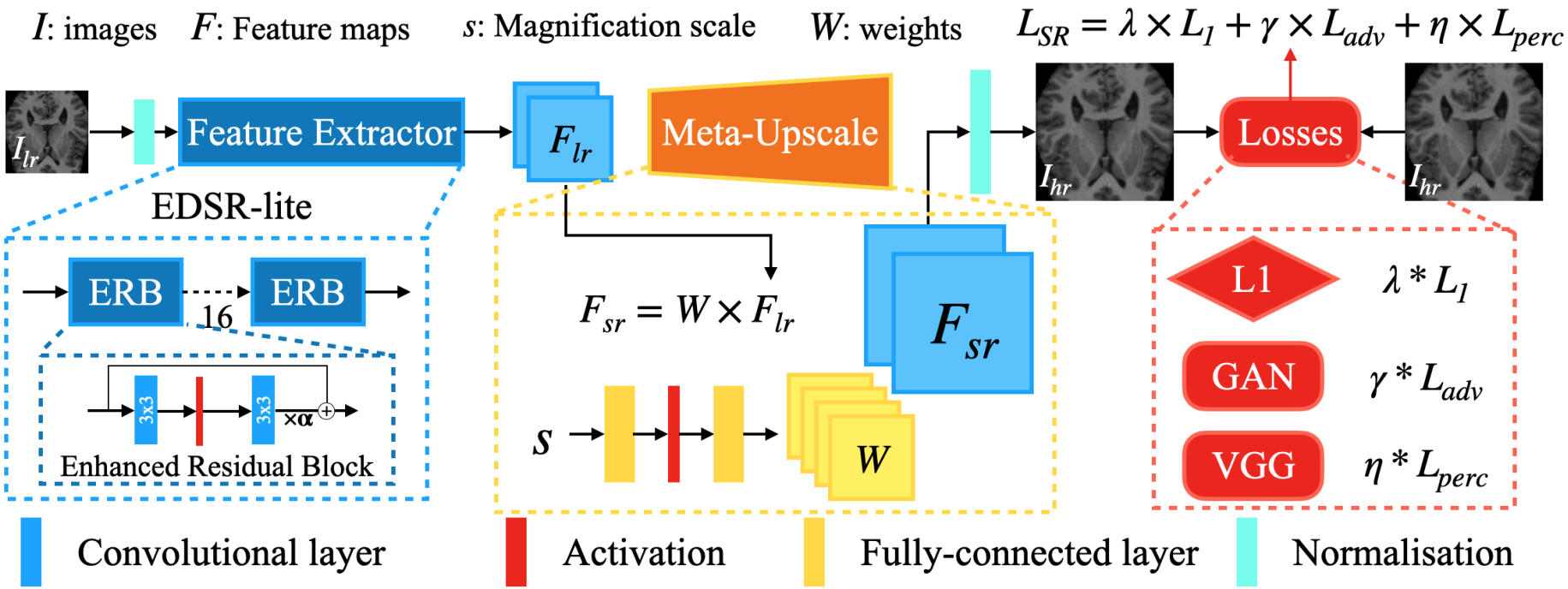


Figure 4.1: The proposed MIASSR consists of an EDSR-lite based low dimension feature extractor and a meta-upscale module. The feature extractor comprises 16 enhanced residual blocks, including two convolution layers and a non-linear activation layer. It extracts low-dimension feature maps $\boldsymbol{F}_{lr}$ of the input LR image $\boldsymbol{I}_{lr}$. 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 by matrix multiplication. Afterwards, super-resolved image $\boldsymbol{I}_{sr}$ is generated from the enlarged feature maps $\boldsymbol{F}_{sr}$. The whole model is trained end-to-end with a combined loss function, including L1, adversarial, and VGG-based perceptual losses.

4.2 Methods

Single image super-resolution aims to restore a high-resolution image $\boldsymbol{I}_{hr}$ from one low-resolution observation $\boldsymbol{I}_{lr}$ of the same object. Generally, the LR image is modelled as [8]:

$$\boldsymbol{I}_{lr} = (\boldsymbol{I}_{hr} \star \boldsymbol{\delta}) \downarrow_s + \boldsymbol{n}, \tag{4.1}$$

where $\boldsymbol{I}_{hr} \star \boldsymbol{\delta}$ denotes the degradation during the image capturing with down-sampling $\downarrow_s$ and noise $\boldsymbol{n}$. SISR aims to inverse this above degradation mapping to recover a super-resolved image $\boldsymbol{I}_{sr}$ from $\boldsymbol{I}_{lr}$[9]:

$$\boldsymbol{I}_{sr} = G(\boldsymbol{I}_{lr}, s; \boldsymbol{\theta}_G), \tag{4.2}$$

where G is a CNN based SR image generator and $\boldsymbol{\theta}_G$ denotes its trainable weights. In each step of training, errors between the approximation $\boldsymbol{I}_{sr}$ and the HR ground truth $\boldsymbol{I}_{hr}$ are measured by a well-designed loss function $\mathcal{L}_{SR}$. The backpropagation passes this loss to the whole network to calculate gradients and update the weights $\boldsymbol{\theta}_G$ [9]:

$$\hat{\boldsymbol{\theta}_G} = arg \min_{\boldsymbol{\theta}_G} \mathcal{L}_{SR}(G(\boldsymbol{I}_{lr}), \boldsymbol{I}_{hr}). \tag{4.3}$$

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

$$\boldsymbol{I}_{sr} = G(\boldsymbol{I}_{lr}) = \mathcal{M}(\mathcal{F}(\boldsymbol{I}_{lr}), s). \tag{4.4}$$

Remember that the input and output layers are ignored to simplify this equation. They normalise and convert the images to the generator's feature domain and vice versa. The rest of this section will introduce each component of the proposed MIASSR (Fig. 4.1) in detail: the feature extraction, the meta-upscale module and the losses for training.

4.2.1 Feature extraction with EDSR-lite

I use an Enhanced Residual Block (ERB) $\mathcal{B}_{ERB}$ based feature extractor, namely EDSR-lite in MIASSR. This residual block (Fig. 4.2) is first proposed in EDSR [133], which consists of two convolution layers $\mathcal{C}$, a non-linear activation layer φ, a residual connection, and no

batch normalisation layer:

$$\begin{aligned}
&\boldsymbol{F}_{out} = \mathcal{B}_{ERB}(\boldsymbol{F}_{in}) \\
&= \boldsymbol{F}_{in} + \mathcal{C}_w^1(\varphi(\mathcal{C}_w^2(\boldsymbol{F}_{in}; \phi_{\mathcal{C}^2})); \phi_{\mathcal{C}^1}) \times \alpha, \\
&\phi_{\mathcal{C}^1}, \phi_{\mathcal{C}^2} \in \boldsymbol{\theta}_G,
\end{aligned} \tag{4.5}$$

where $\boldsymbol{F}_{in}$ and $\boldsymbol{F}_{out}$ 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. Because padding is applied correspondingly with the convolution kernel size in all convolution layers, feature maps can remain the same size. Thus, it becomes very convenient to build deeper neural networks by just stacking up several residual blocks, and the low-dimension feature maps $\boldsymbol{F}_{lr}$ can be extracted from the input LR image $\boldsymbol{I}_{lr}$:

$$\boldsymbol{F}_{lr} = \mathcal{F}(\boldsymbol{I}_{lr}) = \mathcal{B}^b(\boldsymbol{I}_{lr}; \phi_{\mathcal{B}}), \phi_{\mathcal{B}} \in \boldsymbol{\theta}_G. \tag{4.6}$$

In this work, I use a lite version of EDSR, which consists of $b = 16$ enhanced residual blocks (Equation 4.5) and 64 convolution kernels with a size of 3×3 for each convolutional layer. Unlike the handcrafted feature extraction (e.g. texture, shape and morphological features) [303], the feature extraction networks automatically learn relevant hierarchical features ranging from basic edges and textures at initial layers to more complex patterns in deeper layers. Thus, it has better general applicability with image types and pre-processing operations. Section 4.4.2 will illustrate the comparison of widespread CNN-based blocks for LR feature extraction and introduce how the hyper-parameters (e.g. block type, number of blocks and width) are chosen.

4.2.2 Meta-upscale module

The low-dimension feature maps $\boldsymbol{F}_{lr}$ extracted by $\mathcal{F}$ in Equation 4.6 need to be up-sampled to generate HR output from LR input:

$$\boldsymbol{F}_{sr} = \mathcal{M}_s(\boldsymbol{F}_{lr}; \phi_{\mathcal{M}_s}), \tag{4.7}$$

where $\boldsymbol{F}_{sr}$ is the super-resolved feature maps, $\mathcal{M}_s$ is an up-sampler, and $\phi_{\mathcal{M}_s}$ is the group of parameters for the magnification option with scale s. Common single scale up-samplers, such as sub-pixel [189], only learn one group of parameters for a specific SR scale s. In contrast, the meta-upscale module [135] learns to predict a group of weights for each SR scale. Particularly, one pixel with indexes (i, j) in the super-resolved feature maps $\boldsymbol{F}_{sr}$ is

calculated as a weighted sum of all pixels in $\boldsymbol{F}_{lr}$:

$$\boldsymbol{F}_{sr}(i,j) = \mathbf{v}_{i,j} \times \boldsymbol{F}_{lr}, \tag{4.8}$$

where $\boldsymbol{F}_{lr}$ with original shape $[H_{in}, W_{in}]$ is flattened to $[H_{in} \times W_{in}, 1]$ and $\mathbf{v}_{i,j}$ is a vector with a size of $[1, H_{in} \times W_{in}]$. 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):

$$\mathbf{v}_{i,j} = \mathcal{W}(\frac{i}{s} - \left\lfloor \frac{i}{s} \right\rfloor, \frac{j}{s} - \left\lfloor \frac{j}{s} \right\rfloor, \frac{1}{s}), (i,j) \in \boldsymbol{F}_{sr}. \tag{4.9}$$

Accordingly, all pixels in $\boldsymbol{F}_{sr}$ can be achieved via matrix multiplication:

$$\boldsymbol{F}_{sr} = \boldsymbol{W}_s \times \boldsymbol{F}_{lr}, \tag{4.10}$$

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

$$\boldsymbol{F}_{sr} = \mathcal{M}(\boldsymbol{F}_{lr}, s) = \mathcal{W}(s; \phi_{\mathcal{W}}) \times \boldsymbol{F}_{lr}, \phi_{\mathcal{W}} \in \boldsymbol{\theta}_G. \tag{4.11}$$

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 scale, which differs from sub-pixel up-samplers. Thus, it becomes possible to train an end-to-end model for SR tasks with arbitrary magnification scales.

4.2.3 Loss functions

A combined super-resolution loss is used in this work to train MIASSR (Equation 4.2). It consists of the pixel-wise L1 loss $\mathcal{L}_1$, adversarial loss $\mathcal{L}_{adv}$ and perceptual loss $\mathcal{L}_{perc}$:

$$\mathcal{L}_{SR} = \lambda \times \mathcal{L}_1 + \gamma \times \mathcal{L}_{adv} + \eta \times \mathcal{L}_{perc}, \tag{4.12}$$

where λ, γ, and η are scale factors to balance each part of the loss function $\mathcal{L}_{SR}$. This additive form of loss has proven effective for comprehensive objective capture, gradual refinement, training stability and flexibility. However, it is essential to strike a balance between each component. Inappropriate hyper-parameters (e.g. type and scale of loss) may lead to issues such as training instability, local minima/over-fitting and complex evaluation. In section 4.4.3, the ablation study will introduce the impacts of each loss component and how the scales are determined.

L1 loss SISR requires predicting the correct value of each pixel in the super-resolved images. Thus, pixel-wise errors are important for evaluating and training SR networks. In this work, I 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:

$$\mathcal{L}_1(\boldsymbol{I}_{sr}, \boldsymbol{I}_{hr}) = \frac{1}{H * W} \sum_{(i,j)\in I} \|\boldsymbol{I}_{hr}[i,j] - \boldsymbol{I}_{sr}[i,j]\|\,, \tag{4.13}$$

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, it is also limited to generating perceptually realistic textures in medical images because it leads to over-smoothing. In the visual system of human beings, images are not processed as a set of pixels but as patches or as a whole, so generating realistic features is also essential. Therefore, perceptual and adversarial losses are also involved.

VGG based perceptual loss The VGG based perceptual loss $\mathcal{L}_{perc}$ was first introduced in [139] and had been widely used in super-resolution tasks [23, 24, 114, 134]. It presents the mean-square-error (MSE) between the super-resolved images and the HR ground truth images in the feature domain:

$$\mathcal{L}_{perc}(\boldsymbol{I}_{sr}, \boldsymbol{I}_{hr}) = \mathbb{E}(\|\mathcal{V}_l(\boldsymbol{I}_{hr}) - \mathcal{V}_l(\boldsymbol{I}_{sr})\|^2), \tag{4.14}$$

where $\mathcal{V}$ is a pre-trained VGG-19 model and l denotes the feature maps of the specific layer of $\mathcal{V}$. Following the conclusion in ESRGAN [134], I use the feature maps before the non-linear activations of earlier layers to provide more textural information.

Adversarial loss To generate perceptually more realistic images, adversarial learning with Wasserstein GAN is applied in the method. A GAN comprises the SR image generator G and a discriminator D. Both networks are trained jointly. The discriminator aims to correctly recognise any image as real or fake, while the generator aims to produce as real as possible SR images. Thus, the primary adversarial loss function is defined as [101]:

$$\mathcal{L}_{GAN} = -\mathbb{E}_{\boldsymbol{I}_{hr}}\left[\log D(\boldsymbol{I}_{hr})\right] - \mathbb{E}_{\boldsymbol{I}_{lr}}\left[\log(1 - D(G(\boldsymbol{I}_{lr})))\right]. \tag{4.15}$$

Whilst this basic version of adversarial loss has been successfully used in natural image [114] and medical image [23] super-resolution tasks, it is susceptible to problems of training instability and mode collapse. When the discriminator becomes too powerful (especially at the beginning of the training), vanishing gradients may occur with the generator.

Meanwhile, the logarithmic terms can cause sharp regions, saturations and poor gradients in the loss landspace, leading to training instability. Additionally, this loss (Equation 4.15) has no direct correlation with perceptual quality, which means no clear signal of the training procedure. Thus, Wasserstein GAN is proposed to resolve these issues [108]. It measures the Earth Mover's distance (also known as Wasserstein distance) between the two probability distributions of authentic and generated data, representing the cost of transforming one distribution into another. Compared to the Jensen-Shannon divergence used in vanilla GANs, the Wasserstein distance provides meaningful and smooth gradients regardless of the discriminator and generator states. The generator clearly knows how to improve and tends to produce more diverse samples, thereby mitigating mode collapse. In practice, the WGAN also removes the logarithmic terms to provide a smoother loss landscape for more stable training:

$$\mathcal{L}_{WGAN} = \mathbb{E}_{\boldsymbol{I}_{lr}} \left[D(G(\boldsymbol{I}_{lr})) \right] - \mathbb{E}_{\boldsymbol{I}_{hr}} \left[D(\boldsymbol{I}_{hr}) \right]. \tag{4.16}$$

An essential trick is to clip all the discriminator's weights to a constant range $[-c, c]$ for derivable Wasserstein distance in the training of WGAN. However, the clipping strategy makes the weights the minimum or maximum values. As a result, the discriminator behaves like a binary network and declines the non-linear simulating ability of the GAN. Thus, the gradient penalty is proposed to replace the clipping operation [22]. The new trick restricts the gradients of D to not change rapidly by adding a new term in the adversarial loss:

$$\mathcal{L}_{adv} = \mathcal{L}_{WGAN} + \mathbb{E}_{\boldsymbol{I}} \left[\left\| \bigtriangledown_{\boldsymbol{I}} D(\boldsymbol{I}) \right\|_p - 1 \right]^2, \tag{4.17}$$

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

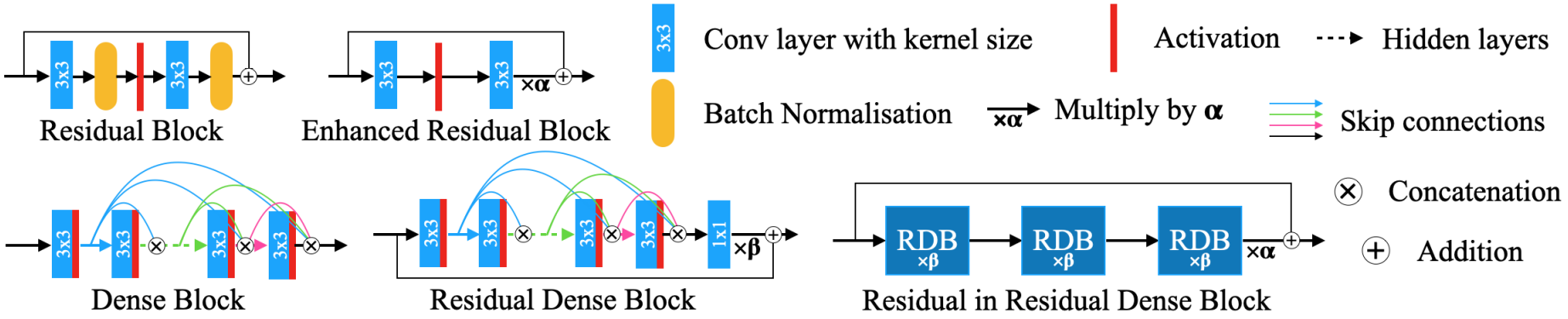


Figure 4.2: Residual and dense blocks. These blocks have convolution and activation layers in common but have various skip connections. They are the basic unit of SOTA SISR networks: Residual Block is used in SRResNet and SRGAN [114]; Enhanced Residual Block is used in EDSR and MDSR [133]; Dense Block is used in SRDenseNet [220]; Residual Dense Block is proposed in RDN [132]; Residual in Residual Dense Block is proposed in ESRGAN [134]. Notice that residual scales α and β are introduced to stable the training of very deep neural networks (e.g., ESRGAN).

4.3 Experiments

In the simulation experiments, the proposed method successfully applies to four different medical image datasets (Section 4.3.1) in super-resolution tasks with arbitrary scales of magnification in $(1, 4]$. HR ground truth (GT) and corresponding LR images are generated from the original slices. To evaluate the method, metrics including PSNR, SSIM and Fréchet Inception Distance (FID) [175] are used to measure the differences between the super-resolved images and GT images in the test set (Section 4.3.2). Afterwards, the mean performance over all scales in $(1, 4]$ is compared with bicubic interpolation and 7 SOTA SISR methods (Section 4.3.4).

4.3.1 Data and pre-processing

LR-HR image pairs for training, validation and testing are generated from the original slices. HR images are achieved by removing the pure-black background margin of the original slices. LR images are generated by down-sampling the corresponding HR images and blurred with a 3×3 Gaussian kernel. I focus on the central regions of each slice because the pure-black background regions only provide useless information and slow down the training process. In the experiments, a suitable margin size for each dataset is carefully chosen to ensure that no non-zero values are removed.

OASIS The open access series of imaging studies (OASIS) [304] dataset consists of a cross-sectional collection of 416 subjects, including individuals with early-stage Alzheimer's Disease (AD). Each subject includes 3 or 4 individual T1-weighted MRI scans obtained within a single imaging session. The brain-masked version of an atlas-registered gain filed-corrected image, namely T88-111, is used for the single modality SR experiments. Due to the limitations of computing resources, from the whole dataset, I randomly select 30 subjects for training, 3 subjects for validation and another 9 subjects for testing. Notice that the validation dataset is only for hyperparameter searching. The original size of each subject is $[176 \times 208 \times 176]$, and only a central area of $[144 \times 180]$ is used. All these experiments are applied on the axial plane.

BraTS The brain tumour segmentation dataset (BraTS) [280–282] 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. I randomly select 50 scans (35 HGG and 15 LGG) for training and 10 scans (7 HGG and 3 LGG) for testing. The original image shape is $[240 \times 240 \times 155]$. Slices on the axial plane are cropped to $[180 \times 170]$ to

focus the training on the brain area.

ACDC The automated cardiac diagnosis challenge (ACDC) [305] was established to encourage the development of algorithms for the automatic segmentation, classification, and analysis of cardiac pathologies using MR images. The open-access dataset includes 1.5T and 3.0T MR scans of 100 subjects with expert annotations of various cardiac structures such as left ventricle, right ventricle and myocardium. It uniformly contains five pathology categories: normal subjects, dilated cardiomyopathy, hypertrophic cardiomyopathy, myocardial infarction and abnormal right ventricles. In this work, I randomly assign the samples to 80 cases for training and 19 cases for testing. All experiments are conducted on the transverse plane, where the slices have various shapes from $[174 \times 208]$ to $[184 \times 288]$. To standardise the slice shapes, only the central areas with a size of $[128 \times 128]$ are cropped. This patch size can ensure all ROIs (i.e. ventricles and myocardium) remain.

COVID-CT The COVID-CT dataset [306] is published for the medical and research community amid the ongoing challenges posed by the SARS-CoV-2 pandemic. It encompasses lung CT scans from 632 patients diagnosed with COVID-19 infections. All patients had a positive Reverse Transcription Polymerase Chain Reaction (RT-PCR) confirmation for the presence of SARS-CoV-2 from a sample obtained within one day of the initial CT. CT scans were conducted without intravenous contrast and utilised a soft tissue reconstruction technique. The images originally in DICOM format were later transformed into the NIfTI format. In this work, I randomly selected the images of 199 patients for training and images of another 25 patients for testing. The original image shape is $[512 \times 512]$. After removing the background, only the $[412 \times 332]$ centre area is used for training or testing. The data was manually inspected to ensure the cropping operation preserved the lungs.

4.3.2 Metrics

Three objective image quality assessment methods measure the fidelity and perceptual quality of the generated SR images compared with ground truth (GT) images in the experiments. First, I use the peak signal-to-noise ratio (PSNR) and structural similarity (SSIM [158]) for local and structural reconstruction accuracy measurement. However, they cannot evaluate perceptual quality. Over-smoothed images have been reported in [24, 114] to achieve higher PSNR and SSIM scores than texture-rich images, but they might be less perceptually realistic. Thus, I also calculate the Fréchet Inception Distance (FID) [175], which is popular to evaluate the perceptual quality of generated images by GANs. It measures the difference between high-level hidden 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 [173]. Notice

that higher scores of PSNR and SSIM represent better fidelity quality, while lower FID indicates more perceptually realistic images.

4.3.3 Implementation details

I use PyTorch [37] to implement this method, NiBabel [307] to load medical data, and OpenCV-python [294] for image resize and blur operations. All experiments are conducted on an Nvidia Quadro RTX 8000 GPU. The details of training tricks and hyperparameters are released as config files on GitHub.

The generator consists of $b = 16$ enhanced residual blocks, in which each convolution layer has $w = 64$ feature maps. A residual scale $\alpha = 1.0$ is used for the residual connections in ERB. The discriminator is similar to in DCGAN [110] and SRGAN [114]. It consists of 7 down-sample blocks, each with one convolution layer with $stride = 1$ for feature expanding and one convolution layer with $stride = 2$ for feature maps down-sampling. No batch normalisation layers are used in either the generator or the discriminator, while leaky-ReLU [308] with $negative\text{-}slope = 0.2$ is chosen as the non-linear activation function. In the experiments without transfer learning, both networks are initialised with Kaiming-uniform [70].

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

$$\begin{aligned} H_{lr}, W_{lr} &= \left\lfloor \frac{H_p}{s} \right\rfloor, \left\lfloor \frac{W_p}{s} \right\rfloor; \\ H_{hr}, W_{hr} &= \lfloor sH_{lr} \rfloor, \lfloor sW_{lr} \rfloor. \end{aligned} \tag{4.18}$$

For each training step, a batch of 16 random patch-pairs with the same SR scale is fed to the model. Firstly in pre-training, the generator of MIASSR is trained with only $\mathcal{L}_1$ for 1×10^5 steps because the warm-up can make the training of GANs more stable [134]. Then I train both generator and discriminator with $\lambda = 1$, $\gamma = 0.001$ and $\eta = 0.006$ for 1×10^5 steps. Adam optimiser [61], with an initial learning rate $\iota = 0.0001$, momentum $= 0.9$ and betas $= (0.9, 0.999)$ is used for backpropagation. The learning rate is halved every 5×10^4 steps. Losses above 1×10^8 are discarded to avoid the gradient explosion.

All the above hyperparameters are chosen based on the validation performance in the experiments on the OASIS dataset. In the transfer learning experiments of the ACDC and COVID-CT datasets, the model pre-trained on the OASIS dataset is fine-tuned for 1×10^4 steps with $\mathcal{L}_{SR}$. Meanwhile, to make MIASSR work with multi-modal scans,

the single-channel input and output layers are modified to 4-channel in the experiments with the BraTS dataset. The four modalities of BraTS (i.e. T1ce, T1, T2 and Flair) are stacked during training and testing. Meanwhile, the loss function $\mathcal{L}_{SR}$ is calculated on each modality respectively and then averaged.

4.3.4 Comparison with SOTA methods

I compare the proposed method with the bicubic interpolation and 7 SOTA SISR methods: SRGAN [114], EDSR [133], SRDenseNet [220], RDN [132], MDSR [133], ESRGAN-L1 [134] and MetaRDN [135]. All these methods are designed for natural images, which have much bigger dimensions than the medical images used in this work. I have re-trained the models with the medical image datasets with adjusted smaller patches to make them work well with the experiments. For a fair comparison, all models are trained with the same steps and learning rate decay policy without model embedding and data augmentation. I use the original loss functions to train most of the models. However, the perceptual loss based on material recognition in ESRGAN [145] hinders the training with medical images, so I use ESRGAN trained with only the MAE loss $\mathcal{L}_1$ (so-called ESRGAN-L1) instead.

All these SISR methods except MetaRDN are designed only for specific integer magnification scales (e.g. $\times 2$, $\times 3$ and $\times 4$). I have used an up-and-down strategy to evaluate their performance on SR tasks with float scales. It contains two steps: in the first up-sampling step, the well-trained model with the ceiling scale $\lceil s \rceil$ is used to generate an over-magnified SR image, then in the down-sampling step, the over-magnified image is resized to the target resolution with the bicubic interpolation.

4.4 Results and discussion

In this section, I first present the performance of MIASSR with four datasets compared with SOTA SISR methods with specific or arbitrary scales. Then, ablations studies will illustrate the impacts of the LR feature extraction networks and the influence of each loss component, explaining how the hyper-parameters of network architecture and the scales of loss components are determined.

4.4.1 MIASSR performance

On the OASIS dataset and model efficiency First of all, in the experiments with the OASIS dataset (Fig. 4.3), MIASSR is compared with SOTA methods with arbitrary SR scales. On average, it achieves the third-best mean PSNR and SSIM scores and the best FID with the fewest parameters (Table 4.1). With all SR scales, MIASSR generates images with comparable fidelity and perceptual quality with SOTA methods (Fig. 4.5). Considering the balance between cost and performance, the mean PSNR score and FID are plotted with the number of parameters of each model in Fig. 4.4. Particularly for SISR methods which only support one SR scale (i.e. EDSR, RDN, SRDenseNet and ESRGAN), the model size denotes the number of all parameters of three models (for $\times 2$, $\times 3$ and $\times 4$ SR respectively) because they are all required in inference. Methods designed for multi-scale (i.e. MDSR) and arbitrary scales (i.e. MetaRDN and MIASSR) intensely narrow the model size by reducing the required models to one.

Notably, MIASSR has the fewest parameters, only 26% of the existing scale-free method MetaRDN and fewer than 1% of EDSR. Although they are learning more challenging transformations from the low-resolution space to the high-resolution space, multi-task methods (i.e. MDSR, MetaRDN and MIASSR) achieve much better PSNR scores than single-scale SR methods. Instead of approximating one mapping with a specific SR scale, they share parameters in approximations of mappings with various SR scales. The parameter sharing between tasks and the diversity of LR-HR image pairs with different magnification scales make the models converge better in the training process. Suppose to divide all the methods into two groups: methods with meta-learning and methods without meta-learning. It is clear to observe the perception-distortion trade-off [309]. The stem of workflow, including SR task design and upscale implementation, decides the primary performance. At the same time, the details of model architectures, loss functions and training tricks only affect the balance between fidelity and perceptual quality.

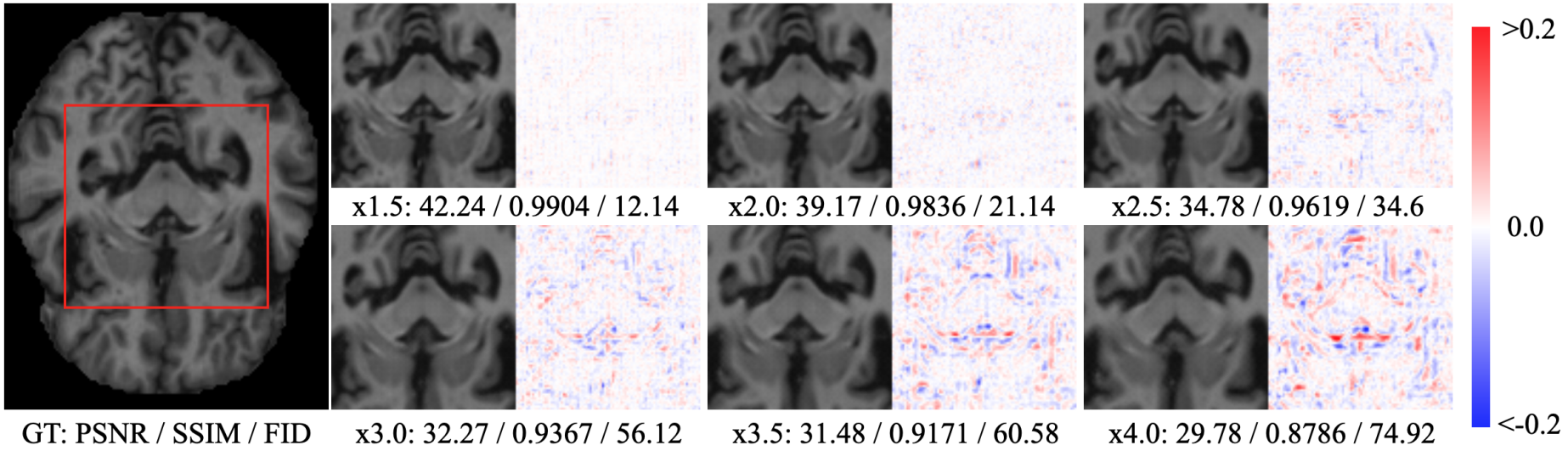


Figure 4.3: An example of super-resolved images with different scales of magnification by the proposed MIASSR. The slice is randomly selected from the OASIS testing dataset. Here I only illustrate the SR images with six scales $(1.5, 2.0, 2.5, 3.0, 3.5, 4.0)$. The method could generate SR images with arbitrary scales in $(1, 4]$. All images are converted to [0, 1]. Differences between SR images and ground truth images are rendered with the colour bar. PSNR, SSIM and FID are used for image quality evaluation. Higher PSNR and SSIM indicate better fidelity quality, while lower FID represents better perceptual quality.

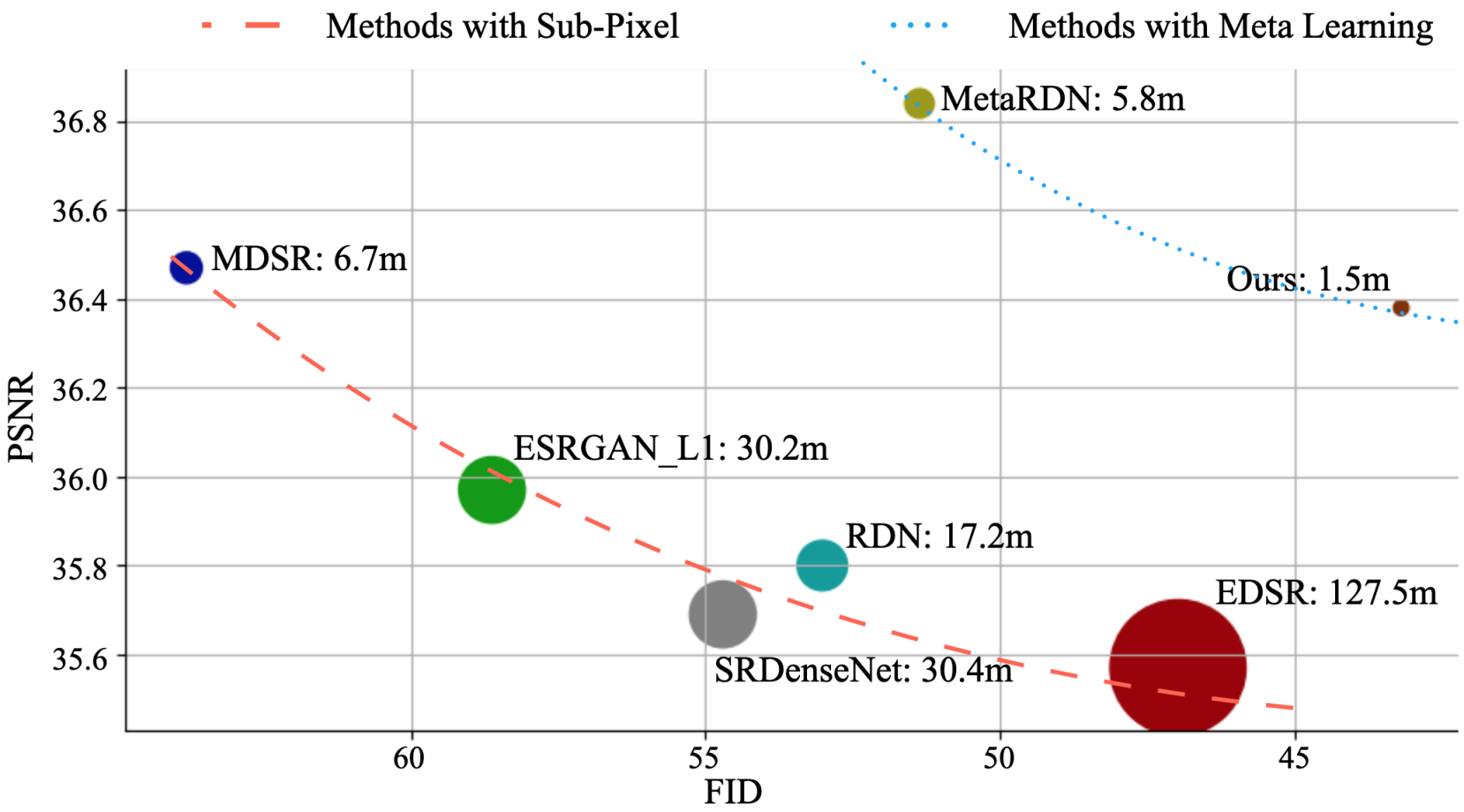


Figure 4.4: Performance and model size of MIASSR compared to SOTA SISR methods. Mean PSNR and mean FID on SR scales between $(1, 4]$ are calculated to evaluate generated images' fidelity and perceptual quality. The bubble size denotes the number of parameters of each model. The plot does not show SRGAN because of its much worse performance ($meanPSNR = 28.15, meanFID = 128.21$) than the other methods. Higher PSNR scores denote better pixel-wise fidelity performance, and lower FID represents better perceptual quality. The proposed method MIASSR has achieved the best FID and the third-best PSNR with the smallest model size. The perception-distortion trade-off [309] is reflected in each group: (a) methods with meta-upscale modules, which directly achieve scale-free magnification results; and (b) methods with sub-pixel modules, which achieve magnification results with float scales by the post-processing of down-sampling. High fidelity and perceptual quality are at odds and impossible to be improved simultaneously in each group.

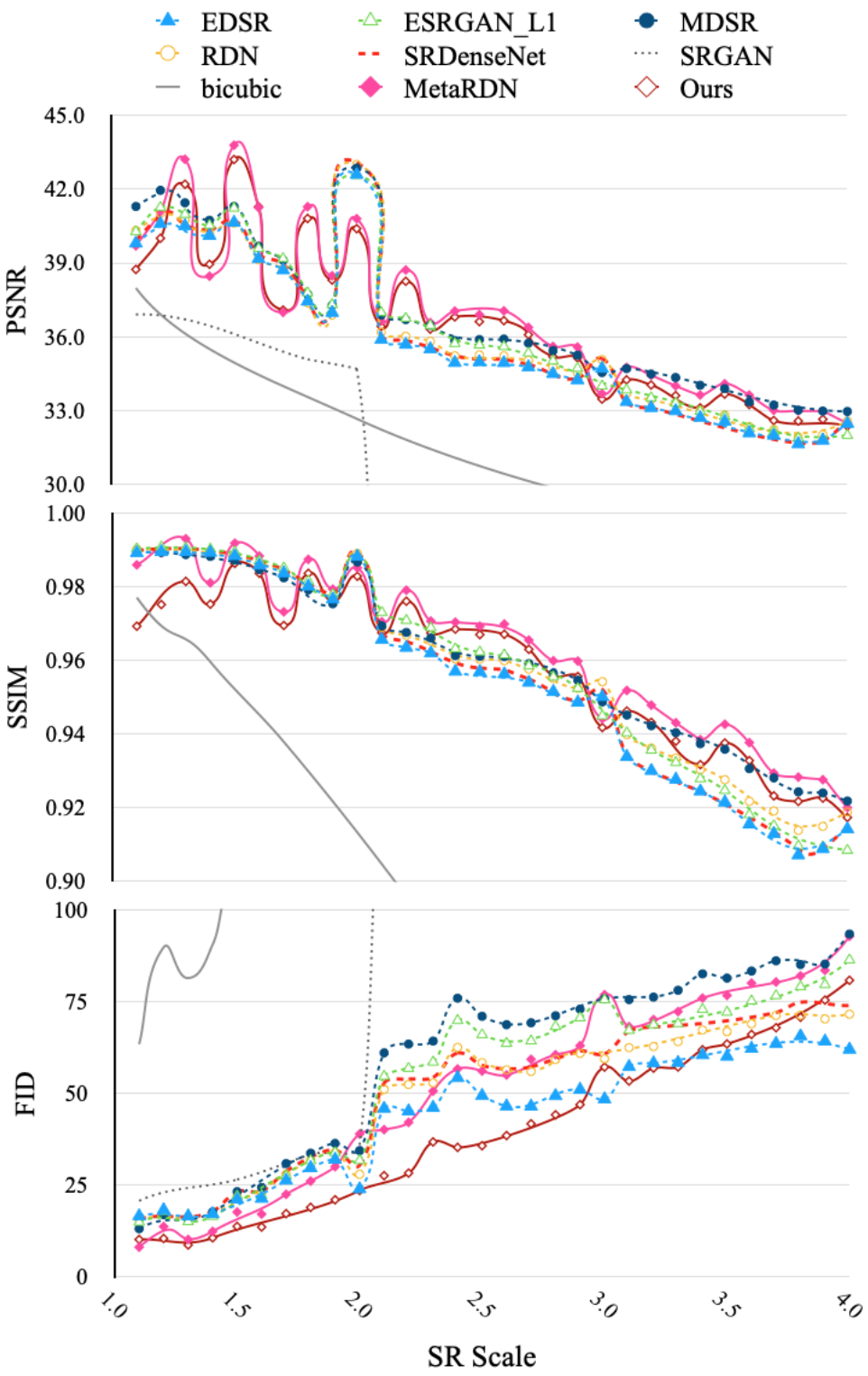


Figure 4.5: Comparing the proposed method with bicubic interpolation and SOTA methods in SR tasks with arbitrary scales in $(1, 4]$. Higher PSNR and SSIM denote better fidelity quality, while lower FID represents better perceptual quality. Results of bicubic interpolation and SRGAN are not fully plotted because of their poor performance. The proposed method has achieved comparable performance with SOTA methods with all SR scales.

Table 4.1: Comparing MIASSR with SOTA methods in SR tasks with arbitrary scales in $(1, 4]$ on the OASIS 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 a million. The proposed method achieves the best FID and comparable PSNR and SSIM scores along with the smallest model size out of all CNN-based methods.

	BiCubic	SRGAN [114]	EDSR [133]	ESRGAN($\mathcal{L}_1$) [134]	MDSR [133]	RDN [132]	SRDenseNet [220]	MetaRDN [135]	MIASSR
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**

Table 4.2: 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 of the 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. Further, it 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 [133]	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 [135]	**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**

On the BraTS, ACDC and COVID-CT datasets The proposed method also performs well on various medical image modalities, including brain MR, cardiac MR and chest CT scans (Table 4.2). Transfer learning, which decreases the training cost on medical data processing [310], also helps to modify the pre-trained model of the OASIS dataset to new single-modality medical image datasets efficiently and effectively. Compared with the bicubic interpolation method, the proposed MIASSR significantly improves the performance in experiments with the ACDC and COVID-CT datasets (Fig. 4.6). Compared with EDSR and MetaRDN, MIASSR generates images with comparable fidelity quality and better perceptual reality, with only one-fifth of the training steps. Indeed, transfer learning reduces the required training steps from 1×10^5 to 2×10^4. Besides, MIASSR can be extended to multi-modality images conveniently. By simply modifying the input and output layers, it successfully works for the cross-modality SR task of the BraTS dataset (Fig. 4.7). Compared with SOTA methods which work for a single modality, it achieves comparable performance on all four modalities (T1, T2, T1ce and Flair).

In summary, the proposed method has good clinical applicability. In the experiments, it successfully applies to various medical image super-resolution tasks with different situations of modalities and diseases. It works with brain and cardiac MR images (i.e. the OASIS and ACDC datasets), chest CT images (i.e. the COVID-CT dataset), and cross-modality MR scans (i.e. the BraTS dataset). Compared with SOTA methods, MIASSR can generate SR images with comparable reconstruction fidelity and better perceptual quality with a smaller model size. With transfer learning, MIASSR can effectively and efficiently extend to new datasets.

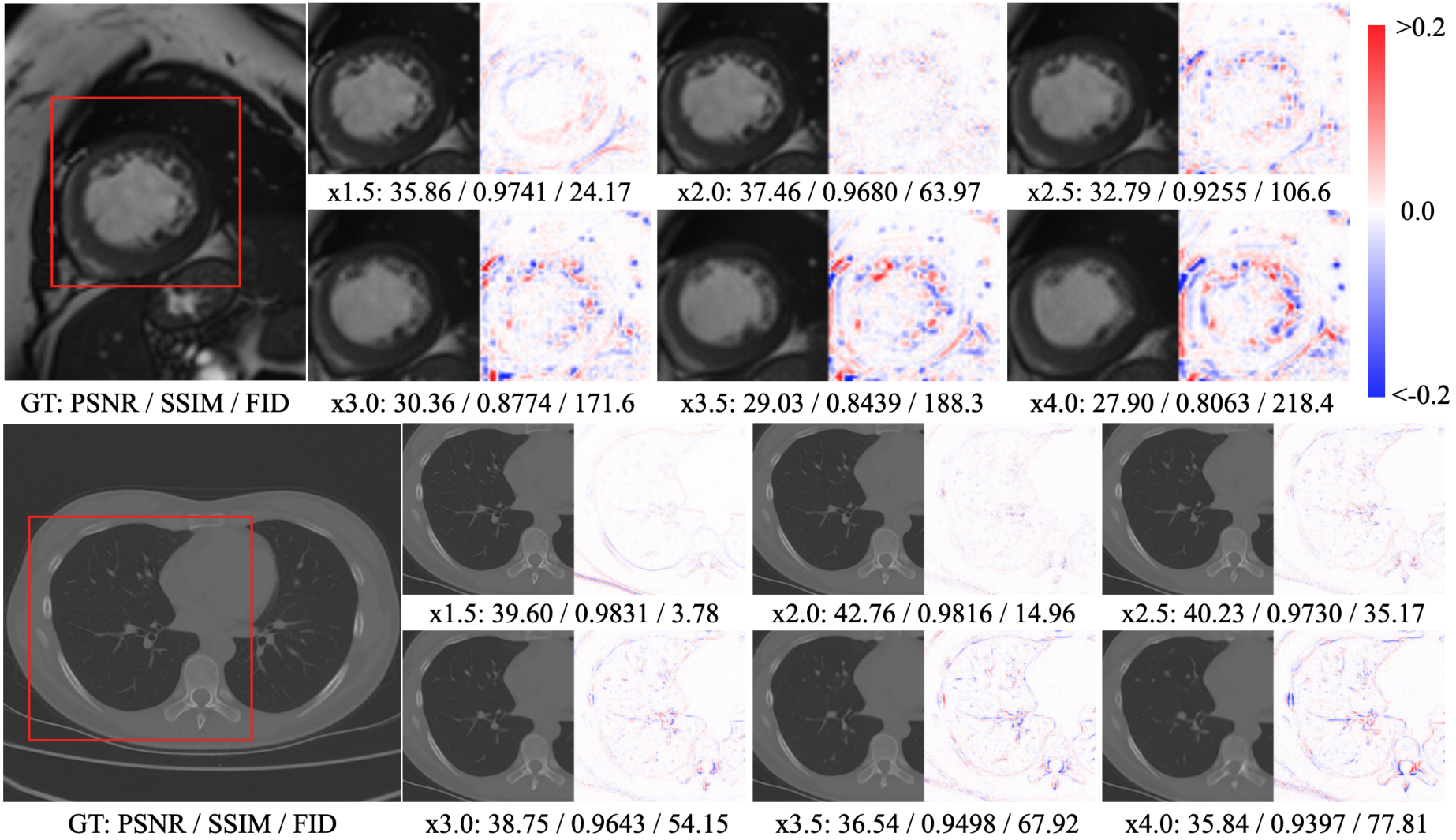


Figure 4.6: With transfer learning, MIASSR can be applied to new medical image modalities conveniently and effectively. I have extended it to cardiac MR images (top) and CT images of COVID-CT patients (bottom). Ground truth image is plotted on the left, while SR images with various magnification scales and pixel-wise errors are on the right.

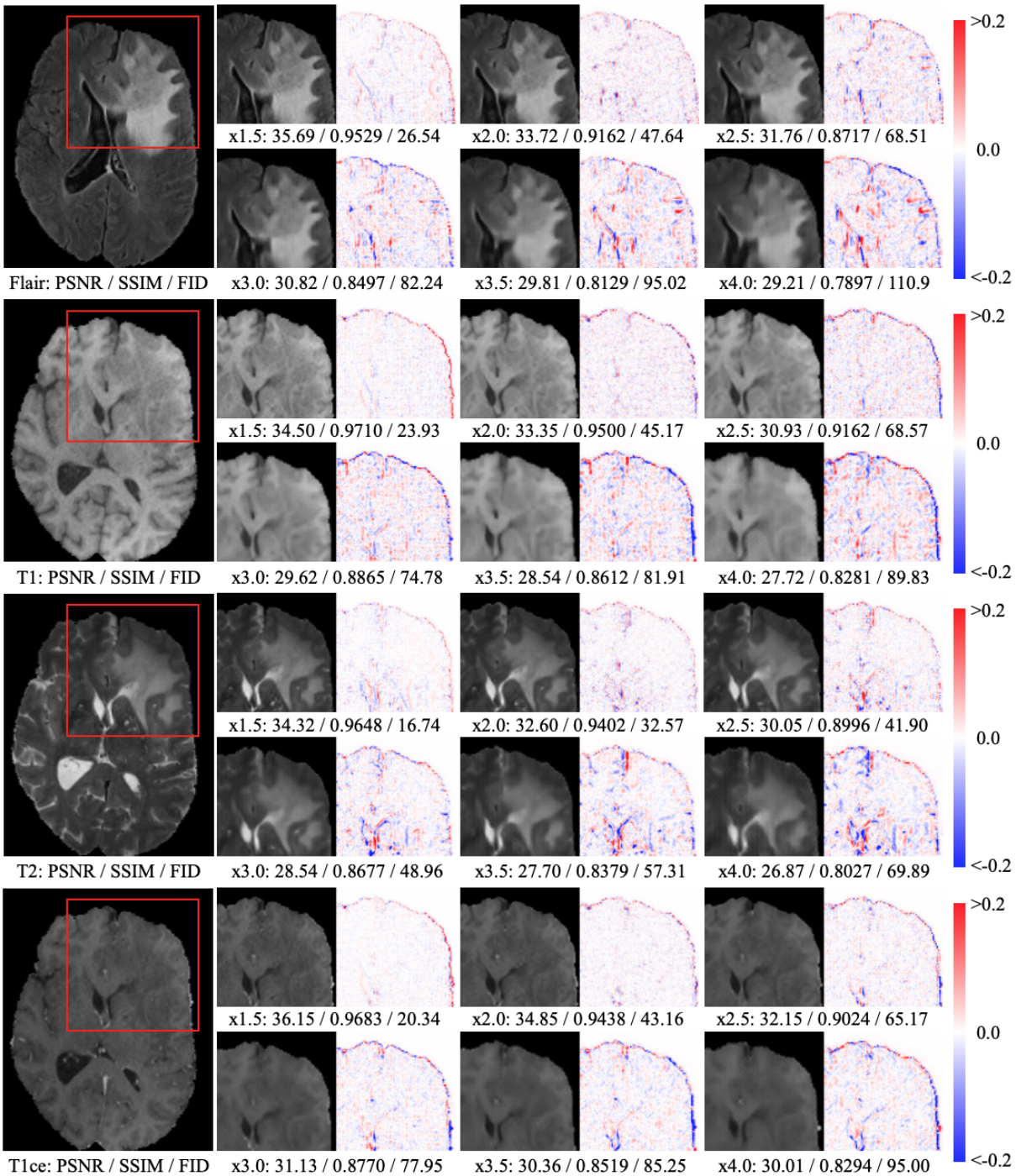


Figure 4.7: MIASSR has successfully worked with the multi-modal brain MR image dataset BraTS. All images are converted to [0, 1]. Differences between SR images and ground truth images are rendered.

4.4.2 Ablation study on the SR image generator

Referring to Equation 4.5 and 4.6, the SR image generator is influenced by three factors: the block structure $\mathcal{B}$, the number of blocks b (i.e. the depth of the network) and the number of convolution kernels of each layer (namely the width w of the network). Briefly speaking, 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. I have designed an ablation study of generator architectures to understand how each factor influences the final performance. Besides, the conclusion also supports the choice of the final hyper-parameters (i.e. block type, number of blocks and width of each layer) in the proposed method with a balanced consideration of SR performance and model efficiency.

Impacts of the network architecture First, I compare a wide range of architectures for LR feature extraction in MIASSR. Six SOTA networks, widely used in computer vision and achieved high performance on SISR tasks, are tested. These networks contain various residual and dense blocks (Fig. 4.2) and behave differently in simulating the LR-to-HR transformation. To match MIASSR, I replace their scale-specific up-samplers with the meta-upscale module, as in MetaSR [135], and keep all the other original settings (e.g. depth and width). To compare the performance of these generators, I train them with only the L1 loss $\mathcal{L}_1$ for 1×10^5 steps with the same training settings. The generated SR images' reconstruction fidelity and visual qualities are evaluated synchronously (Fig. 4.8).

When training with only the pixel-wise loss $\mathcal{L}_1$, the performance divides these variations into two groups. Methods with more layers always perform better because the deeper structures provide more capability for simulating non-linear transformations. Other structures, such as skip connections in dense blocks and the width of models, lead to no significant differences. However, suitable structures for minimising pixel-wise errors might not fit the needs of generating perceptually realistic textures. Thus, I further compare RDN, EDSR and EDSR-lite (i.e. the one used in the proposed MIASSR) with adversarial learning (Fig.4.9). Although MetaRDN performs best with L1 loss, further adversarial learning brings no additional improvements. In contrast, the other two models gain from adversarial and perceptual loss. However, comparing the original EDSR with the EDSR-lite, extra feature maps of each convolution layer only slightly improve the performance, although nearly a hundred times more parameters are used.

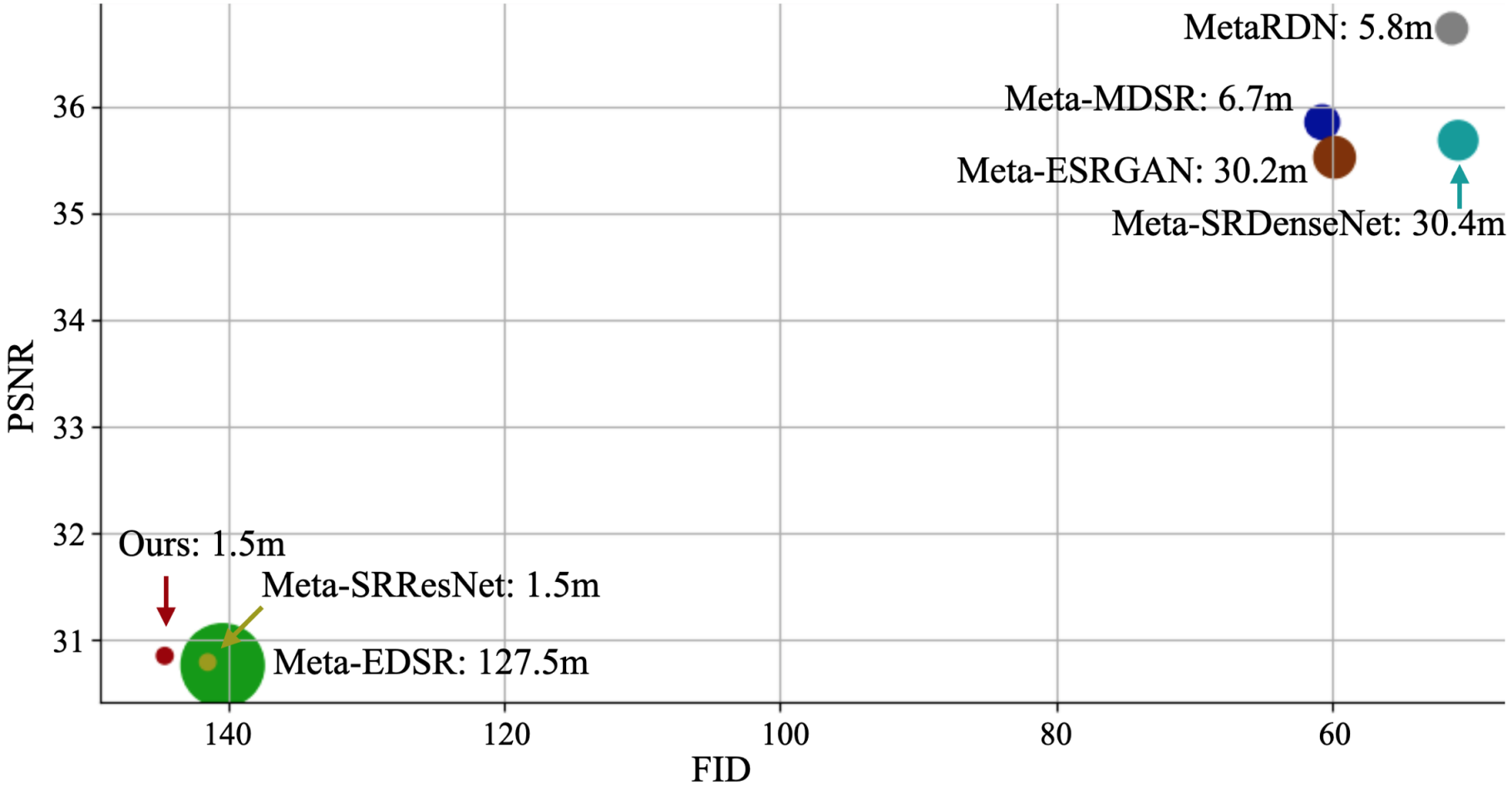
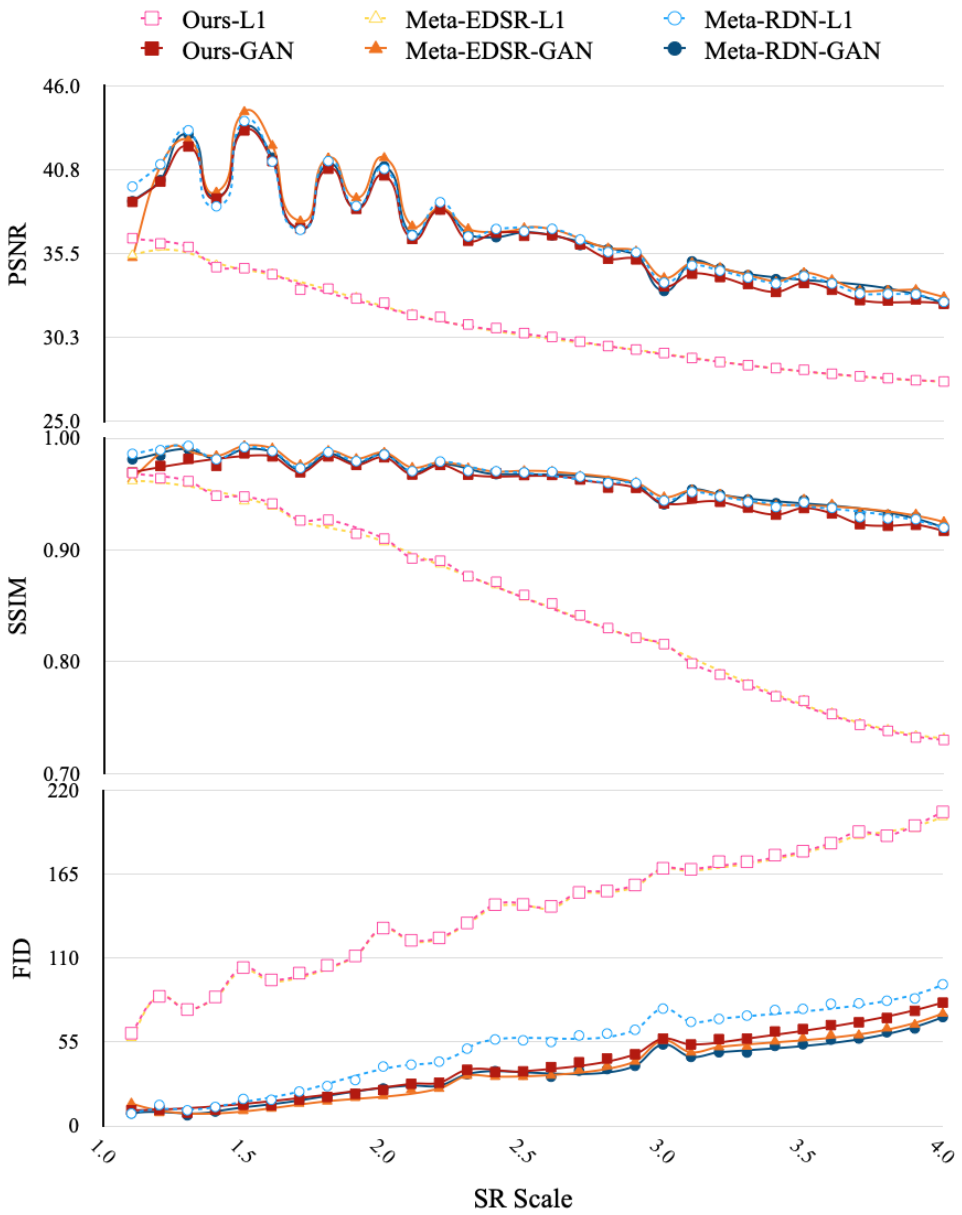


Figure 4.8: Comparison of SR image generators in MIASSR (training with $\mathcal{L}_1$). Higher PSNR indicates better fidelity quality, while lower FID represents more perceptually realistic SR images. Bubble size denotes the number of parameters. The basic block design, such as skip connections, rarely affects the final performance, but the depth of networks impacts the performance a lot. Deeper networks (MetaRDN, Meta-MDSR, Meta-ESRGAN, and Meta-SRDenseNet) perform better than others.



	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

Figure 4.9: The sensitivity analysis of SR image generators to GAN-based adversarial learning. Three networks, EDSR-lite, EDSR and RDN are trained with $\mathcal{L}_1$ only and compared with the extra-trained models with additional $\mathcal{L}_{adv}$ and $\mathcal{L}_{perc}$. 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.

Impacts of network width and depth Additionally, I research 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 lead to more challenging optimisation of Equation 2.2.1 and over-fitting. Meanwhile, the number of parameters grows linearly with the depth and quadratically with the width, so the cost and performance balance should also be considered. In the network's width experiments, the model consisting of 64 convolution kernels in each layer achieves the best PSNR and FID (Table. 4.3). Wider networks do not improve the performance (e.g. $w = 128$) and even crash the training when $w = 256$. Similarly, in the experiments of the depth of the network, additional residual blocks over 16 rarely help (Table. 4.4).

The experiments of apposite architectures searching for LR feature extraction might explain the impacts of skip connections, depth and width of networks in medical image SR tasks. First, using extra skip connections is a double-edged sword. These connections in dense blocks, such as RDB and RRDB, add more information flows. These extra pathways pass gradients more efficiently and effectively to each layer during backpropagation. Thus, the model (e.g., RDN) can perform very well, especially for minimising straightforward errors such as L1 loss. However, the dense connections also make the model liable to getting stuck in specific points and insensitive to uncertain losses such as GANs. As a result, smaller models with fewer connections, such as EDSR-lite, achieve comparable representation ability. Second, wider models are not necessary for medical images. RDN and the original EDSR have much more feature maps than EDSR-lite in each convolutional layer, which should be more potent in simulating and extracting features of nature images in SR tasks. However, too many feature maps are overqualified for medical images with limited size and relatively lower contrast information. As a result, I use EDSR-lite, consisting of 16 residual blocks and 64 convolution kernels in each layer, because it has the fewest parameters and achieves equal performance with bigger models.

Table 4.3: 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 hyperparameter of the method. I have chosen $w = 64$ because of its best PSNR and FID scores.

$w =$	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 4.4: Effects of the depth of the network, which represents the number of residual blocks b in the network. The unit of the number of parameters is in a million. Higher PSNR, SSIM and lower FID denote better performance. Bold texts represent the best performance, and the grey column denotes the hyperparameter of the method. I finally choose $b = 16$, because extra blocks rarely improve the performance but lead to more parameters linearly.

$b =$	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

4.4.3 Ablation study on the loss functions

Referring to Equation 4.2 and 4.12, the joint loss function $\mathcal{L}_{SR}$ plays an essential role in the training of MIASSR. In the three components, $\mathcal{L}_1$ represents the pixel-wise errors, while $\mathcal{L}_{perc}$ and $\mathcal{L}_{adv}$ denote the visual dissimilarity of the entire images. Particularly, the perceptual loss $\mathcal{L}_{perc}$ considers the general visual features of images because it relies on the well-trained VGG network with plenty of normal images. In contrast, the adversarial loss $\mathcal{L}_{adv}$ focuses more on the training dataset's inner features. Notice that there are two stages during training: "warm-up" and the GAN training. In the first stage, only the generator is trained with the reconstruction accuracy loss (i.e. L1 loss), leading to a basic understanding of the data distribution. This is good for training stability, preventing early discriminator dominance and providing more informative feedback on the GANs. The ablation study on the loss components only relates to the second training stage. Thus, all experiments start with the same generator after warm-up.

To achieve the best performance and understand each component's impact well, I have compared the SR performances with various scales of each loss. I first set $\lambda = 1$, then test different values of γ and η. Regarding the perceptual loss, I test $\eta = (0, 0.006, 0.01, 0.1, 1)$ and $infinity$ (Table. 4.5). 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 $infinity$ are tested for the adversarial loss. None of these values in the experiments leads to the best PSNR, SSIM and FID simultaneously. However, when $\gamma = 0.001$ and $\eta = 0.006$, it performs well on both fidelity and perceptual evaluations.

Regarding the variations of adversarial loss, I have also compared four popular GAN variations, which are widespread in SOTA SISR studies: vanilla GAN [101, 114], RaGAN [106, 134], WGAN [24, 108] and WGANGP [22, 24]. They are trained with the same hyper-parameters from the same start point of a pre-trained generator but with different designs of $\mathcal{L}_{adv}$. The adversarial losses of the vanilla GAN, WGAN and WGANGP are

defined in Equation 4.15, 4.16 and 4.17, respectively. The adversarial loss of RaGAN is:

$$\begin{aligned}\mathcal{L}_{RaGAN} &= \mathbb{E}_{\boldsymbol{I}_{hr}}\left[\log D(\boldsymbol{I}_{hr}) - \mathbb{E}_{\boldsymbol{I}_{lr}}\left[\log D(G(\boldsymbol{I}_{lr}))\right]\right] \\ &+ \mathbb{E}_{\boldsymbol{I}_{lr}}\left[\log(D(G(\boldsymbol{I}_{lr}))) - \mathbb{E}_{\boldsymbol{I}_{hr}}\left[\log D(\boldsymbol{I}_{hr})\right]\right].\end{aligned} \tag{4.19}$$

In the experiments (Fig. 4.10), WGANGP helps MIASSR achieve the best performance with all three metrics, so it is chosen in this work.

Table 4.5: Effects of the perceptual loss. In this experiment, I 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 hyperparameters of the 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 4.6: Effects of the adversarial loss. In this experiment, I 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 hyperparameters of the 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

4.4.4 Training tricks

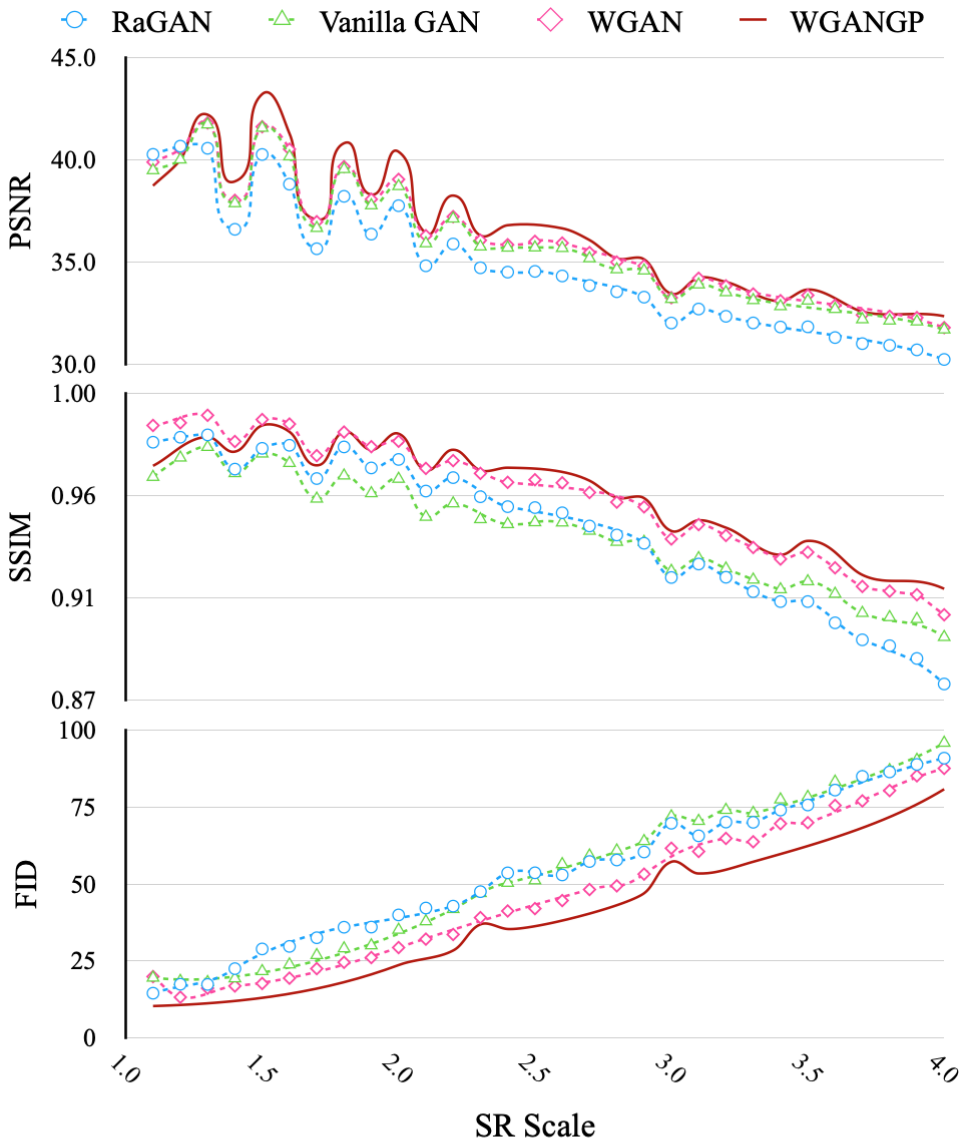
Data processing tricks can significantly affect model training and the final performance. First, transferring the well-trained model on the OASIS dataset to new datasets can accelerate the training process, although it does not improve the final performance. Second, when trained from scratch, the networks, including the generator and discriminator, must be initialised by uniform functions (e.g., Kaiming-Uniform [70]). Initialising networks with a normal distribution (e.g., Kaiming-Normal [70]) crashes the training process in the experiments. Third, batch normalisation [72] should not be used, although it has succeeded in a wide range of image processing tasks. As the only method with batch normalisation in the comparison study, SRGAN performs poorly in the patient-wise experiment (Table. 4.1) probably because the normalisation operation distorts the patient-wise contrast information

during training and leads to poor performance on the test set. Finally, hyper-parameters such as patch size and image size also affect the final performance. In Fig. 4.5, although all methods tend to perform better with smaller magnification scales than bigger ones, the best PSNR and SSIM scores always appear with the $\times 1.5$ SR task. A potential reason might be that 1.5 is the minor scale to fit the training patch size [96×96] and testing image shape [144×180] synchronously.

4.4.5 Limitations and future work

The dataset used for this study is limited to single-modality 2D radiology images (i.e. MR and CT) with simulated down-sampling and additive noise. This restricts the general applicability of the findings to clinical medical image enhancement applications. It is worth exploring the method for a broader range of medical image modalities and data types. Modifying MIASSR with techniques such as 3D convolution and recurrent networks may extend it to 3D images [311] and temporal scans [312]. Meanwhile, I have tasted cross-modality with naive transfer learning (e.g. from brain MR images to cardiac MR and chest CT images) and multi-task training (multi-modal brain MR scans). It is worth exploring with more specific cross-modality applications. Thirdly, the conclusions and findings may also benefit other medical image analysis research studies such as synthesis [15], reconstruction [14] and segmentation [313]. Finally, conclusions of training tricks in Section 4.4.4 are also restricted with the dataset and simulation experiments. For example, although the conclusion of avoiding batch normalisation is the same as other SR works [133], it may be incorrect with other scenarios [314].

Although FID has been used to represent the visual reality of generated images, it is still an open problem of perceptual quality evaluation. With this work published, there were rare research works involving FID in medical image analysis or discussing the credence of using FID for medical image evaluation. Fortunately and excitingly, in the following research of human assessment and image quality metrics on GAN-generated MR images [315], FID has shown a good correspondence with human behaviour. Meanwhile, further research on the perception-distortion trade-off (Fig. 4.4) will be helpful for research and clinical applications. However, explaining the FID score within the clinical process is unclear. Finding more straightforward and task-specific measurements of medical image perceptual quality is always desired. For example, evaluating the images in specific tasks such as AD diagnosis [2] and Parkinson disease identification [316] could be a possible way. In the following Chapter 5, I will introduce my work on involving segmentation accuracy for SR image evaluation.



	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**

Figure 4.10: Comparing GAN variations in MIASSR. Higher PSNR and SSIM indicate better fidelity quality, while lower FID denotes more perceptual realistic images are generated. WGANGP has achieved the best mean performance (the table at the bottom) and performed the best with almost all SR scales.

4.5 Chapter summary

This chapter presents a CNN-based framework for medical image arbitrary-scale super-resolution tasks. It is the first attempt to develop a meta-learning scheme with adversarial learning for this problem. The proposed method has reduced the model size (only 26% parameters compared with the Meta-SR) by using a lightweight EDSR model as the LR image feature extractor and achieved comparable reconstruction fidelity of SR images with SOTA methods. Meanwhile, it involves WGAN-GP to improve the perceptual quality of the generated images. Moreover, this approach has obtained good practical applicability. It was successfully applied to T1-weighted brain MR images, and multi-modal brain MR scans in the experiments. Furthermore, with transfer learning, the pre-trained model on brain images has been efficiently modified to cardiac MR and chest CT scans. I have also discussed the findings and understanding of model architecture design, training tricks and adversarial learning in the comparison studies.

CHAPTER 5

Residual Dense Swin Transformers for Medical Image Super-Resolution

5.1 Introduction

It is still an open problem to introduce vision transformers (ViTs) to medical image SISR tasks, which achieve state-of-the-art performance on a wide range of natural image restoration tasks [117] and medical image analysis tasks [118, 317]. To preserve sensitive information and to enhance the structures of interest for radiologists and physicians in medical image super-resolution tasks, existing vision transformers for natural images must be modified on training datasets, loss functions, evaluation metrics and architecture design [7]. Thus, the robustness, capacity, efficiency and limitation of vision transformers on medical image SR tasks will be discussed in this section.

First, acknowledged tricks of CNN architecture design, such as localisation operation, residual connection and feature fusion, are worth introducing to CNN-ViT hybrid models for potential performance improvement. For example, inspired by the shared weights and localisation operations of CNNs, a shifted window vision transformer (Swin Transformer [119]) is proposed for high-level image processing tasks. The novel Swin layers are then applied in natural image restoration [98, 138] and segmentation [96, 318].

Additionally, prior knowledge of related medical image tasks, such as segmentation, can benefit the upstream super-resolution task. On the one hand, the challenges of performing image quality assessment (IQA) on enhanced images [156] and on medical images [17] still exist. Generally, IQA of generated and super-resolved natural images mainly

includes reconstruction accuracy and human perception. Since current SISR methods are getting close to the limitation of signal fidelity metrics [319], perceptual quality assessment methods [157] have become more and more critical. However, various artefacts in medical images, mainly caused by the hardware of imaging systems and the body motion of individuals, are never seen in natural images. Peak signal-to-noise ratio (PSNR) and structure similarity (SSIM [158]) are prevalent in almost every medical image SR work. However, directly and only using the IQA methods designed for natural images may not be reliable in medical image SR tasks. In a supplemental manner, researchers evaluate the quality of SR images with the performance of downstream medical image analysis tasks such as segmentation [20]. Although the quality measurement of medical images does not equal diagnostic accuracy [19], radiologists and medical consultants always prefer high-quality images for accurate diagnosis. In addition to the measurement of machine perception, the prior knowledge of pre-trained segmentation models can also benefit the training of medical image super-resolution models, similar to the existing perceptual losses [139, 145].

In summary, this chapter has two main research questions. First, how to achieve superior SR performance and improved model efficiency with CNN-ViT hybrid networks? Second, how to involve the prior knowledge of segmentation tasks for super-resolution network training and result evaluation? I explore the possibility of extending successful architectures in CNNs to vision transformers to improve the single-image super-resolution performance of medical images efficiently and robustly. I propose a Residual Dense Swin Transformer (RDST) as a novel backbone for SR tasks by introducing residual dense connections [91, 220] and local feature fusion [132, 134] to SOTA vision transformers. Meanwhile, I take segmentation as a typical medical image analysis task in the clinic and connect it with the upstream super-resolution task for model training and result evaluation. I present a perceptual loss based on the prior knowledge of the pre-trained segmentation U-Net [92] and extend its variants to a wide range of SOTA SISR models, including CNNs and ViTs. I focus on supervised super-resolution with a single magnification scale (i.e. $\times 4$) in this work. Meanwhile, the proposed method can extend to semi-/un-supervised SR tasks with multi or arbitrary magnification scales. For a comprehensive comparison with SOTA SISR methods, I run experiments on four big and small public medical image datasets, including brain MR images, cardiac MR images and CT scans of COVID patients. Ablation studies are also designed to discuss the impacts of critical characteristics of the proposed model architecture, perceptual loss and training tricks.

Based on a comparison study of state-of-the-art single-image super-resolution methods on four public medical image datasets, I claim that the main contributions of this work

include:

- A Residual Dense Swin Transformer (RDST) is proposed by introducing the residual dense connections to vision transformers. In the $\times 4$ SR experiments on four medical image datasets, it achieves the best PSNR scores of 6 modalities among 7 modalities in total. It leads to +0.09 dB PSNR improvement on average than the SOTA SISR method SwinIR with only 38% parameters. Additionally, the SR results of RDST achieve the best segmentation accuracy of 8 sub-regions among all 15 target regions in the downstream segmentation tasks and increase the dice coefficient by 0.0029 on average than the SR results of SwinIR.

- The lite version RDST-E further improves the model efficiency with hyper-parameter modification. It achieves comparable performance with the SOTA method SwinIR on both SR image quality (+0.06 dB PSNR on average of 7 medical image modalities) and downstream segmentation accuracy (-0.0026 dice coefficient on average of 15 target regions) but has only 20% parameters of SwinIR and is 46% faster than SwinIR on inference.

- Two variants of SR perceptual loss are proposed with pre-trained segmentation U-Nets, dramatically improving the SR image quality by transferring prior knowledge of medical images in segmentation tasks to super-resolution tasks. The proposed losses successfully extend to various SOTA SISR methods, including CNNs and ViTs. Compared with the native L1 loss, the novel loss variant for SR fidelity (i.e. $\mathcal{L}_{E(1)}$) results in a noticeable improvement of +0.14 dB PSNR on average, while the proposed loss variant for machine perception (i.e. $\mathcal{L}_{HRL}$) leads to an improvement of +0.0023 dice coefficient on average in the downstream segmentation task.

This chapter is organised as follows: in Section 5.2, I introduce the proposed residual dense vision transformer and the segmentation-based perceptual loss; in Section 5.3, I describe the experiment settings; in Section 5.4, I illustrate the qualitative and quantitative results and discuss the essential characteristics of the proposed method in contrast with SOTA SISR methods; and in Section 5.5 I provide concluding remarks of this work. All related publications and code are publicly realised on `https://github.com/GinZhu/RDST`.

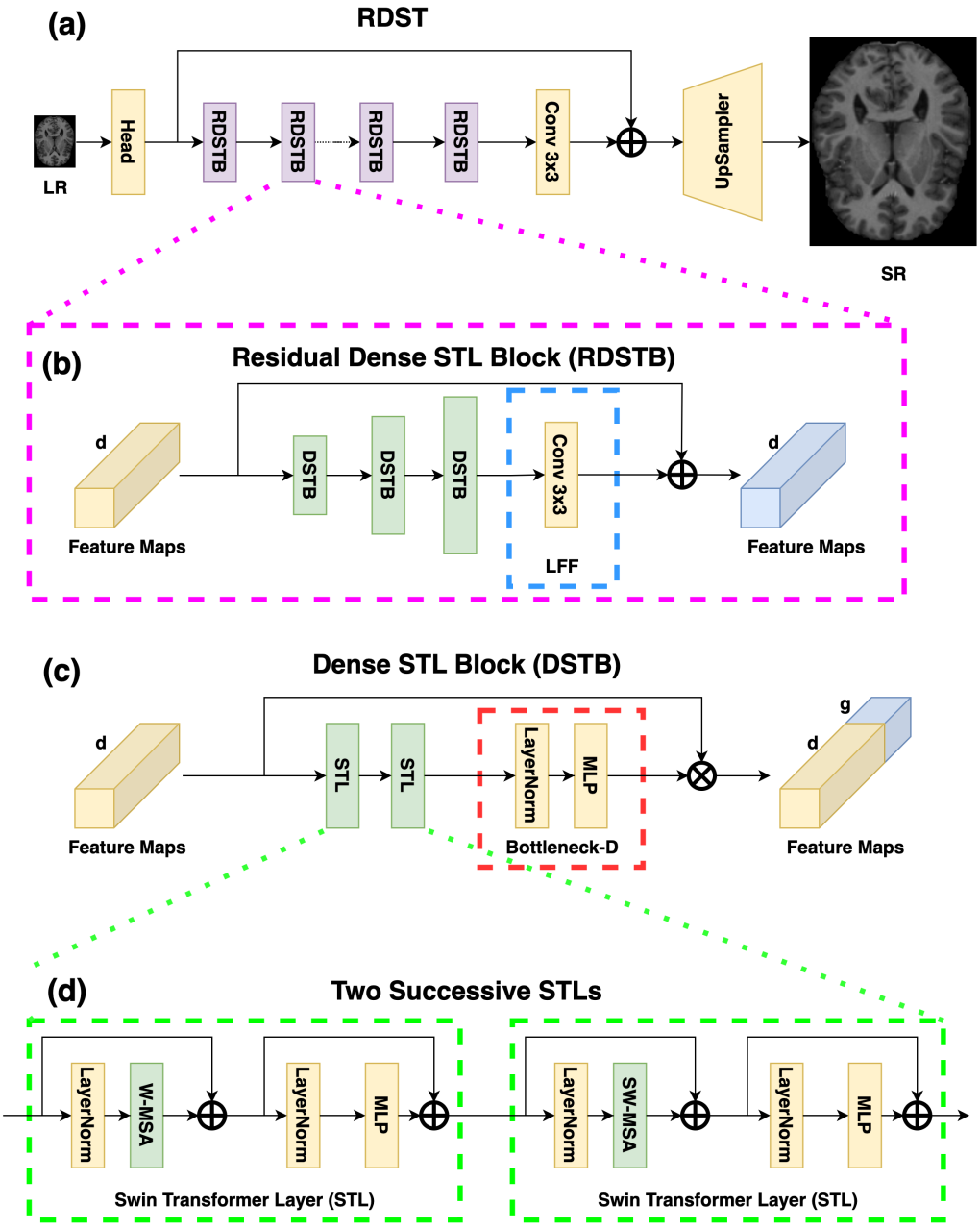


Figure 5.1: Framework of the proposed RDST network. (a): the proposed RDST (notation represented with Eq. 5.2, 5.3 and 5.4) consists of a convolution layer head for shallow feature extraction, a SubPixel-based UpSampler, N RDSTB modules and a global residual connection. (b): A residual dense Swin transformer block (notation presented with Eq. 5.7) is composed of three DSTB modules and a local feature fusion module (LFF), which compress the feature maps from $(3 \times g + d)$ to d. (c): A Dense STL Block (notation presented with Eq. 5.6) is composed of two successive swin transformer layers, a bottleneck layer and a concatenation operator. (d): The two successive shifted-window transformer layers (notation presented in Eq. 5.5). Each STL consists of two-layer normalisation layers, one multi-head self-attention layer with regular or shifted windowing configurations (W-MSA and SW-MSA), one MLP layer and skip connections. The patch embedding and un-embedding operations are ignored for a brief illustration. They convert feature maps from $[N \times d \times H \times W]$ to $[N_w \times P \times d]$ and vice visa to adjust linear layers and the convolution layers.

5.2 Methods

In this work, I mainly focus on single image super-resolution tasks with certain magnification scales (e.g. $\times 4$), which can be represented as:

$$\boldsymbol{I}_{sr} = G_s(\boldsymbol{I}_{lr}; \boldsymbol{\theta}_G), \tag{5.1}$$

where s is the magnification scale, $\boldsymbol{I}_{lr}$ and $\boldsymbol{I}_{sr}$ are a pair of one input image with a low resolution of $[H \times W \times C]$ and its super-resolved output with a high resolution of $[sH \times sW \times C]$. Following the most popular and successful architecture of SISR networks, the proposed residual dense Swin transformer consists of three components: a convolutional layer consisting of shallow feature extraction head $\mathcal{H}(\cdot)$, a feature map up-sampler $\mathcal{M}$ and a main body for deep feature extraction (Fig. 5.1-a). Mathematically, the reconstruction of a super-resolved image with increased resolution can be represented as three steps. First, the LR input is embedded in shallow features with the CNN-based head:

$$\boldsymbol{F}_{lr} = \mathcal{H}(\boldsymbol{I}_{lr}), \tag{5.2}$$

where $\boldsymbol{F}_{lr}$ is the shallow feature maps with shape $[H \times W \times d]$ with the embedding dimension d. Then, the CNN-ViT hybrid stem extracts deeper features with global feature fusion:

$$\boldsymbol{F}_d = \boldsymbol{F}_{lr} + \mathcal{C}_{k\times k}(\mathcal{B}^n(\boldsymbol{F}_{lr})), \tag{5.3}$$

where $\boldsymbol{F}_d$ is the deep feature maps remain the same shape as $[H \times W \times d]$, $\mathcal{C}_{k\times k}$ is a convolutional layer for global feature fusion with kernel size k and $\mathcal{B}^n$ indicates n successive blocks. Finally, the up-sampler increases the resolution of the feature maps and reconstructs the output image with the higher resolution $[sH \times sW \times C]$:

$$\boldsymbol{I}_{sr} = \mathcal{M}_s(\boldsymbol{F}_d). \tag{5.4}$$

Similar to SOTA SISR methods, I use one convolutional layer as the head and a Sub-Pixel [189] module as the up-sampler for super-resolution image generation. Regarding the main body for deep feature extraction, I use a skip connection for global residual learning and propose a residual dense swin transformer block (RDSTB), which will be introduced in detail in the following.

5.2.1 Residual dense swin transformer block

Shifted-windows transformer layer (STL) is used as the most basic unit in the proposed residual dense swin transformer block. To reduce the computation cost in the vision

transformer, it splits the input feature maps of size $[H \times W]$ to windows of size $[M \times M]$ first and then applies standard multi-head self-attention localised in each window. To connect these local windows, in two successive STLs (Fig. 5.1-d), the first STL applies regular window partition from top-left, while the second STL shifts the feature maps by $(\frac{M}{2}, \frac{M}{2})$ pixels before partition:

$$\begin{aligned}
\hat{\boldsymbol{F}}^{i} &= \mathcal{A}_{\mathcal{W}}(\mathcal{N}(\boldsymbol{F}^{i-1})) + \boldsymbol{F}^{i-1}, \\
\boldsymbol{F}^{i} &= \mathcal{P}(\mathcal{N}(\hat{\boldsymbol{F}}^{i})) + \hat{\boldsymbol{F}}^{i}, \\
\hat{\boldsymbol{F}}^{i+1} &= \mathcal{A}_{\mathcal{S}}(\mathcal{N}(\boldsymbol{F}^{i})) + \boldsymbol{F}^{i}, \\
\boldsymbol{F}^{i+1} &= \mathcal{P}(\mathcal{N}(\hat{\boldsymbol{F}}^{i+1})) + \hat{\boldsymbol{F}}^{i+1},
\end{aligned} \tag{5.5}$$

where $\mathcal{A}_{\mathcal{W}}$ and $\mathcal{A}_{\mathcal{S}}$ are multi-head self-attention layers with regular and shifted window configurations, respectively. $\mathcal{N}$ is layer normalisation, and $\mathcal{P}$ is a multi-layer perceptron (MLP) consisting of two fully-connected layers with GELU non-linearity [51] in between. Skip connections with pixel-wise addition are applied after each module. The key advantages of STL are the localisation operation and shared weights, just like convolutional layers. Additionally, with the reshaping operation, the size of its output feature maps remains the same as the input feature maps (i.e. $[N \times d \times H \times W]$). Thus, it behaves like a convolutional layer, and successful designs in CNN-based SISR models can be easily introduced in STL-based models.

First, I introduce dense connection [91, 220] to STLs. As shown in Fig. 5.1-c, a dense swin transformer block (DSTB) consists of two successive STLs $\mathcal{S}^2$ and an MLP-based bottleneck module $\mathcal{T}_{d \to g}$. Before concatenating to the input feature maps $\boldsymbol{F}_d^{i-1}$, the new feature maps are compressed from $[N \times H \times W \times d]$ to $[N \times H \times W \times g]$ to reduce the computation cost further. Thus, the output of the i-th DSTB is computed as:

$$F_{d+g}^{i} = cat[\boldsymbol{F}_d^{i-1}, \mathcal{T}_{d \to g}(\mathcal{S}^2(\boldsymbol{F}_d^{i-1}))]. \tag{5.6}$$

Then, the residual dense swin transformer block (RDSTB, Fig. 5.1-b) is proposed by applying local feature fusion (LLF) [132] after stacking several DSTBs. One RDSTB consists of three successive DSTBs and a 3×3 convolutional layer for local feature fusion. As reported in SRDenseNet [220] and RDN [132], combining dense connections and LLF can preserve the feed-forward nature and extract local features without high computational costs and training problems. The convolution-based LLF controls the output information by reducing the number of feature maps from $(3 \times g + d)$ to d. As a result, the output feature maps $\boldsymbol{F}^i$ of the i-th RDSTB block remains the same shape $[N \times d \times H \times W]$ as

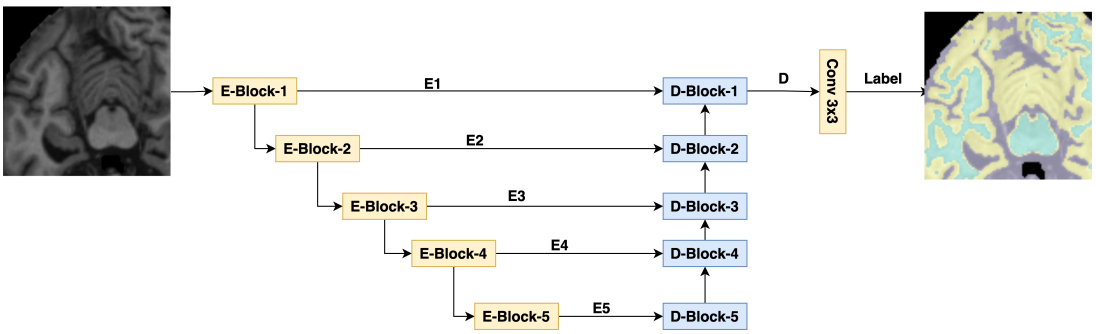


Figure 5.2: The segmentation U-Net [92] is used in this work for two purposes: for perceptual losses and segmentation-based SR evaluation. It consists of 5 levels of ResNet-based encoders and decoders, which are paired with skip connections. E1 to E5 indicate the output feature maps of each encoder block correspondingly, while D denotes the output of the last decoder.

its input feature maps $\boldsymbol{F}^{i-1}$:

$$\boldsymbol{F}_d^i = \boldsymbol{F}_d^{i-1} + \mathcal{T}_{3\times g+d\rightarrow d}(\mathcal{D}^3(\boldsymbol{F}_d^{i-1})), \tag{5.7}$$

where $\mathcal{D}^3$ is a group of three successive DSTBs and $\mathcal{T}_{3\times g+d\rightarrow d}$ is the bottleneck module for local feature fusion.

5.2.2 Segmentation U-Net based perceptual loss

The proposed method RDST is trained in two stages: basic training with $\mathcal{L}_1$ loss and a fine-tuning stage with perceptual loss. In the first stage, the parameters are optimised by minimising the native pixel-wise L1 distance between the output SR images and HR ground truth images:

$$\mathcal{L}_1(G(\boldsymbol{I}_{lr}), \boldsymbol{I}_{hr}) = \frac{1}{sH \times sW \times C} \left\| G(\boldsymbol{I}_{lr}) - \boldsymbol{I}_{hr} \right\|_1, \tag{5.8}$$

where G is a RDST model, s is the magnification scale, $\boldsymbol{I}_{lr}$ is the input low resolution image with shape $[H \times W \times C]$ and $\boldsymbol{I}_{hr}$ is the corresponding ground truth high resolution image.

In the second stage, a U-Net [92] based perceptual loss is proposed to fine-tune the parameters of the RDST after stage 1. Depending on the dataset, a U-Net model has been first trained for medical image segmentation with the same training data $\boldsymbol{I}_{hr} \in \mathbb{H}^{sH\times sW\times C}$ and the corresponding segmentation labels $\boldsymbol{L}_{hr}$:

$$\hat{\boldsymbol{\theta}_U} = arg \min_{\boldsymbol{\theta}_U} \mathcal{L}_{seg}(U(\boldsymbol{I}_{hr}), \boldsymbol{L}_{hr}), \tag{5.9}$$

where U is the segmentation model, $\boldsymbol{\theta}_U$ represents its trainable parameters and $\mathcal{L}_{seg}$ is a loss function for segmentation tasks.

Inspired by previous work on perceptual losses for SISR tasks [114, 134, 139], I define the segmentation-based perceptual loss between two images $(\boldsymbol{I}_{sr}, \boldsymbol{I}_{hr})$ as the L1 distance between their feature maps of specific layers in the pre-trained U-Net. Layers of the U-Net have learned various features at different levels from the segmentation task. Earlier encoders usually capture basic geometric shapes such as edges and textures. Deeper encoders represent more abstract features like organ parts and contextual information. The final output directly indicates the semantic label of each pixel. Feeding the generated and ground-truth images to the U-Net and calculating the distance between their hidden features can represent the semantic and perceptual similarities. This perceptual loss can have various formats, depending on the desired output quality and which level of feature maps are used (Fig. 5.2):

$$\mathcal{L}_{E(i)} = \mathcal{L}_1(U[E_i](G(\boldsymbol{I}_{lr})) - U[E_i](\boldsymbol{I}_{hr})), \tag{5.10}$$

$$\mathcal{L}_{D} = \mathcal{L}_1(U[D](G(\boldsymbol{I}_{lr})) - U[D](\boldsymbol{I}_{hr})), \tag{5.11}$$

where $U[E_i]$ indicates the i-th block of the encoder and $U[D]$ is the decoder. Furthermore, the differences between the predicted segmentation labels of $\boldsymbol{I}_{sr}$ and $\boldsymbol{I}_{hr}$ can represent the expected performance of the SR results in the downstream segmentation task, which is crucial in the clinical process. During training, the gradients will be calculated with a segmentation loss $\mathcal{L}_{seg}$ passed by the U-Net to the SR network:

$$\mathcal{L}_{HRL} = 1 - \mathcal{L}_{seg}(U(\boldsymbol{I}_{sr}), U(\boldsymbol{I}_{hr})). \tag{5.12}$$

In this work, I use dice coefficient [320] as $\mathcal{L}_{seg}$ to evaluate the distance between two binary segmentation labels X and Y, because it is popularly used in medical image segmentation tasks [321]:

$$Dice(\boldsymbol{X}, \boldsymbol{Y}) = \frac{2\,|\boldsymbol{X} \bigcap \boldsymbol{Y}|}{|\boldsymbol{X}| + |\boldsymbol{Y}|}. \tag{5.13}$$

During model training and fine-tuning, these segmentation-based perceptual loss variants can be used with the native $\mathcal{L}_1$ loss:

$$\mathcal{L}_{SR} = \alpha\mathcal{L}_1 + \lambda\mathcal{L}_U, \tag{5.14}$$

where α and λ are scale factors and $\mathcal{L}_U$ can be one or a combination of $\mathcal{L}_{E(i)}$, $\mathcal{L}_D$ and $\mathcal{L}_{HRL}$.

5.3 Experiments

5.3.1 Data and pre-processing

Four public medical image datasets are used in this work to evaluate the SR performance and robustness of the proposed method in simulating the clinical situation as widely as possible. Experiments are designed and applied on: the OASIS [304] dataset of single-modality brain MR scans; the BraTS [280–282] dataset of multi-modal brain MR scans; the ACDC [305] dataset of cardiac MR images; and the COVID [322] dataset of chest CT scans. Notice that experiments of ablation studies are mainly conducted with the OASIS dataset because it is clean and representative in the discussion of model architectures, hyper-parameters, loss functions and model efficiency.

OASIS I randomly select 39 subjects (30 for training and 9 for testing) from the OASIS-brain dataset[1] for the $\times 4$ super-resolution simulation experiments. Each subject includes 3 or 4 T1-weighted MRI scans and corresponding segmentation labels of one patient with early-stage Alzheimer's Disease (AD). Only one scan (T88-111) is used in this work, with an original size of $[176 \times 208 \times 176]$. It includes plenty of black background regions, providing useless information and slowing down the training process. Thus, the original scans and their corresponding segmentation labels are first rotated to the axial plane and centrally cropped to 145 slices of size $[160 \times 128]$. As a result, the OASIS training dataset includes 4350 slices, and the testing dataset includes 1305 images.

BraTS The BraTS dataset[2] consists of multi-modal MRI scans of 285 patients, including 210 cases with glioblastoma and 75 cases with lower grade glioma. Scans of each patient include 4 registered MR modalities: native (T1), post-contrast T1-weighted (T1Gd), T2-weighted (T2) and T2 FLuid Attenuated Inversion Recovery (FLAIR). Manual annotations of the enhancing tumour (ET), the peritumoral edema (ED) and the necrotic and non-enhancing tumour core (NCR/NET) are also provided with each scan. In this work, I randomly select 120 patients from the BraTS dataset for training and 30 other patients for testing. In pre-processing, only slices with tumours are chosen for a fair comparison in the downstream segmentation task. As a result, there are 7333 slices for training and 1853 slices for testing. To remove the pure black background, all slices are centrally cropped to $[192 \times 192]$.

ACDC The ACDC dataset[3] includes 1.5T and 3.0T cardiac MR scans of 150 patients

[1] OASIS: https://www.oasis-brains.org/
[2] BraTS: https://www.med.upenn.edu/cbica/brats2020/data.html
[3] ACDC: https://www.creatis.insa-lyon.fr/Challenge/acdc/databases.html

consisting of 5 evenly divided subgroups: normal subjects, previous myocardial infarction, dilated cardiomyopathy, hypertrophic cardiomyopathy and abnormal right ventricle. Additionally, the contours of the left ventricle (LV), right ventricle (RV) and myocardium are manually drawn and double-checked by two independent experts with more than 10 years of experience. These segmentation labels of 100 patients are released to the public. In this work, I randomly divide these 100 patients for training (80 patients with 1462 slices) and testing (20 patients with 373 slices). For a fair comparison, all slices are first centrally cropped to $[128 \times 128]$.

COVID-19 CT The COVID-19 CT dataset[4] includes 3D CT scans with left lung, right lung, and infection annotations of 20 COVID-19 patients. The proportion of infections in the lungs ranges from 0.01% to 59%. Annotations of the left lung, right lung and infection are manually labelled by experienced radiologists. In total, there are 300+ infections with 1800+ slices of various shapes. In this work, I uniformly crop a $[512 \times 512]$ region in the centre of each slice and randomly divide all scans to the training dataset (16 scans with 2264 slices) and the testing dataset (4 scans with 588 slices).

HR-LR image pair generation The original slices are used as high-resolution ground truth images $\boldsymbol{I}_{hr} \in \mathbb{H}^{H \times W \times c}$, and the corresponding low-resolution images are generated by down-sampling:

$$\boldsymbol{I}_{lr} = (\boldsymbol{I}_{hr} \star \boldsymbol{\delta}) \downarrow_s + \boldsymbol{n}, \forall \boldsymbol{I}_{hr} \in \mathbb{H}^{H \times W \times c} \tag{5.15}$$

where $\boldsymbol{\delta}$ is a bicubic down-sampling kernel and $\boldsymbol{n}$ is an additive Gaussian noise. I focus on $\times 4$ super-resolution tasks in this work, so the HR and LR patches are cropped with size $[96 \times 96]$ and $[24 \times 24]$, respectively.

5.3.2 Evaluation metrics

In addition to Peak Signal-to-Noise Ration (PSNR) and Structural Similarity (SSIM), the SR results of the proposed RDST and SOTA methods are also evaluated in downstream segmentation tasks. As described in Section 5.2.2, I first train a segmentation U-Net for each dataset with HR images in the training subset and their corresponding ground truth segmentation labels. Take the OASIS dataset as an example. The pre-trained U-Net achieves reliable segmentation performance on HR images in the testing dataset so that the segmentation-based SR performance measurement can be reliable. To evaluate the SR results of SOTA methods and the proposed RDST variants, I use dice coefficients of the whole region and tissues depending on each dataset. Notice that all datasets have expert

[4]COVID-19 CT: https://zenodo.org/record/3757476

annotations as ground truth segmentation labels, so the dice scores in the downstream segmentation tasks can correctly represent how SR images improve the segmentation accuracy of each organ/lesion compared to LR images. For example, the experiments with the OASIS dataset involve the dice coefficient scores on the whole brain (Dice-T), grey matter (Dice-G), white matter (Dice-W) and cerebrospinal fluid (Dice-CSF):

$$
\begin{aligned}
&\boldsymbol{P}_{sr} = U(\boldsymbol{I}_{sr}),\\
&\text{Dice-T} = Dice(\boldsymbol{P}_{sr}, \boldsymbol{L}_{GT}),\\
&\text{Dice-G} = Dice(\boldsymbol{P}_{sr}[C_G], \boldsymbol{L}_{GT}[C_G]),\\
&\text{Dice-W} = Dice(\boldsymbol{P}_{sr}[C_W], \boldsymbol{L}_{GT}[C_W]),\\
&\text{Dice-CSF} = Dice(\boldsymbol{P}_{sr}[C_{CSF}], \boldsymbol{L}_{GT}[C_{CSF}]),
\end{aligned} \tag{5.16}
$$

where C_G, C_W, C_{CSF} are label indexes of grey matter, white matter and CSF, respectively. Similarly, I use dice coefficient scores of the left ventricular cavity (Dice-LV), the right ventricular cavity (Dice-RV), the myocardium (Dice-MC) and the whole region (Dice-T) for the ACDC dataset and the dice coefficient scores of the left lung (Dice-LL), the right lung (Dice-RL), the lesion (Dice-Lesion) and the whole region (Dice-T) for the COVID dataset. In the experiments with the BraTS dataset, I use dice coefficient scores of the enhancing tumour (Dice-ET, including ET only), the tumour core (Dice-TC, including ET and NCR/NET) and the whole tumour (Dice-WT, including ET, ED and NCR/NET).

5.3.3 Implementation details

The proposed RDST and SOTA models are implemented with PyTorch [37]. All experiments performed on an Nvidia Quadro RTX 8000 GPU. Inspired by SwinIR [138], the window size, attention head number and the basic feature embedding dimension are set to $[8 \times 8]$, 6 and 60, respectively. As mentioned in Section 5.2.1, each RDSTB module consists of 3 DSTBs with a growth rate of 30. Each DSTB module consists of 2 STLs. In this work, I propose two RDST variants. The original one consists of 8 RDSTBs, while the more efficient version (RDST-E) consists of only 4 RDSTBs. In the experiments, all parameters are initialised by Kaiming-uniform [70] and optimised by the Adam optimiser [61]. I set the batch size to 32 for each step for both training stages. In the first training stage, the initial learning rate is set to 0.0002 with no decay, and the RDST is trained for 100k steps with only native $\mathcal{L}_1$ loss. The fine-tuning stage includes 20k steps. Its learning rate is initialised as 0.0001 and halved at $[10\text{k}, 15\text{k}, 17.5\text{k}]$. The scale factors in Equation 5.14 are set as $\alpha = 1, \lambda = 10$ to ensure the segmentation-based perceptual loss dominates the fine-tuning stage. The L1 distance between feature maps of the first encoder block

$\mathcal{L}_{E(1)}$ is mainly used as the perceptual loss, and other variations of $\mathcal{L}_U$ are used in ablation study.

The segmentation U-Net model is implemented with Segmentation-Models-PyTorch [323]. It consists of 5 ResNet [90] based encoder blocks and a decoder of native convolutional layers (Fig. 5.2). The channel number is set to 64 basically and doubled after each encoder block. This model is trained with Dice loss (Equation 5.12) for 100k steps with the Adam optimiser. The learning rate is initialised as 0.0001 and halved at $[50\text{k}, 75\text{k}]$.

5.4 Results and discussion

In this section, I first illustrate the superior performance of the proposed RDST variants compared with SOTA SISR methods on four medical image datasets, then discuss the key factors of RDST twofold: first, how the novel CNN-ViT hybrid dense block improves the model efficiency; and second, how the proposed segmentation-based perceptual loss results in controllable improvements of reconstruction accuracy and perceptual quality.

5.4.1 Comparing RDST with SOTA methods

Two variations of RDST are compared with 5 popular and representative state-of-the-art SISR methods, including: (1). pure convolutional methods EDSR [133] and RDN [132]; (2) CNN based attention models RCAN [136] and HAN [137]; and (3) self-attention based vision transformers SwinIR [138]. Additionally, I have done experiments with related SR methods in a wider range, such as zero-shot super-resolution [210], scale-free super-resolution [27, 135] and pre-trained ViT [127]. However, I finally decided not to involve these methods in the comparison because of their mediocre performance.

RDST variants achieve the best image quality on all four datasets. Specifically, RDST-E achieves the best PSNR performance on the ACDC dataset, and RDST achieves the best PSNR scores on the OASIS dataset, the COVID dataset and every MR modality in the BraTS dataset. On average, RDST increases by 0.09dB in PSNR, and RDST-E leads to a 0.06dB increase compared with the most recent SOTA method SwinIR. Meanwhile, I clearly show that improving image quality can lead to significantly better performance in the downstream segmentation tasks. With the well-trained U-Net segmentation models and introducing the segmentation-based SR results evaluation, SOTA SISR methods considerably narrow the gap of segmentation accuracy between HR GT images and $\times 4$ bicubic interpolated images. In summary, RDST achieves the best dice coefficient scores of 8 targeted regions among all 15 regions. Detailed results of each dataset are as follows.

Performance on the OASIS dataset RDST achieves the best performance of almost all metrics in the $\times 4$ super-resolution experiment with the OASIS dataset (Fig. 5.3). It brings noticeable improvements in image quality (+0.18dB PSNR and +0.0012 SSIM) to SwinIR. In the downstream segmentation task, SR results of RDST get the best dice coefficient scores of the whole brain, the grey matter and the white matter (Table 5.1). The pre-trained U-Net achieves reliable segmentation performance on HR GT images. It clearly shows a notable decline with native bicubic interpolation SR results: $[0.1395, 0.2210, 0.1262, 0.1201]$ on whole brains, grey matters, white matters and CSFs, respectively. The proposed RDST narrows these gaps by $[0.0774, 0.1335, 0.0551, 0.0667]$ respectively. Notice that

room for improvement exists as the best segmentation dice scores of all SR images are still significantly lower than HR images ($[-0.0621, -0.0887, -0.0711, -0.0518]$). On the other hand, the smallest model RDST-E achieves the second-best PSNR and SSIM scores with a slight decrease in segmentation performance. While visualising the SR results and their corresponding segmentation predictions, the segmentation labels can help determine the differences between SR images of SOTA methods, which are difficult to recognise with only the images. Vision transformers (i.e. SwinIR and RDST) achieve superior image quality on edges than CNNs (green box in Fig. 5.4). On the other hand, it is still challenging for all methods to reconstruct rich textures of small regions (red box in Fig. 5.4).

Model efficiency Additionally, I compare the model efficiency in the experiment on the OASIS dataset by calculating the number of parameters and the multi-add calculations (MACs) with an $[1 \times 40 \times 32 \times 1]$ input (Table 5.1). RDST-E is the smallest and requires the fewest MACs, while RDST is the second smallest and requires fewer calculations than SOTA methods. Compared with SwinIR, RDST has only 38% parameters, and RDST-E has only 20% parameters. As a result, they reduce the computational cost by 58% and 76%, respectively.

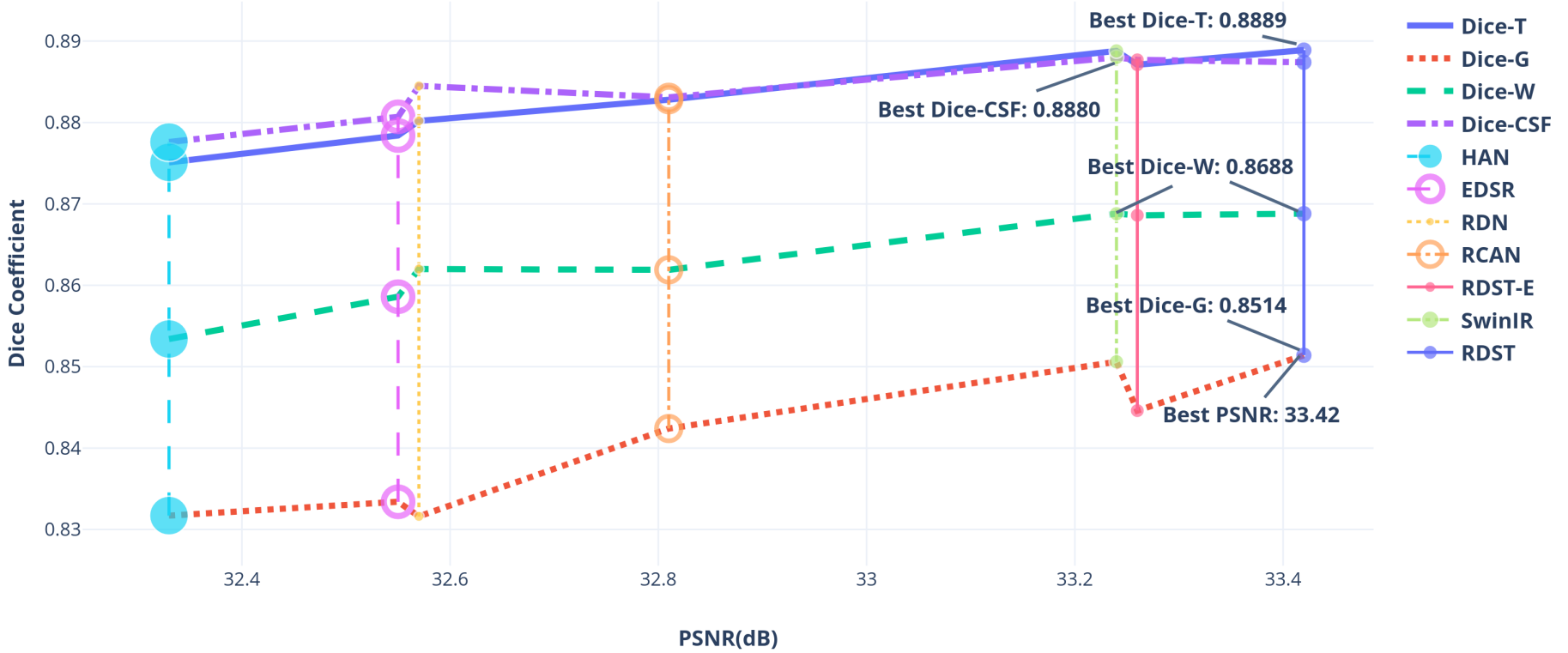


Figure 5.3: Comparing RDST and RDST-E with SOTA SISR methods on $\times 4$ super-resolution task with the OASIS dataset. The PNSR (the X-axis) indicates the SR image quality, while the dice scores (the Y-axis) represent the downstream segmentation performance. Dice-[T, G, W, CSF] specify the dice coefficient of whole brains, grey matter, white matter and cerebrospinal fluid, respectively. The best PSNR and dice scores are pointed out. Bigger symbols indicate more parameters of the models.

Table 5.1: Compare RDST with SOTA methods on the OASIS dataset. The best and second-best scores are highlighted in red and blue respectively. MACs are calculated with an input size $[1 \times 40 \times 32 \times 1]$.

Mean(std)	PSNR(dB)↑	SSIM↑	Dice-T↑	Dice-G↑	Dice-W↑	Dice-CSF↑	MACs(G)↓	params(M)↓
HR			0.9520(0.012)	0.9401(0.040)	0.9399(0.013)	0.9408(0.015)		
Bicubic	29.88(2.7)	0.8574(0.052)	0.8125(0.0088)	0.7179(0.075)	0.8137(0.021)	0.8207(0.016)		
EDSR [133]	32.55(3.5)	0.9184(0.039)	0.8784(0.0098)	0.8334(0.055)	0.8586(0.021)	0.8807(0.015)	64.22	43.08
RDN [132]	32.57(3.2)	0.9241(0.036)	0.8802(0.010)	0.8316(0.059)	0.8620(0.021)	0.8845(0.015)	7.95	5.76
RCAN [136]	32.81(3.6)	0.9224(0.038)	0.8828(0.0094)	0.8424(0.054)	0.8619(0.021)	0.8831(0.013)	41.34	32.03
HAN [137]	32.33(3.7)	0.9120(0.042)	0.8751(0.0091)	0.8317(0.055)	0.8534(0.022)	0.8776(0.015)	83.86	64.19
SwinIR [138]	33.24(3.7)	0.9287(0.036)	**0.8888(0.0097)**	**0.8506(0.054)**	**0.8688(0.020)**	**0.8880(0.015)**	14.68	11.47
RDST-E	**33.26(3.4)**	**0.9291(0.035)**	0.8871(0.0096)	0.8446(0.057)	**0.8686(0.020)**	**0.8877(0.015)**	**3.53**	**2.35**
RDST	**33.42(3.7)**	**0.9299(0.035)**	**0.8889(0.0097)**	**0.8514(0.054)**	**0.8688(0.021)**	0.8874(0.015)	**6.17**	**4.40**

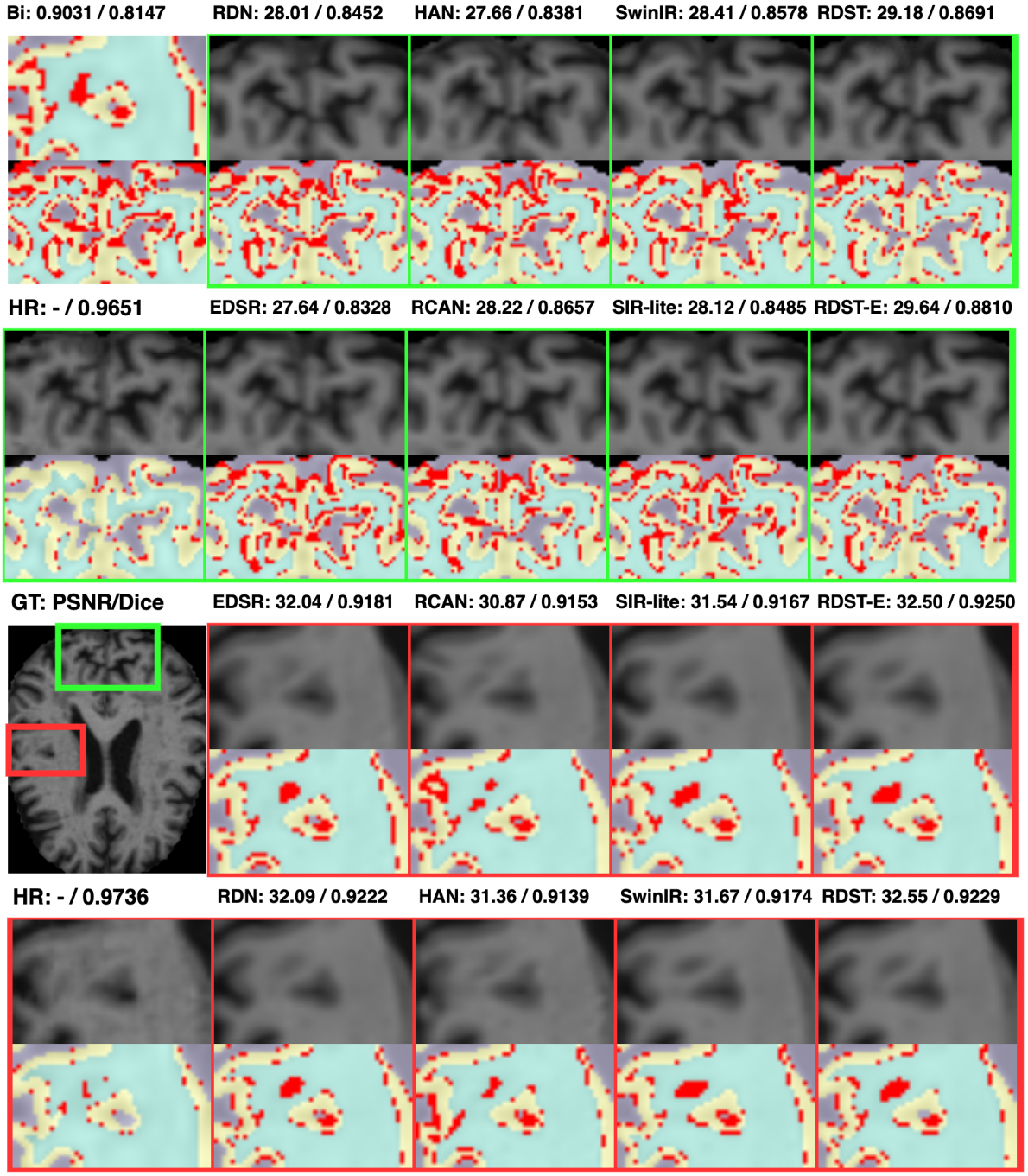


Figure 5.4: Comparing RDST with SOTA methods on $\times 4$ super-resolution task with OASIS dataset. SR results and corresponding segmentation predictions of a randomly selected slice are shown with PSNR and dice coefficient of the whole brain. In the segmentation labels, grey, yellow and cyan indicate grey matters, white matters and CSFs, respectively, and segmentation errors are marked as red.

Performance on the BraTS dataset Doing $\times 4$ super-resolution in the multi-modal brain tumour segmentation dataset is more challenging for the following reasons. First, the super-resolution method must be synchronously applied on four registered MR scans (i.e. T1Gd, T1, T2 and T2-FLAIR) so the input and output layers of RDST variants and SOTA methods are modified. Second, the downstream multi-modal tumour issue segmentation is challenging, and the U-Net with poor segmentation performance may misdirect the fine-tuning stage. In this experiment, the pre-trained U-Net has achieved dice coefficients of $[0.7830, 0.6919, 0.6820]$ on the whole tumour, the enhancing tumour and the tumour core, respectively. Compared with bicubic interpolation, SOTA deep neural networks significantly improve the SR performance (Table. 5.2) on image quality and downstream segmentation performance. The proposed RDST achieves the best PSNR scores on all modalities, and the efficient version RDST-E achieves the second-best scores on three modalities (i.e. T1Gd, T1 and T2-FLAIR). SwinIR, the SOTA vision transformer for SISR tasks, also performs better than CNN models. It dominates the SSIM scores with RDST. On average, RDST gets +0.13dB higher PSNR than SwinIR and equal SSIM (-0.0003) of the four modalities. In the downstream tumour segmentation task, RDST achieves the best dice coefficients of the whole tumours and the enhancing tumours. EDSR achieves the best segmentation performance of the tumour cores. Interestingly, SR results of SwinIR and RDST can even provide more accurate segmentation labels than HR GT images, probably because the down-sample and super-resolve process removes some noises and misleading textures. Similar to the experiments on the OASIS dataset, the segmentation labels help to recognise the most challenging part in the SR image generation: reconstructing the blur edges and irregular textures (Fig. 5.5).

Table 5.2: Comparing RDST with SOTA methods in the ×4 super-resolution task on the BraTS dataset. PSNR and SSIM scores are calculated on each modality (i.e. one of T1Gd, T1, T2, and T2-FLAIR) and averaged. Dice coefficients of the whole tumour (WT), the enhancing tumour (ET) and the tumour core (TC) are used in the downstream segmentation evaluation with the pre-trained 2D U-Net. The best and the second best scores are highlighted in red and blue, while ∗ indicates superior segmentation performance than HR GT images.

Mean(std)	Dice↑			PSNR↑				SSIM↑			
	WT	**ET**	**TC**	**T1Gd**	**T1**	**T2**	**Flair**	**T1Gd**	**T1**	**T2**	**Flair**
HR	0.7833(0.13)	0.6919(0.13)	0.6820(0.16)								
Bicubic	0.7614(0.12)	0.5226(0.20)	0.5878(0.18)	29.97(1.9)	29.04(2.3)	28.19(2.1)	29.33(2.4)	0.8571(0.035)	0.8661(0.037)	0.8509(0.037)	0.8362(0.047)
EDSR [133]	0.7800(0.12)	0.6854(0.12)	**0.6776(0.16)**	32.63(2.0)	32.36(2.4)	31.08(1.9)	31.89(2.4)	0.9143(0.024)	0.9271(0.025)	0.9156(0.024)	0.8968(0.033)
RDN [132]	0.7806(0.12)	**0.6874(0.13)**	**0.6729(0.17)**	32.97(2.0)	32.73(2.4)	31.48(2.0)	32.29(2.4)	0.9193(0.024)	0.9320(0.024)	0.9218(0.024)	0.9039(0.032)
RCAN [136]	0.7815(0.12)	0.6808(0.12)	0.6674(0.17)	32.84(2.0)	32.59(2.4)	31.29(2.0)	32.14(2.3)	0.9183(0.023)	0.9310(0.024)	0.9202(0.023)	0.9022(0.032)
HAN [137]	0.7801(0.12)	0.6833(0.12)	0.6722(0.17)	32.56(1.9)	32.24(2.4)	31.03(1.9)	31.84(2.3)	0.9155(0.024)	0.9281(0.024)	0.9168(0.024)	0.8976(0.033)
SwinIR [138]	**0.7836*(0.12)**	0.6816(0.12)	0.6710(0.16)	33.23(2.1)	33.03(2.5)	**31.73(2.0)**	32.54(2.4)	**0.9226(0.023)**	**0.9350(0.024)**	**0.9253(0.023)**	**0.9082(0.031)**
RDST-E	**0.7831(0.12)**	0.6832(0.12)	0.6713(0.18)	**33.32(2.0)**	**33.13(2.4)**	31.71(2.1)	**32.59(2.4)**	0.9218(0.024)	**0.9339(0.025)**	0.9231(0.024)	0.9066(0.032)
RDST	**0.7836*(0.12)**	**0.6883(0.12)**	0.6713(0.17)	**33.37(2.0)**	**33.22(2.5)**	**31.81(2.1)**	**32.63(2.4)**	**0.9227(0.023)**	**0.9350(0.024)**	**0.9248(0.023)**	**0.9075(0.032)**

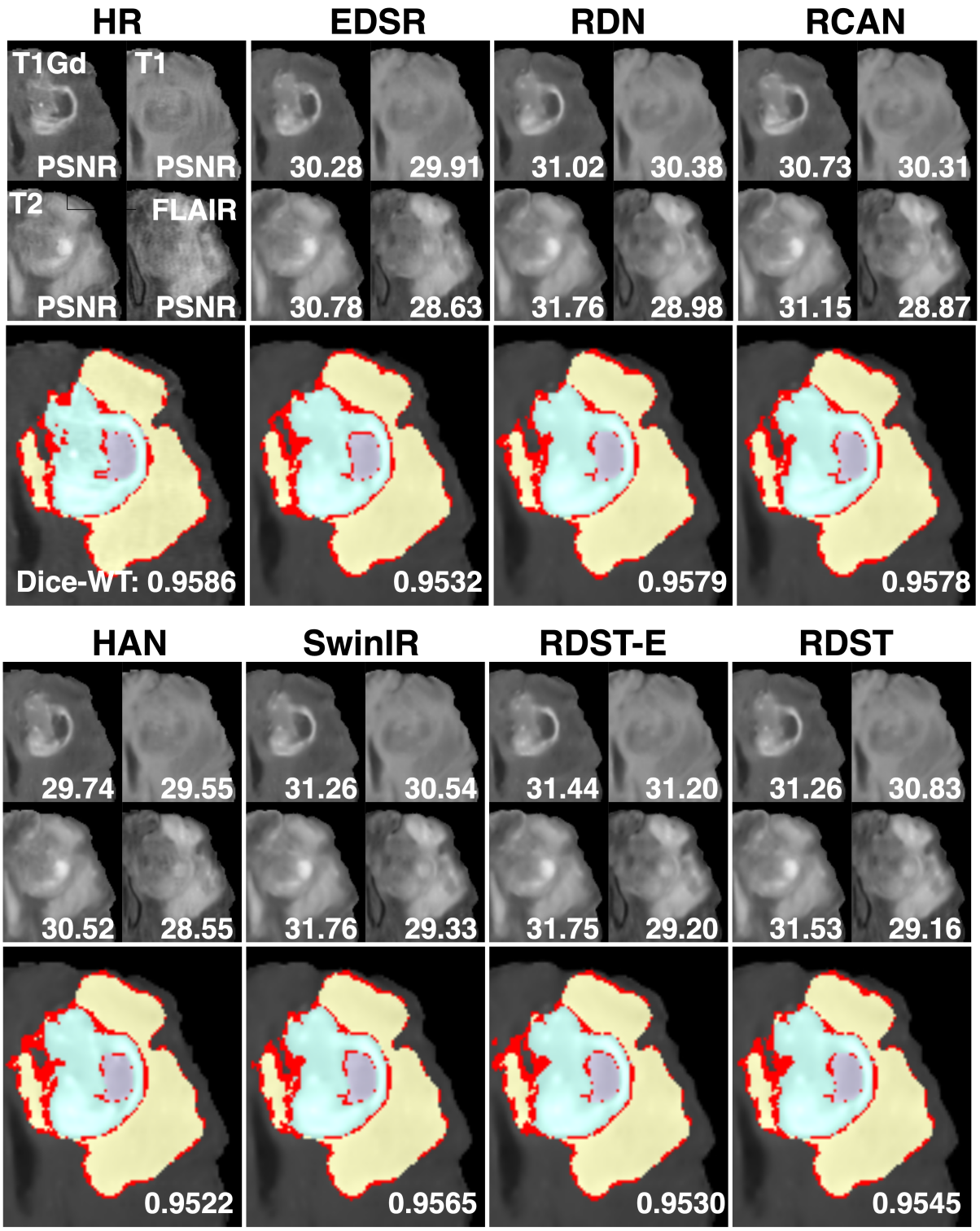


Figure 5.5: SR results of a random slice in the testing subset of BraTS. SR results of whole slices are generated, but only the regions within the tumour of the four modalities are plotted with predicted labels in the downstream segmentation task for better comparison. Annotations of tumour sub-regions are the necrotic and the non-enhancing (NCR & NET) parts of the tumour in yellow, the peritumoral edema (ED) in cyan and the enhancing tumour in grey. Segmentation errors are indicated in red. PSNR and dice coefficient of the whole tumour (Dice-WT) are also displayed.

Performance on the ACDC dataset Single image SR of cardiac MR images is challenging because the motion artefacts caused by patients' atrial fibrillation during the scanning procedure are individual and hard to resolve. As a result, data-driven deep learning methods can rarely bring a dramatic improvement in PSNR than traditional interpolation methods. In the comparison study of $\times 4$ magnification (Table. 5.3), SOTA methods, including RDST variants, increase PSNR from +0.80 dB to +1.10 dB than bicubic interpolation. In contrast, SOTA SISR methods can easily lead to more than 3 dB improvement of PSNR on other datasets. Additionally, because the training data is limited (only 1462 slices with size $[128 \times 128]$), over-fitting may happen. As a result, the smallest model RDST-E achieves a significant advantage of PSNR (+0.20 dB higher than other methods). However, it is hard to evaluate the SR performance with only PSNR because most methods achieve very close scores from 27.03 dB to 27.06 dB. I guess the low PSNR scores of RDST and SwinIR are caused by the over-fitting of background noise, which leads to substantial pixel-wise errors but rare impacts in segmentation and visualisation (Fig. 5.6). In contrast, evaluating SR results with the dice coefficient scores and SSIM scores is more effective and robust. The segmentation U-Net is well-trained for the ACDC dataset with reliable performance on HR GT images. Meanwhile, the segmentation-based evaluation represents the global structure reconstruction accuracy in SR results. In this experiment, vision transformers (i.e. SwinIR, RDST-E and RDST) achieve the highest SSIM and perform the best in the downstream segmentation task, so I claim that they are better than the CNN-based methods.

Table 5.3: Comparing RDST with SOTA methods in the $\times 4$ super-resolution task on the ACDC dataset. In addition to PSNR and SSIM, dice coefficients of the whole region (T), the left ventricular cavity (LV), the right ventricular cavity (RV), and the myocardium (MC) in the downstream segmentation task are used for evaluation. The best and the second-best scores are highlighted in red and blue, respectively.

Mean(std)	PSNR↑	SSIM↑	Dice-T↑	Dice-LV↑	Dice-RV↑	Dice-MC↑
HR			0.8932(0.027)	0.9184(0.034)	0.7390(0.11)	0.8783(0.036)
Bicubic	26.14(2.7)	0.7501(0.053)	0.8096(0.051)	0.8717(0.059)	0.5927(0.15)	0.7697(0.058)
EDSR [133]	26.94(3.1)	0.7722(0.064)	0.8599(0.030)	0.8950(0.044)	0.6897(0.12)	0.8354(0.044)
RDN [132]	**27.07(3.0)**	0.7854(0.058)	0.8624(0.029)	**0.8979(0.038)**	0.6946(0.12)	0.8376(0.039)
RCAN [136]	27.06(3.0)	0.7819(0.061)	0.8657(0.030)	0.8947(0.044)	0.6867(0.14)	0.8399(0.042)
HAN [137]	27.06(3.4)	0.7738(0.069)	0.8666(0.030)	0.8946(0.043)	0.6971(0.13)	0.8419(0.046)
SwinIR [138]	27.04(3.1)	**0.7876(0.059)**	**0.8705(0.026)**	**0.9001(0.043)**	0.6964(0.12)	**0.8456(0.039)**
RDST-E	**27.24(3.0)**	**0.7940(0.057)**	**0.8691(0.025)**	0.8954(0.043)	**0.7008(0.12)**	**0.8454(0.035)**
RDST	27.03(3.1)	0.7875(0.059)	**0.8691(0.027)**	0.8959(0.043)	**0.7071(0.11)**	0.8433(0.035)

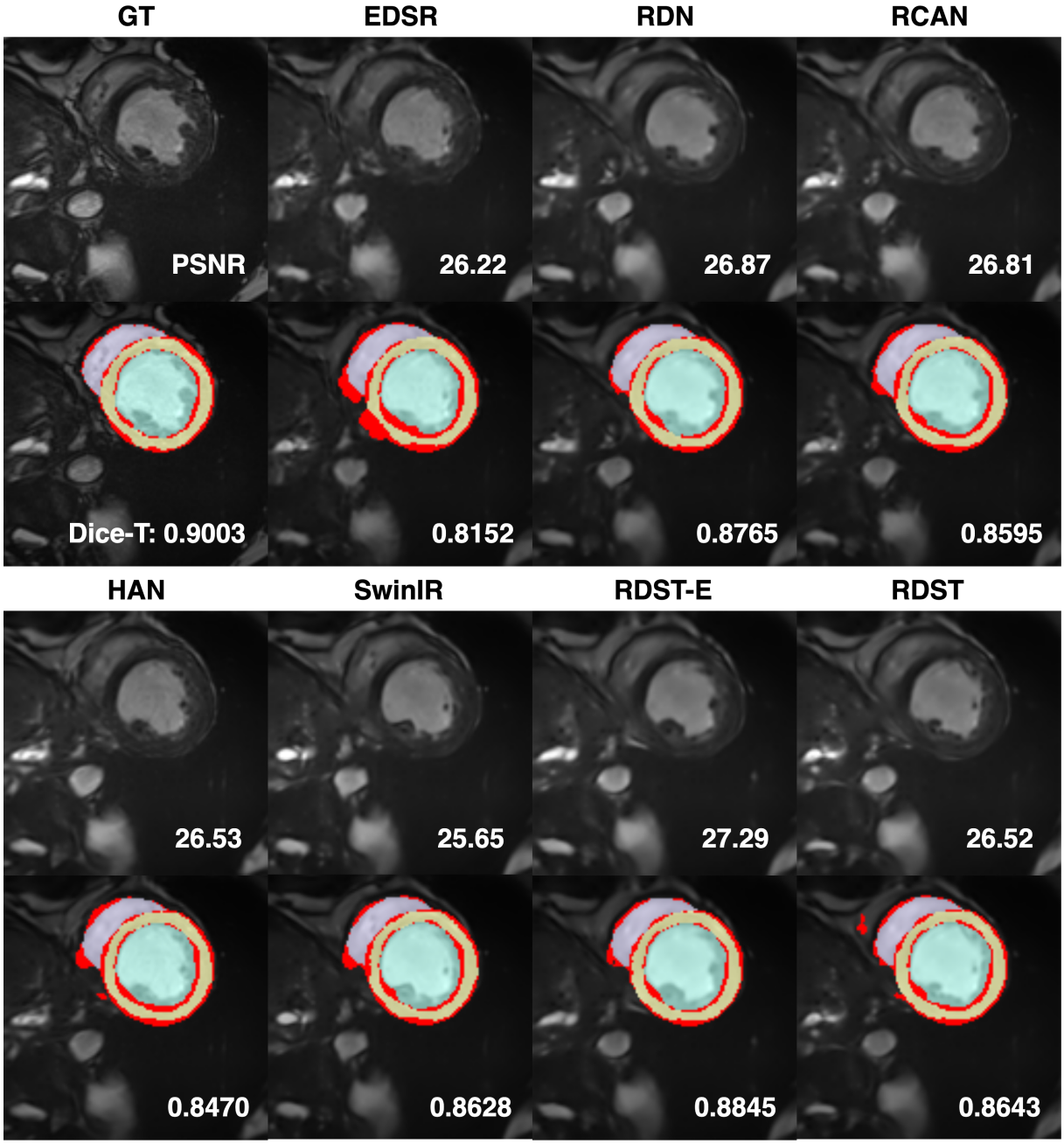


Figure 5.6: SR results of a random slice in the testing dataset of ACDC. Accurate segmentation predictions are annotated in grey (the RV cavity), yellow (the myocardium) and cyan (the LV cavity), while wrong predictions are annotated in red. PSNR scores and dice coefficients of the whole region are shown.

Performance on the COVID dataset I also extend the proposed method to CT images. Compared with MR scans, CT scans are with higher resolution (e.g. $[512 \times 512]$) and fewer artefacts. All methods achieve very close PSNR (from 34.56 dB to 34.70 dB) and SSIM (from 0.8678 to 0.8707) scores in the $\times 4$ SR experiment (Table. 5.4). Specifically, RDST achieves the highest PSNR score and the best segmentation results of the whole region and the right lung. Meanwhile, RCAN achieves the best SSIM, and HAN achieves the best segmentation accuracy of the left lung and the lesion. Notice that both methods are with very deep architectures (> 800 layers) and with more than 30M parameters, which may benefit the reconstruction of the textures in lesion areas (Fig. 5.7).

Table 5.4: Comparing RDST with SOTA methods in the ×4 super-resolution task on the COVID-CT dataset. In addition to PSNR and SSIM, dice coefficients of the total region (T), the left lung (LL), the right lung (RL), and the lesion in the downstream segmentation task are used for evaluation. The best and the second-best scores are highlighted in red and blue, respectively.

Mean(std)	PSNR↑	SSIM↑	Dice-T↑	Dice-LL↑	Dice-RL↑	Dice-Lesion↑
HR			0.8762(0.11)	0.8553(0.17)	0.8905(0.10)	0.6554(0.035)
Bicubic	28.45(3.0)	0.7760(0.15)	0.8092(0.12)	0.7386(0.17)	0.8430(0.11)	0.3848(0.14)
EDSR [133]	34.59(4.4)	0.8694(0.15)	0.8315(0.18)	0.8265(0.20)	**0.8713(0.13)**	0.5881(0.14)
RDN [132]	34.56(4.4)	0.8678(0.15)	0.8196(0.19)	0.8167(0.21)	0.8590(0.14)	0.5674(0.13)
RCAN [136]	**34.69(4.4)**	**0.8707(0.15)**	**0.8365(0.17)**	0.8175(0.21)	0.8684(0.13)	**0.6207(0.11)**
HAN [137]	34.65(4.4)	**0.8700(0.15)**	0.8323(0.18)	**0.8294(0.19)**	0.8695(0.13)	**0.6247(0.18)**
SwinIR [138]	34.66(4.5)	0.8698(0.15)	0.8299(0.18)	0.8201(0.20)	0.8678(0.13)	0.5846(0.12)
RDST-E	34.62(4.5)	0.8678(0.15)	0.8264(0.19)	0.8134(0.21)	0.8620(0.15)	0.5709(0.088)
RDST	**34.70(4.5)**	0.8687(0.15)	**0.8445(0.16)**	**0.8281(0.19)**	**0.8761(0.13)**	0.5884(0.085)

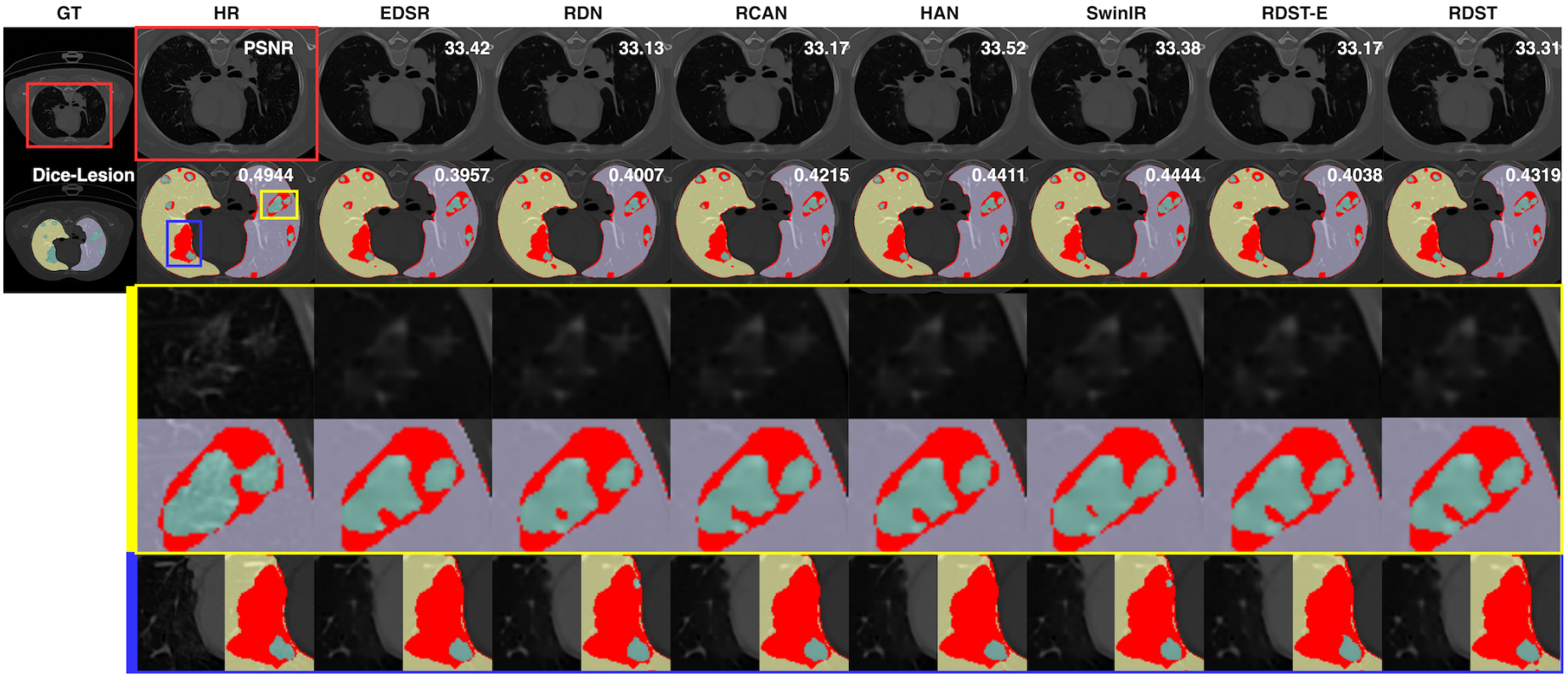


Figure 5.7: SR results of a random slice in the testing dataset of COVID-CT. PSNR scores and dice coefficients of the lesion are shown. Accurate segmentation predictions are annotated in grey (the left lung), yellow (the right lung) and cyan (the lesion), while wrong predictions are annotated in red. Patches of two sub-regions are zoomed in for better visualisation.

5.4.2 Network architecture, attention and inference efficiency

An ablation study is designed to figure out the critical factors of RDST variants achieve superior performance than SOTA SISR methods on the OASIS dataset (Tabel. 5.5 and Fig. 5.9). PSNR, SSIM and dice coefficient of the whole brain are used as SR image evaluation metrics. Meanwhile, the numbers of MACs and parameters and the throughout frame rate during inference are used to measure the model efficiency. All methods are trained with the same settings (100k steps with $\mathcal{L}_1$ only) to avoid the impacts of the segmentation-based perceptual loss and additional training steps.

Network architecture Comparing RDST variants and SOTA methods, the main differences in network architecture are fourfold. First, a model can be created with only convolutional layers or a hybrid of transformers (e.g. STL) and CNNs. Second, the window size, which presents the receptive field of each layer, can be different. Third, the model widths, which indicate the feature map dimensions, vary from 64 to 256. Fourth, one network can be shallow (e.g. EDSR-lite with only 32 layers) or deep (e.g. HAN with 1610 layers). Thus, I divide all methods into four groups: (1) native CNN SISR methods, including EDSR and RDN; (2) native CNN models with large window size, which are implemented based on ConvNet [324]; (3) deep CNN models with more than 800 layers, i.e. RCAN and HAN; and (4) STL based vision transformers such as SwinIR and RDST. I propose the ConvNet-based SISR method to discuss the impacts of the receptive field [42], which is considered the critical factor for the success of vision transformers [325]. Meanwhile, the lite versions of EDSR, SwinIR and ConvNet are also involved in comparing small SISR models with RDST-E.

STL or CNN? First, vision transformers show more dramatically improved SR image quality and segmentation accuracy than CNN methods. Notice that SwinIR and RDST variants are all based on the Swin transformer layer, which learns from the shared weights and localised operation of convolutional layers. Additionally, CNN layers are used at intervals of STL blocks which further ensures the training stability of these hybrid methods and leads to superior performance than pure CNN methods. Second, the larger window size failed to extend the success with transformers to CNN models. Both versions of the ConvNet-based methods perform worse than all the other methods. In deep networks, the receptive field depends on the window size in one layer and the number of layers. Thus, deeper networks with small window sizes can have an equal receptive field with shallow networks with large window sizes. For CNNs, the former works better probably because more non-linear activation is essential for feature extraction. Third, deeper models consistently achieve better results in each group with similar network architectures. Notice that both increases in width and depth lead to an increase in parameters and an efficiency

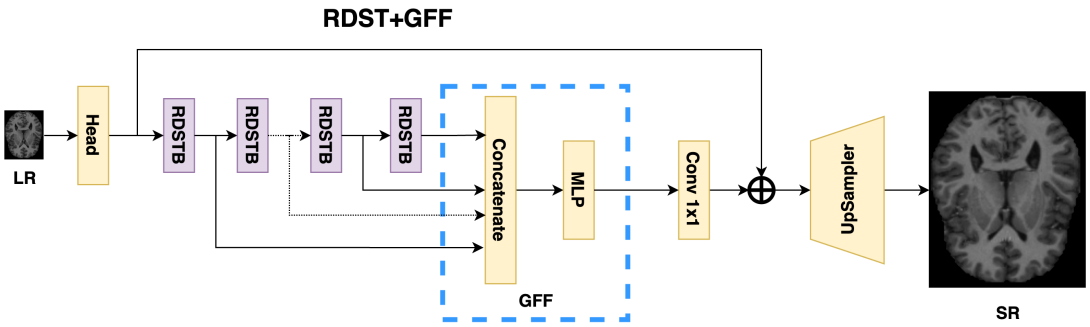


Figure 5.8: A RDST variant with MLP-based global feature fusion (GFF).

decrease. In this comparison study on medical image SR tasks, increasing the number of blocks is more effective. For example, RDST has more layers and a smaller width than SwinIR, leading to fewer computational costs and parameters. It finally achieves equal PSNR (+0.01 dB) and SSIM (-0.0001) scores with only 38% parameters of SwinIR.

Hyperparameters Additionally, experiments are designed on RDST-E to figure out the impacts of hyperparameters of RDST (Fig. 5.10). In RDST there is a gradual growth of feature map dimension because of the dense connections. I take the dimension after local feature fusion as the model width. On the other hand, a deeper RDST can be created by increasing three factors: the number of Swin transformer layers in the DSTB modules, the number of DSTBs and the number of RDSTBs. Briefly speaking, increasing the number of RDSTB modules, the window size and the model width in an adequate range can improve SR image quality.

Attention The attention mechanism is briefly regarded as threefold depending on where it works: channel attention, spatial attention and layer attention (Table. 5.5). Notice that the global feature fusion in RDN [132] is considered elementary layer attention, and the self-attention in transformers is considered a super-set of both channel and spatial attention. Attention in CNN models raises training stability and leads to very deep networks but rarely improves the medial image super-resolution performance. For example, RCAN has 811 layers, and HAN has 1610 layers, but neither performs superior to other methods. On the other hand, the self-attention in vision transformers, which introduce dynamic parameters to the whole computation process, dramatically improves the SR performance as in SwinIR[138] and RDST variations. Meanwhile, the GFF module has no noticeable improvement or decline in SR results, so it is abandoned in the final design of RDST.

Inference frame rate In addition to the number of parameters and calculations (MACs), I introduce the inference frame rate (i.e. frame-per-second FPS) as a more straightforward

evaluation of model efficiency. Compared with MR and CT scanning in the clinic, all methods can almost real-time (at least 6 slices per second on an Nvidia RTX 8000 GPU for $[40 \times 32 \times 1] \rightarrow [160 \times 128 \times 1]$ SR. In addition to MACs and parameters, the inference efficiency depends on GPU acceleration and model architecture. Thus, transformers are generally slower than CNNs, although they have fewer parameters. Additionally, the depth of each model plays a crucial role in inference efficiency because it cannot be executed in parallel. For example, RCAN is the slowest model as it has the most layers. Among the vision transformers, RDST has fewer parameters and requires less computation but is slower than SwinIR because of more layers. The main advantage of vision transformers is that self-attention activates the capability of parameters more adequately, so superior performance is achieved with less computation and shallower networks. With potential hardware acceleration support in the future, transformers may be smaller and faster than CNNs. Specifically, RDST-E has a similar size to the lite version model (e.g. EDSR-lite and SwinIR-lite) but achieves comparable performance with regular SISR methods such as SwinIR.

Table 5.5: Ablation study on the OASIS dataset to answer why transformers perform better than CNNs. PSNR, downstream segmentation dice score of the whole brain (Dice-T) and inference frame rate (FPS) are used as evaluation metrics for both image quality and model efficiency. The best (in red) and second-best (in blue) scores are in bold. All methods are divided into four groups: (1). native CNN methods include EDSR and RDN; (2). the proposed ConvNet-based SR methods with the large receptive fields in CNN; (3). very deep CNNs with attention; and (4). methods with Swin transformer layers. Notice that in RDN and RDST variants, the global feature fusion (GFF) is considered elementary hierarchical attention, while self-attention in transformers is considered a superset of both channel and spatial attention. Very wide (dimensions ≥ 128) and deep (layers ≥ 128) networks are in bold.

	Network Architecture				Attention			Performance			Efficiency		
	Layer	**Window**	**Width**	**Depth**	**Channel**	**Spatial**	**Layer**	**PSNR↑**	**SSIM↑**	**Dice-T↑**	**MACs(G)↓**	**params(M)↓**	**FPS↑**
EDSR-lite [133]	CNN	3×3	64	32	✗	✗	✗	32.40(3.1)	0.9192(0.037)	0.8766(0.0099)	2.51	**1.52**	**325.02**
EDSR [133]	CNN	3×3	**256**	64	✗	✗	✗	32.55(3.5)	0.9184(0.039)	0.8784(0.0098)	64.22	43.08	88.00
RDN [132]	CNN	3×3	**64→256**	**142**	✗	✗	✓	32.57(3.2)	0.9241(0.036)	0.8802(0.010)	7.95	5.76	70.89
ConvNet-lite [324]	CNN	7×7	64	48	✗	✗	✗	31.92(2.9)	0.9111(0.039)	0.8669(0.010)	**1.69**	**0.88**	**213.54**
ConvNet-large [324]	CNN	7×7	**192**	96	✗	✗	✗	32.14(3.3)	0.9130(0.040)	0.8733(0.010)	21.00	12.43	98.79
RCAN [136]	CNN	3×3	64	**1610**	✓	✗	✗	32.81(3.6)	0.9224(0.038)	0.8828(0.0094)	41.34	32.03	6.38
HAN [137]	CNN	3×3	**128**	**811**	✓	✓	✓	32.33(3.7)	0.9120(0.042)	0.8751(0.0091)	83.86	64.19	17.01
SwinIR-lite [138]	STL+CNN	8×8	60	101	✓	✓	✗	32.87(3.2)	0.9252(0.037)	0.8836(0.0095)	**1.15**	**0.88**	30.12
SwinIR [138]	STL+CNN	8×8	**180**	**150**	✓	✓	✗	**33.24(3.7)**	**0.9287(0.036)**	**0.8888(0.0097)**	14.68	11.47	19.45
RDST-E($\mathcal{L}_1$)	STL+CNN	8×8	**60→150**	114	✓	✓	✗	33.06(3.3)	0.9278(0.035)	0.8869(0.0091)	3.53	2.35	28.47
RDST($\mathcal{L}_1$)	STL+CNN	8×8	**60→150**	**226**	✓	✓	✗	**33.25(3.5)**	0.9286(0.035)	0.8880(0.0097)	6.17	4.40	13.93
RDST+GFF($\mathcal{L}_1$)	STL+CNN	8×8	**60→150**	**228**	✓	✓	✓	33.23(3.5)	**0.9290(0.035)**	**0.8887(0.0095)**	6.17	4.40	13.89

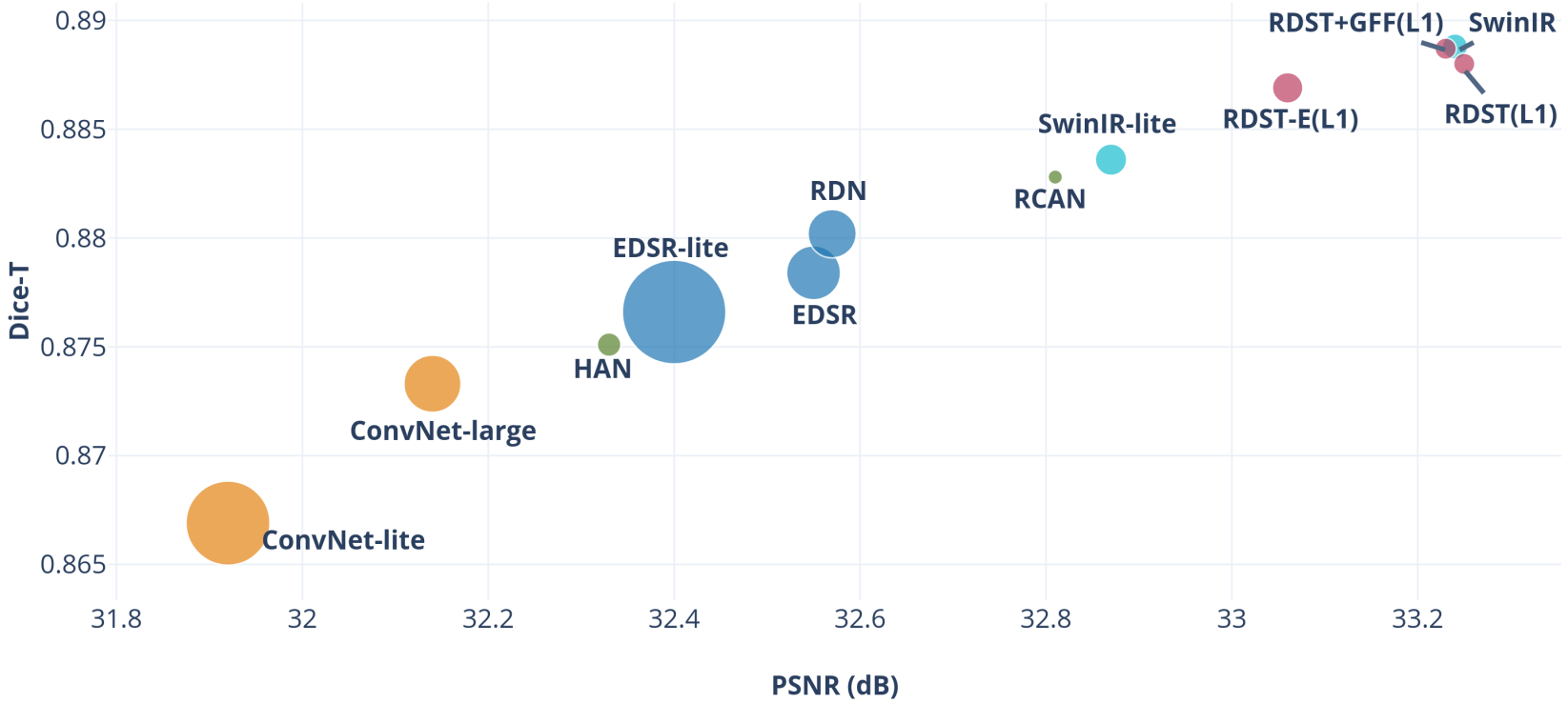


Figure 5.9: Ablation study on model architectures and attention mechanisms. All models are divided into 5 groups: native CNNs in blue, CNNs with a large receptive field in yellow, CNN-based attention models in green, SwinIR variations in light blue and the proposed RDST variations in red. PSNR scores (the x-axis) and dice coefficients of the whole brain in the downstream segmentation task (Dice-T, the y-axis) are used as evaluation metrics for SR image quality. Bigger symbols indicate higher inference throughout frame rates. For a fair comparison, all models apply the same training settings ($\mathcal{L}_1$ only for 100k steps with no extra fine-tuning).

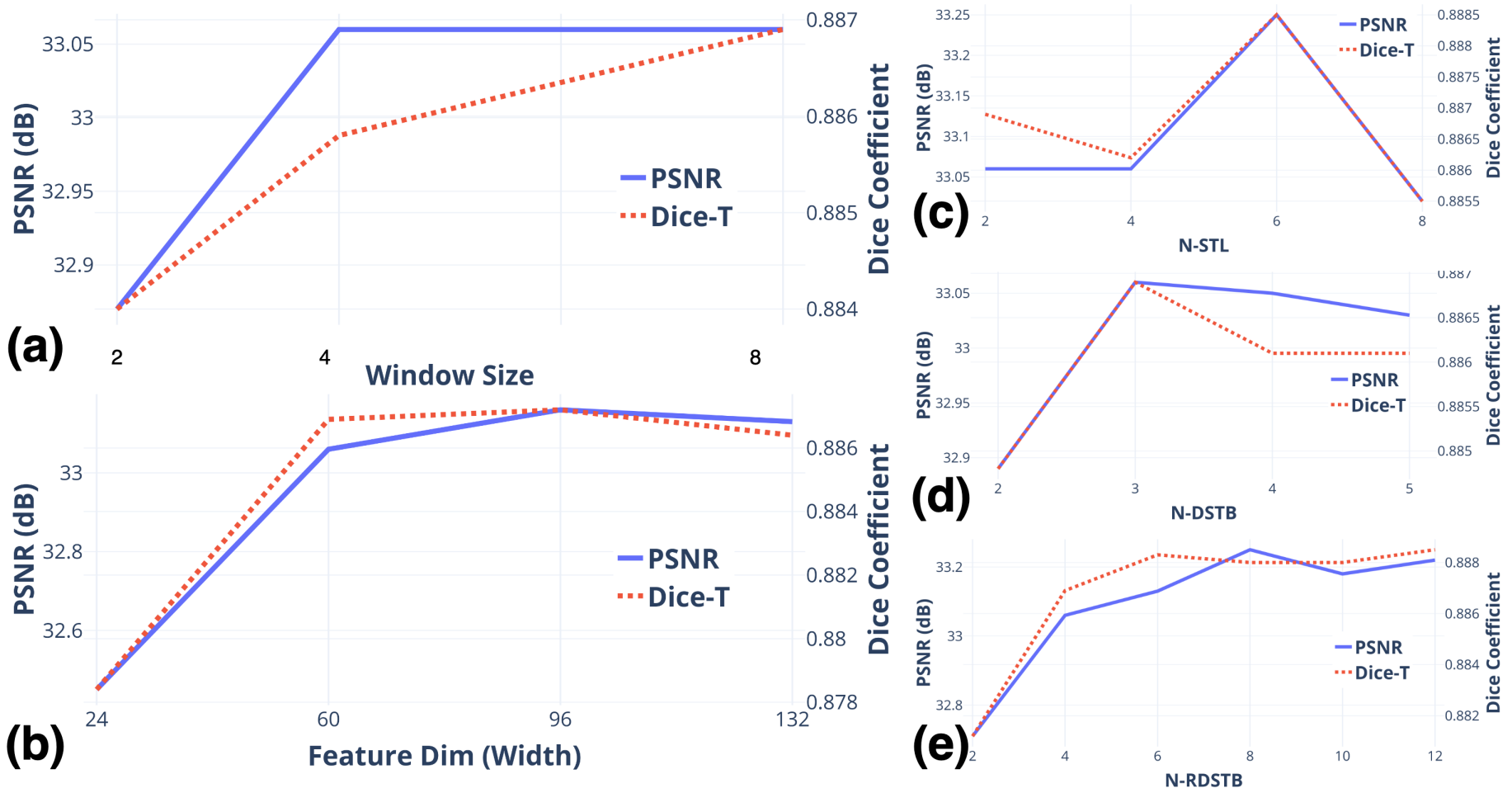


Figure 5.10: An ablation study on the hyper-parameters of RDST: (a) impacts of the window size, varying in [2, 4, 8]; (b) impacts of the feature map embedding dimension, varying from 24 to 132; (c) impacts of the number of swin transformer layers (STLs) in the dense block, varying from 2 to 8; (d) impacts of the number of DSTBs in each RDSTB, varying from 2 to 5; (e) impacts of the number of RDSTBs in RDST, varying from 2 to 12.

5.4.3 Impacts of the segmentation-based perceptual loss

In this section, I discuss the impacts of the proposed segmentation-based perceptual loss $\mathcal{L}_U$ with the following questions (Table. 5.6):

1. Is the segmentation-based perceptual loss better than existing perceptual and adversarial losses?
2. Feature maps of which layer in the pre-trained U-Net should be used?
3. What connection has been created with this perceptual loss between super-resolution and segmentation tasks?
4. Can this perceptual loss improve SR performance with other SISR methods?
5. How can this segmentation-based perceptual loss be extended to datasets without segmentation labels?

To answer these questions, an RDST-E model is trained for 100k steps with only $\mathcal{L}_1$ as the **Baseline** and then fine-tuned for extra 20k steps with different combinations of $\mathcal{L}_1$ and perceptual losses. To avoid the impacts of the additional training steps, I further train an RDST-E with $\mathcal{L}_1$ for the same steps (so-called "$\mathcal{L}_1$ **only**" in Table. 5.6).

Comparison with L1, VGG and WGANGP The shallow feature map based perceptual loss $\mathcal{L}_{E(1)}$ achieves the best SR image quality in all models. It results in significant increases of PSNR (+0.20 dB) and SSIM (+0.0013) than the Baseline (100k-steps training with only $\mathcal{L}_1$) and +0.14 dB PSNR and +0.0005 SSIM comparing with the 120k-steps $\mathcal{L}_1$ model, so the improvement is mainly caused by the proposed loss function but not the extra training steps. In contrast, neither the VGG-based perceptual loss nor the WGANGP loss combination can lead to better image quality. Actually, in the experiments, I have tested a big range ($0.0001 < \gamma < +\infty$) of the scale factor of the VGG-based perceptual loss when using it with the $\mathcal{L}_1$ loss ($\mathcal{L}_1 + \gamma\mathcal{L}_{VGG}$) and find that all the combinations decline the image quality in the medical image SR task.

Comparison of $\mathcal{L}_U$ variations As mentioned in Section. 5.2.2, variants of the segmentation-based perceptual loss are defined depending on the choice of feature maps in the U-Net. Generally, researchers agree that shallow layers represent local and basic features while deep layers represent global and semantic information in networks. I conduct experiments to compare the $\mathcal{L}_U$ variations. Narrowing the distance between the feature maps of the first encoder block (i.e. $\mathcal{L}_{E(1)}$) in the segmentation U-Net is an effective restriction of pixel-wise and structure reconstruction, leading to an increase of PSNR and SSIM scores.

Table 5.6: Ablation study of the segmentation-based perceptual loss variations. $\mathcal{L}_{E(i)}$ indicates the native L1 distance between the feature maps of the i-th encoder block, while $\sum_1^5 \mathcal{L}_{E(i)}$ indicates the sum. L1 distance between the outputs of the decoder ($\mathcal{L}_D$) and the dice loss of predicted labels ($\mathcal{L}_{HRL}$) are also tested. The best (in red) and the second best (in blue) scores are in bold.

Mean(std)	PSNR↑	SSIM↑	FID↓	Dice-T↑	Dice-G↑	Dice-W↑	Dice-CSF↑
Baseline	33.06(3.3)	0.9278(0.035)	81.63	0.8869(0.0091)	0.8438(0.057)	0.8685(0.020)	0.8877(0.014)
$\mathcal{L}_1$ only	**33.12(3.4)**	**0.9286(0.035)**	81.30	0.8874(0.0094)	0.8453(0.057)	0.8688(0.020)	0.8880(0.014)
WGANGP+VGG	33.06(3.4)	0.9259(0.037)	**72.64**	0.8885(0.0094)	0.8480(0.055)	**0.8692(0.020)**	0.8883(0.014)
VGG only	**33.12(3.4)**	0.9277(0.036)	**70.83**	0.8880(0.0092)	0.8471(0.056)	0.8689(0.020)	0.8882(0.014)
$\mathcal{L}_{E(1)}$	**33.26(3.4)**	**0.9291(0.035)**	82.13	0.8871(0.0096)	0.8446(0.057)	0.8686(0.020)	0.8877(0.015)
$\mathcal{L}_{E(2)}$	32.56(3.3)	0.9165(0.040)	73.07	0.8858(0.0089)	0.8484(0.052)	0.8650(0.020)	0.8854(0.015)
$\mathcal{L}_{E(3)}$	32.53(3.3)	0.9181(0.040)	74.83	0.8855(0.0087)	0.8483(0.051)	0.8643(0.021)	0.8843(0.014)
$\mathcal{L}_{E(4)}$	32.31(3.2)	0.9138(0.041)	82.77	0.8845(0.0087)	0.8473(0.051)	0.8630(0.021)	0.8849(0.014)
$\mathcal{L}_{E(5)}$	32.04(3.2)	0.9061(0.041)	87.01	0.8830(0.0079)	0.8473(0.049)	0.8609(0.021)	0.8821(0.013)
$\mathcal{L}_{\sum_1^5 E(i)}$	32.30(3.3)	0.9144(0.040)	77.02	0.8833(0.0085)	0.8469(0.050)	0.8612(0.021)	0.8821(0.014)
$\mathcal{L}_D$	32.28(3.3)	0.8976(0.038)	95.91	**0.8893(0.0086)**	**0.8522(0.050)**	0.8687(0.021)	**0.8897(0.012)**
$\mathcal{L}_{HRL}$	32.78(3.3)	0.9230(0.037)	78.02	**0.8899(0.0082)**	**0.8532(0.050)**	**0.8693(0.021)**	**0.8890(0.013)**

On the other hand, decreasing the dice coefficient of the predicted labels (i.e. $\mathcal{L}_{HRL}$) and the distance between the output of decoders (i.e. $\mathcal{L}_{D}$) help semantic information recovery, leading to better segmentation performance. For example, the model fine-tuned with $\mathcal{L}_{HRL}$ achieves the best dice scores of the whole brain, the grey matter and the white matter and the second best dice score of the CSF with a slight decline of PSNR and SSIM. Meanwhile, the experiments show that outputs of both ends in the U-Net are more useful for the SR task, but the feature maps of hidden layers seem useless. Perceptual losses based on the feature maps of the second to the fifth encoder block neither increase the PSNR and SSIM scores nor improve the segmentation performance.

SR for humans or machines? In addition to the distortion-perception trade-off of SR results, I find a human and machine perceptual difference in the experiments of perceptual losses. I admit that both PSNR and SSIM are essential for SR image evaluation of fidelity. However, neither can represent human perception or the performance in potential downstream image analysis tasks. Similarly, Fréchet Inception distance (FID [175]) has been used in previous works of medical image analysis tasks [104, 105] to evaluate the perceptual quality. However, its significance for medical images is also doubtful because the metric is designed for and pre-trained with natural images. In clinics, medical images are mainly for doctors to view and for machines to auto analysis. However, it is hard to explain how PSNR, SSIM and FID represent the perceptual performance in both cases. Inspired by [20], I take segmentation as a typical downstream task to discuss the difference between human perception (with FID) and machine perception (with dice coefficients) of super-resolved medical images. Based on the conclusions and discussions in previous works [27], I agree that PSNR and SSIM represent the fidelity of SR images and assume that FID denotes human perception quality. Additionally, I take the segmentation dice coefficient scores as the measurement for machine perception. In general, PSNR and SSIM closely correspond, but they are independent with either FID or dice scores. There is a trade-off between these three aspects, and none of the models can achieve superior performance in more than two directions. Thus, I suggest SR models be customised to suit the particular task. For example, the RDST variant fine-tuned with $\mathcal{L}_{E(1)}$ is proper for general purpose, and the variations fine-tuned with $\mathcal{L}_{VGG}$ and $\mathcal{L}_{WGANGP}$ are recommended for human viewing. Furthermore, I suggest using the segmentation label-based loss $\mathcal{L}_{HRL}$ for SR model fine-tuning to meet the particular needs of downstream segmentation tasks.

Table 5.7: Fine-tuning RDST variations with $\mathcal{L}_{HRL}$ and comparing with SOTA methods in the downstream segmentation tasks. The mean and standard deviations of the dice coefficients of the whole organs (e.g. brain or tumour) of SR results are compared. The best and the second best scores are highlighted in red and blue, respectively, while $*$ indicates better segmentation performance than HR ground truth images. Red and green cells indicate improved and declined scores compared to $\mathcal{L}_{E(1)}$, respectively.

Dice-T/WT↑	**HR**	**Bicubic**	**EDSR** [133]	**RCAN** [136]	**RDST-E($\mathcal{L}_{E(1)}$)**	**RDST($\mathcal{L}_{E(1)}$)**
OASIS	0.9520(0.012)	0.8125(0.0088)	0.8784(0.0098)	0.8828(0.0094)	0.8871(0.0096)	0.8889(0.0097)
BraTS	0.7833(0.13)	0.7614(0.12)	0.7800(0.12)	0.7815(0.12)	0.7831(0.12)	**0.7836*(0.12)**
ACDC	0.8932(0.027)	0.8096(0.051)	0.8599(0.030)	0.8657(0.030)	0.8691(0.025)	0.8691(0.027)
COVID-CT	0.8762(0.11)	0.8092(0.12)	0.8315(0.18)	0.8365(0.17)	0.8264(0.19)	**0.8445(0.16)**
Dice-T/WT↑		**RDN** [132]	**HAN** [137]	**SwinIR** [138]	**RDST-E($\mathcal{L}_{HRL}$)**	**RDST($\mathcal{L}_{HRL}$)**
OASIS		0.8802(0.010)	0.8751(0.0091)	0.8888(0.0097)	**0.8899(0.0082)**	**0.8906(0.0081)**
BraTS		0.7806(0.12)	0.7801(0.12)	**0.7836*(0.12)**	0.7825(0.13)	**0.7842*(0.13)**
ACDC		0.8624(0.029)	0.8666(0.030)	0.8705(0.026)	**0.8745(0.024)**	**0.8732(0.024)**
COVID-CT		0.8196(0.19)	0.8323(0.18)	0.8299(0.18)	0.8347(0.18)	**0.8411(0.16)**

SR for segmentation I fine-tune RDST and RDST-E with $\mathcal{L}_{HRL}$ and both models achieve superior segmentation performance than SOTA methods on all datasets (Table. 5.7). Compared with $\mathcal{L}_{E(1)}$ fine-tuned RDST variants, the RDST-E (with $\mathcal{L}_{HRL}$) improves the segmentation performance of the whole regions in the experiments with the OASIS, the ACDC and the COVID-CT datasets and increases the dice coefficient scores by 0.0040 on average of all four datasets, while the RDST (with $\mathcal{L}_{HRL}$) improves the segmentation performance on the OASIS, the BraTS and the ACDC datasets and increases the dice coefficient scores by 0.0008 on average of all four datasets. In addition, I choose one SOTA method with the best segmentation performance for each dataset and compare it with the $\mathcal{L}_{HRL}$ fine-tuned RDST variants in detail (Table. 5.8). Among the 15 sub-regions in total, both RDST and RDST-E (with $\mathcal{L}_{HRL}$) achieve the best scores on 7 sub-regions (with one overlap) and RCAN achieves the best segmentation performance of the right lung and lesion of the COVID-CT dataset.

Table 5.8: Dice coefficients of each tissue in the downstream segmentation tasks of SR results. RDST variations (fine-tuned with $\mathcal{L}_{HRL}$) and one SOTA method with the best performance are compared for each dataset. The highest scores are highlighted in red.

OASIS	**Dice-T**↑	**Dice-G**↑	**Dice-W**↑	**Dice-CSF**↑
SwinIR [138]	0.8888(0.0097)	0.8506(0.054)	0.8688(0.020)	0.8880(0.015)
RDST-E($\mathcal{L}_{HRL}$)	0.8899(0.0082)	0.8532(0.050)	**0.8693(0.021)**	**0.8890(0.013)**
RDST($\mathcal{L}_{HRL}$)	**0.8906(0.0081)**	**0.8567(0.049)**	**0.8693(0.021)**	0.8885(0.013)
BraTS	**Dice-WT**↑	**Dice-ET**↑	**Dice-TC**↑	
SwinIR [138]	0.7836(0.12)	0.6816(0.12)	0.6710(0.16)	
RDST-E($\mathcal{L}_{HRL}$)	0.7825(0.13)	**0.7039(0.12)**	**0.6922(0.17)**	
RDST($\mathcal{L}_{HRL}$)	**0.7842(0.13)**	0.6970(0.12)	0.6869(0.17)	
ACDC	**Dice-T**↑	**Dice-LV**↑	**Dice-RV**↑	**Dice-MC**↑
SwinIR [138]	0.8705(0.026)	0.9001(0.043)	0.6964(0.12)	0.8456(0.039)
RDST-E($\mathcal{L}_{HRL}$)	**0.8745(0.024)**	**0.9005(0.044)**	0.7068(0.12)	**0.8501(0.032)**
RDST($\mathcal{L}_{HRL}$)	0.8732(0.024)	0.8996(0.036)	**0.7197(0.11)**	0.8476(0.036)
COVID-CT	**Dice-T**↑	**Dice-LL**↑	**Dice-RL**↑	**Dice-Lesion**↑
RCAN [136]	0.8365(0.17)	0.8175(0.21)	**0.8684(0.13)**	**0.6207(0.11)**
RDST-E($\mathcal{L}_{HRL}$)	0.8347(0.18)	0.8237(0.21)	0.8632(0.14)	0.5974(0.097)
RDST($\mathcal{L}_{HRL}$)	**0.8411(0.16)**	**0.8265(0.20)**	0.8585(0.15)	0.6173(0.089)

Extending to SOTA methods The proposed segmentation-based perceptual loss variations can successfully extend to other SOTA methods (Tabel. 5.9 and Fig. 5.11). To verify the universality and usability, $\mathcal{L}_{E1}$ and $\mathcal{L}_{HRL}$ are used to fine-tune three popular SOTA SISR methods, including CNNs and vision transformers: RDN [132], RCAN [136] and SwinIR [138]. In contrast to the baselines (trained with $\mathcal{L}_1$ for 100k steps), extra training steps with only $\mathcal{L}_1$ bring limited improvements of PSNR (+0.04 dB) and segmentation performance (+0.0005 Dice-T/WT) on average. Conversely, the proposed segmentation-based perceptual loss variations significantly boost both image quality (+0.16 dB PSNR on average with $\mathcal{L}_{E(1)}$) and segmentation accuracy (+0.0028 Dice-T/WT on average with $\mathcal{L}_{HRL}$). Similar to the above conclusion in this section, models fined-tuned with $\mathcal{L}_{HRL}$ achieve the best dice scores in the downstream segmentation tasks in all cases, as expected. On the other hand, models fine-tuned with $\mathcal{L}_{E1}$ achieve the best PSNR for all cases and the second-best dice scores in most cases.

Figure 5.11: Extend the segmentation based perceptual loss variations $\mathcal{L}_{E(1)}$ and $\mathcal{L}_{HRL}$ to three popular SOTA methods: RDN [132], RCAN [136] and SwinIR [138]. For each method, three fine-tuned variations (with $\mathcal{L}_1$, $\mathcal{L}_{E(1)}$ and $\mathcal{L}_{HRL}$, respectively) are compared with the baseline (trained for 100k steps with $\mathcal{L}_1$ only).

Table 5.9: Extending the proposed segmentation-based perceptual losses to SOTA methods. For each method, the baseline model is trained with $\mathcal{L}_1$ only for 100k steps. Three fine-tuned variations are additionally trained for 20k steps with: (1) $\mathcal{L}_1$ only; (2) $\mathcal{L}_1$ with $\mathcal{L}_{E(1)}$; and (3) $\mathcal{L}_1$ with $\mathcal{L}_{HRL}$. In each group of the same method, the best and the second-best scores are in highlighted in red and blue, respectively.

Model	Training	PSNR↑	SSIM↑	Dice-T↑	Dice-G↑	Dice-W↑	Dice-CSF↑
RDN [132]	100k-$\mathcal{L}_1$	32.57(3.2)	**0.9241(0.036)**	0.8802(0.010)	0.8316(0.059)	0.8620(0.021)	**0.8845(0.015)**
	+20k-$\mathcal{L}_1$	**32.62(3.3)**	0.9217(0.037)	0.8810(0.010)	0.8343(0.058)	0.8621(0.021)	0.8840(0.015)
	+20k-$\mathcal{L}_{E(1)}$	**32.78(3.4)**	**0.9227(0.037)**	**0.8815(0.0099)**	**0.8361(0.058)**	**0.8622(0.021)**	0.8830(0.014)
	+20k-$\mathcal{L}_{HRL}$	32.30(3.1)	0.9153(0.038)	**0.8843(0.0088)**	**0.8440(0.052)**	**0.8640(0.021)**	**0.8859(0.014)**
RCAN [136]	100k-$\mathcal{L}_1$	**32.81(3.6)**	**0.9224(0.038)**	0.8828(0.0094)	0.8424(0.054)	0.8619(0.021)	0.8831(0.013)
	+20k-$\mathcal{L}_1$	**32.81(3.6)**	0.9221(0.038)	0.8827(0.0094)	0.8419(0.055)	0.8619(0.021)	0.8828(0.013)
	+20k-$\mathcal{L}_{E(1)}$	**32.94(3.7)**	**0.9231(0.038)**	**0.8833(0.0093)**	**0.8432(0.054)**	**0.8623(0.021)**	**0.8836(0.014)**
	+20k-$\mathcal{L}_{HRL}$	32.64(3.5)	0.9187(0.038)	**0.8851(0.0082)**	**0.8477(0.050)**	**0.8637(0.021)**	**0.8841(0.013)**
SwinIR [138]	100k-$\mathcal{L}_1$	33.24(3.7)	0.9287(0.036)	0.8888(0.0097)	0.8506(0.054)	0.8688(0.020)	**0.8880(0.015)**
	+20k-$\mathcal{L}_1$	**33.33(3.7)**	**0.9295(0.036)**	0.8891(0.010)	**0.8523(0.054)**	0.8688(0.021)	0.8872(0.015)
	+20k-$\mathcal{L}_{E(1)}$	**33.46(3.7)**	**0.9303(0.036)**	**0.8893(0.011)**	0.8521(0.054)	**0.8692(0.021)**	0.8874(0.015)
	+20k-$\mathcal{L}_{HRL}$	33.08(3.6)	0.9248(0.037)	**0.8908(0.0086)**	**0.8573(0.048)**	**0.8697(0.021)**	**0.8899(0.014)**
RDST	100k-$\mathcal{L}_1$	33.25(3.5)	0.9286(0.035)	0.8880(0.0097)	0.8489(0.055)	0.8679(0.021)	**0.8881(0.015)**
	+20k-$\mathcal{L}_1$	**33.29(3.6)**	**0.9293(0.035)**	**0.8890(0.0094)**	0.8512(0.054)	**0.8690(0.020)**	0.8877(0.015)
	+20k-$\mathcal{L}_{E(1)}$	**33.42(3.7)**	**0.9299(0.035)**	0.8889(0.0097)	**0.8514(0.054)**	0.8688(0.021)	0.8874(0.015)
	+20k-$\mathcal{L}_{HRL}$	33.01(3.5)	0.9243(0.037)	**0.8906(0.0081)**	**0.8567(0.049)**	**0.8693(0.021)**	**0.8885(0.013)**

To datasets without segmentation labels In the above experiments, the segmentation-based perceptual loss $\mathcal{L}_{E(1)}$ has demonstrated its robust applicability and effectiveness. In most cases, it significantly improves image fidelity quality (i.e. PSNR and SSIM) with comparable segmentation performance. Although I train a segmentation model for each dataset independently, the training is not a limitation. To use $\mathcal{L}_{E(1)}$, a U-Net can be trained on one dataset with segmentation labels and straightly extended to a new dataset without segmentation labels. In the simulation experiment, I use the pre-trained U-Net models with the OASIS, the ACDC and the COVID-CT datasets to fine-tune the RDST-E model of every dataset and achieve equal performance (Table. 5.10). Compared with the baselines (120k steps training with $\mathcal{L}_1$), they result in obvious improvement of PSNR and dice coefficient scores in almost all cases. The prior knowledge of the pre-trained segmentation model with one medical image dataset can be effectively transferred to new datasets. Proposed segmentation perceptual loss $\mathcal{L}_{E(1)}$ can become regular in a wider range of medical image low-level tasks.

Table 5.10: A transfer learning study on the segmentation-based perceptual loss in the fine-tuning stage. For each testing dataset of OASIS, ACDC and COVID, the RDST is trained for 120k steps with only $\mathcal{L}_1$ as the baselines to avoid the impacts of extra training steps. In the fine-tuning stage, pre-trained U-Net models with OASIS, ACDC and COVID datasets are used respectively to calculate $\mathcal{L}_{E(1)}$. Scores in red indicate better performance than baselines, and scores in green indicate worse performance.

Testing	**Baseline**	**Which U-Net for $\mathcal{L}_{E(1)}$**		
PSNR↑	**120k-$\mathcal{L}_1$**	**OASIS**	**ACDC**	**COVID**
OASIS	33.12(3.4)	33.26(3.4)	33.27(3.4)	33.26(3.4)
ACDC	27.23(3.0)	27.24(3.0)	27.24(3.0)	27.24(3.0)
COVID	34.58(4.4)	34.60(4.4)	34.62(4.4)	34.62(4.5)
Dice-T↑	**120k-$\mathcal{L}_1$**	**OASIS**	**ACDC**	**COVID**
OASIS	0.8874(0.0094)	0.8871(0.0096)	0.8874(0.0094)	0.8875(0.0094)
ACDC	0.8681(0.026)	0.8684(0.025)	0.8691(0.025)	0.8683(0.025)
COVID	0.8241(0.19)	0.8284(0.18)	0.8301(0.18)	0.8264(0.19)

5.4.4 Limitations and future works

Although the above experiments illustrate the superior performance and robustness of the proposed method, three limitations are worth to be noticed. First of all, all results and comparisons are achieved in simulation SR tasks. The degradation and noise formulation I used for HR-LR image pair generation in Section 5.3.1 may not represent the actual condition of various medical modalities in the clinic. Thus, it is worth exploring the capacity of the proposed method in enhancement tasks with clinical medical images in the future. Second, it is challenging to avoid over-fitting in medical image SR tasks. In Section 5.4.1, the proposed method RDST achieves the best performance (i.e. PSNR) for three datasets but achieves worse performance than the lite version RDST-E with the small dataset ACDC. I deduce that the decline is caused by over-fitting because fewer training steps of regular-size models (e.g. RDST and SwinIR) result in higher PSNR scores in the experiments of ACDC. Thus, developing data-driven early-stopping methods [326] for each case of medical image dataset is necessary and significant. Third, in the ablation study of model efficiency (Section 5.4.2), vision transformers are much slower in inference than CNN models with similar sizes. Although RDST variants have achieved the smallest model size and fewest parameters, non-attention CNN methods (i.e. EDSR and RDN) are still more than twice faster as our proposed method. Exploring more efficient vision transformers [327, 328] with the remaining SR performance will be very interesting.

Additional feature works can also be arranged in the following two directions. On the one hand, the proposed method can be a potential backbone for broader low-level medical image analysis tasks. For example, the residual dense vision transformer can be extended to MR and CT synthesis [329] tasks with shallow feature extraction and up-sampler layers modifications. Meanwhile, the proposed perceptual losses can be alternatives in MR imaging reconstruction and image registration [330, 331]. On the other hand, super-resolution tasks can be integrated into more downstream medical image analysis tasks than segmentation. Thus, the proposed method can be introduced to more medical modalities in addition to radiology scans. For example, novel perceptual losses may be designed with pre-trained models of retinal image classification [332] and improve the performance of retinal image synthesis [333].

5.5 Chapter Summary

In this chapter, I aim to improve the single-image super-resolution performance and efficiency of supervised vision transformers on medical images by introducing popular mechanisms in previous CNNs to transformers and transferring prior knowledge of segmentation tasks to SR tasks. I propose an efficient and robust single-image super-resolution method RDST for medical images by successfully introducing the residual dense connection and local feature fusion to vision transformers. The proposed RDST and its efficient version RDST-E have achieved superior or equal performance to the SOTA SISR methods of both SR image fidelity quality and downstream auto segmentation tasks. In the simulation experiments of four public medical image datasets, including MR and CT scans, the proposed RDST variants have resulted in averaged improvements of +0.09 dB and +0.06 dB PSNR receptively with only 38% and 20% parameters of SwinIR. Meanwhile, I present a perceptual loss for SR tasks based on the prior knowledge of pre-trained segmentation models and successfully extend its variants to SOTA methods, including CNNs and ViTs. The perceptual loss variant for reconstruction fidelity has led to an improvement of +0.14 dB PSNR on average, and the variant for machine perception (i.e. downstream segmentation tasks) has led to an improvement of 0.0023 dice coefficient on average. In summary, this work has introduced a framework with novel and practical designs on model architecture, loss function, training tricks and evaluation metrics. It has achieved SOTA performance in super-resolution tasks with various medical image modalities. It is also a potential backbone for more medical image low-level tasks such as reconstruction and synthesis.

CHAPTER 6

Conclusion and Future Works

This dissertation presents several deep learning-based frameworks with applications on various medical image modalities, incorporating network architectures, loss functions and training tricks for robust medical image super-resolution. In this chapter, I will summarise the work presented in this dissertation and outline possible directions for the future.

6.1 Contribution summary

This dissertation explores deep neural networks for efficient and robust medical image super-resolution. Besides incorporating SOTA algorithms designed for natural images, I implement novel network frameworks based on convolution neural networks, generative adversarial networks and vision transformers for medical image applications. More specifically, by introducing experts' opinions and in-clinic medical image analysis tasks, I conduct comprehensive comparison studies of current methods, with subjective and objective evaluations of the super-resolved images' reconstruction fidelity and perceptual quality. I present general improvements of super-resolution loss functions with experienced tricks and corresponding settings for stable and efficient training, potentially benefiting other low-level image processing tasks in the medical domain. The proposed approaches achieve state-of-the-art super-resolution results on various public and private medical image modalities with these modifications. Specifically, the main contributions include:

1. **Chapter 3** presents a multi-scale GAN for the challenging super-resolution task of rich perceptually realistic texture generation with large magnification scales (e.g. $\times 4$). I implement a progressive super-resolution framework with a lesion-focused training strategy to decline the convergence difficulty of training large-scale SR GANs and avoid the adverse effects of non-ROIs (e.g. meaningless and noisy backgrounds and unrelated organs). When this work was published, I first introduced Wasserstein distance with gradient penalty into medical single-image super-resolution for advanced

adversarial learning, leading to more efficient and stable training with no requirement of 'warm-up'. In addition to the widespread image quality assessment metrics PSNR and SSIM, a mean-opinion-score evaluation is applied to super-resolution results in a comparison study on brain and cardiac MR images. This approach has achieved comparable perceptual quality with ground truth HR images with significant improvements in PSNR, SSIM and the MOS scores than existing SISR methods when publishing.

2. In **Chapter 4**, I focus on efficient arbitrary-scale super-resolution with meta learning, adversarial learning and transfer learning on medical images. I implement a CNN-based lite version feature extraction network correlating to a scale-free upscale module that relies on weight prediction. In the simulation experiments on brain and cardiac MR images and chest CT scenes, the proposed method has achieved comparable performance on local-/global- accuracy reconstruction accuracy (i.e. PSNR and SSIM) and objective perceptual quality (i.e. FID) with much fewer parameters than SOTA SISR networks (e.g. EDSR, RDN and MetaSR). Additionally, I illustrate the impacts of various widespread residual blocks in the SR image generator and the consequences of GAN variants. I also discuss the distortion-perception trade-off of high-quality image reconstruction in a comprehensive ablation study of network and loss function components.

3. In **Chapter 5**, I present a backbone framework of vision transformers and a general-purpose perceptual loss for superior performance in medical image super-resolution tasks. I implement a residual dense transformer by introducing dense connections and local feature fusion to shifted-window attention transformers. This CNN-transformer hybrid model improves the representation capability with gradual-growing feature maps and efficient residual learning, leading to decreased trainable parameters and inference costs. In the experiments on four medical image datasets, it achieves superior performance with only 38% parameters of the SOTA methods SwinIR. Meanwhile, potential reasons behind the success of vision transformers over CNNs are discussed in ablation studies. Besides, I present a novel perceptual loss for low-level medical image processing by incorporating the prior knowledge of in-clinic segmentation. With the manual selection of corresponding variants, this perceptual loss can lead to desired improvements in reconstruction fidelity or segmentation accuracy in downstream medical image analysis tasks. The experiments declare that this segmentation-based perceptual loss significantly increases the PSNR scores of SOTA SISR networks.

6.2 Future works

In previous chapters, I have discussed the limitations of the presented works, which lead to some potential directions for the future.

Generation models Exploring and applying more advanced generation models will always be fundamental to super-resolution tasks. This dissertation has noticed how the developing generation methods (i.e. CNNs, GANs and transformers) benefit super-resolution performance. Their evolution will further benefit SR tasks shortly in the following aspects. First, novel techniques may help CNN-based backbones achieve superior performance with efficient computation, such as sizeable receptive field [324, 334] and activation-free block [335]. The second, efficient design of vision transformers [327, 328] may accelerate the inference for better clinical applicability. Third, it is worth exploring the capability of diffusion models [161–164] in super-resolution tasks for new SOTA performance.

Clinical Applications Simulation experiments can rarely estimate the authentic artefacts during medical image acquisition in the clinic. In this dissertation, I concentrate on the generated LR-HR images for a fair comparison with SOTA methods and the general applicability in super-resolution tasks. Although this dissertation involves various medical image modalities, exploring the proposed approaches with clinical needs is worth studying. Besides the radiology images (e.g. MR and CT) that I mainly focus on in this dissertation, super-resolution is also necessary for other medical images with very different signal statistics. For example, endoscopy super-resolution [336] requires real-time techniques with frame-wise recurrent learning [337]. In addition to super-resolution tasks, I expect the proposed frameworks (based on CNNs, GANs and ViTs) can be potential backbones or components for a broad range of medical image processing tasks such as reconstruction [14, 330] and translation [113]. Meanwhile, the clinic may require an end-to-end pipeline which embeds multi models of successive medical image analysis tasks. It will be a great research topic for evaluating and applying image enhancement in such a framework.

General-purpose foundation models The emerging general-purpose multimodal foundation models may bring a promising feature of medical image analysis, including medical image super-resolution. These large-scale pre-trained models have recently dominated high-level computer vision and vision-text tasks [338, 339], but low-level image synthesis tasks are still challenging [340]. In contrast, there are rare reports of foundation models for low-level image processing tasks. Specifically, the potential general-purpose cross-modality models may benefit medical image enhancement tasks for the advanced representability of medical data and the flexible applicability of new tasks with limited training data. Notably,

the multimodal medical data may provide crucial prior knowledge for high-quality image restoration with preserved structural information and generated rich details. On the other hand, it is hard to obtain coordinated image pairs with high-/low- resolution and quality, limiting the applicability of medical image analysis networks. Foundation models may solve this issue by pre-training on a large dataset and fine-tuning for specific applications. Meanwhile, the aligned cross-modality representations in the large-scale foundation models [341] may play an alternative perceptual loss for image generation tasks [342, 343]. Thus, the challenges in general-purpose foundation models for low-level medical image processing are worth discussing, such as data encoding and explainability problems [344].

6.3 Conclusion

In summary, I have presented my research on deep neural networks for efficient and robust medical image super-resolution tasks in this dissertation. These proposed approaches achieved state-of-the-art performance on a broad range of medical image datasets when published. Besides the novel network architectures, applicable training techniques and clinically significant image quality evaluation in super-resolution, the methods and findings also benefit other low-level image processing tasks. In the future, they could apply in hospitals for advanced clinical processes with proper case-specific modifications and supplementary techniques. Moreover, the discussion and ablation studies provide exciting future research directions.

Bibliography

[1] Herb Brody. Medical imaging. *Nature*, 502(7473):S81–S81, 2013.

[2] Golrokh Mirzaei, Anahita Adeli, and Hojjat Adeli. Imaging and machine learning techniques for diagnosis of alzheimer's disease. *Reviews in the Neurosciences*, 27(8):857–870, 2016.

[3] Golrokh Mirzaei and Hojjat Adeli. Resting state functional magnetic resonance imaging processing techniques in stroke studies. *Reviews in the Neurosciences*, 27(8):871–885, 2016.

[4] Matthew Leming, Juan Manuel Górriz, and John Suckling. Ensemble deep learning on large, mixed-site fmri datasets in autism and other tasks. *International journal of neural systems*, 30(7):2050012, 2020.

[5] Diego Castillo-Barnes, Francisco J Martinez-Murcia, Andres Ortiz, Diego Salas-Gonzalez, Javier RamÍrez, and Juan M Górriz. Morphological characterization of functional brain imaging by isosurface analysis in parkinson's disease. *International journal of neural systems*, 30(9):2050044–2050044, 2020.

[6] Ilker Ozsahin, Boran Sekeroglu, Musa Sani Musa, Mubarak Taiwo Mustapha, and Dilber Uzun Ozsahin. Review on diagnosis of covid-19 from chest ct images using artificial intelligence. *Computational and Mathematical Methods in Medicine*, 2020, 2020.

[7] Y Li, Bruno Sixou, and F Peyrin. A review of the deep learning methods for medical images super resolution problems. *Irbm*, 42(2):120–133, 2021.

[8] Wenming Yang, Xuechen Zhang, Yapeng Tian, Wei Wang, Jing-Hao Xue, and Qingmin Liao. Deep learning for single image super-resolution: A brief review. *IEEE Transactions on Multimedia*, 21(12):3106–3121, Dec 2019.

[9] Zhihao Wang, Jian Chen, and Steven CH Hoi. Deep learning for image super-resolution: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 2020.

[10] Dawa Chyophel Lepcha, Bhawna Goyal, Ayush Dogra, and Vishal Goyal. Image super-resolution: A comprehensive review, recent trends, challenges and applications. *Information Fusion*, 2022.

[11] Dinggang Shen, Guorong Wu, and Heung-Il Suk. Deep learning in medical image analysis. *Annual review of biomedical engineering*, 19:221, 2017.

[12] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017.

[13] Heang-Ping Chan, Ravi K Samala, Lubomir M Hadjiiski, and Chuan Zhou. Deep learning in medical image analysis. *Deep Learning in Medical Image Analysis*, pages 3–21, 2020.

[14] Hai-Miao Zhang and Bin Dong. A review on deep learning in medical image reconstruction. *Journal of the Operations Research Society of China*, 8(2):311–340, 2020.

[15] Tonghe Wang, Yang Lei, Yabo Fu, Jacob F Wynne, Walter J Curran, Tian Liu, and Xiaofeng Yang. A review on medical imaging synthesis using deep learning and its clinical applications. *Journal of applied clinical medical physics*, 22(1):11–36, 2021.

[16] Sameera V Mohd Sagheer and Sudhish N George. A review on medical image denoising algorithms. *Biomedical signal processing and control*, 61:102036, 2020.

[17] Li Sze Chow and Raveendran Paramesran. Review of medical image quality assessment. *Biomedical signal processing and control*, 27:145–154, 2016.

[18] Katarzyna Krupa and Monika Bekiesińska-Figatowska. Artifacts in magnetic resonance imaging. *Polish journal of radiology*, 80:93, 2015.

[19] Christine Cavaro-Menard, Lu Zhang, and Patrick Le Callet. Diagnostic quality assessment of medical images: Challenges and trends. In *2010 2nd European Workshop on Visual Information Processing (EUVIP)*, pages 277–284, 2010.

[20] Yan Xia, Nishant Ravikumar, John P Greenwood, Stefan Neubauer, Steffen E Petersen, and Alejandro F Frangi. Super-resolution of cardiac mr cine imaging using conditional gans and unsupervised transfer learning. *Medical Image Analysis*, 71:102037, 2021.

[21] Athanasios Voulodimos, Nikolaos Doulamis, Anastasios Doulamis, and Eftychios Protopapadakis. Deep learning for computer vision: A brief review. *Computational intelligence and neuroscience*, 2018, 2018.

[22] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. *Advances in neural information processing systems*, 30:5767–5777, 2017.

[23] Jin Zhu, Guang Yang, and Pietro Lio. Lesion focused super-resolution. In *Medical Imaging 2019: Image Processing*, volume 10949, pages 401–406. SPIE, 2019.

[24] Jin Zhu, Guang Yang, and Pietro Lio. How can we make gan perform better in single medical image super-resolution? a lesion focused multi-scale approach. *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, Apr 2019.

[25] Jin Zhu, Guang Yang, Pedro Ferreira, Andrew Scott, Sonia Nielles-Vallespin, Jennifer Keegan, Dudley Pennell, Pietro Lio, and David Firmin. A roi focused multi-scale super-resolution method for the diffusion tensor cardiac magnetic resonance. In *Proceedings of the 27th Annual Meeting (ISMRM)*, pages 11–16. International Society for Magnetic Resonance in Medicine Concord, CA, USA, 2019.

[26] Jin Zhu, Guang Yang, Tom Wong, Raad Mohiaddin, David Firmin, Jennifer Keegan, and Pietro Lio. A single-image super-resolution method for late gadolinium enhancement cmr. In *Proceedings of International Society for Magnetic Resonance in Medicine (ISMRM).*, page 2028, 2019.

[27] Jin Zhu, Chuan Tan, Junwei Yang, Guang Yang, and Pietro Lio'. Arbitrary scale super-resolution for medical images. *International Journal of Neural Systems*, 31(10):2150037, 2021.

[28] Chuan Tan, Jin Zhu, and Pietro Lio'. Arbitrary scale super-resolution for brain mri images. In *Artificial Intelligence Applications and Innovations: 16th IFIP WG 12.5 International Conference, AIAI 2020, Neos Marmaras, Greece, June 5–7, 2020, Proceedings, Part I*, pages 165–176. Springer, 2020.

[29] Jin Zhu, Guang Yang, and Pietro Lio. A residual dense vision transformer for medical image super-resolution with segmentation-based perceptual loss fine-tuning, 2023.

[30] Jin Zhu, Guang Yang, Tom Wong, Raad Mohiaddin, David Firmin, Jennifer Keegan, and Pietro Lio. Usr-net: A simple unsupervised single-image super-resolution method for late gadolinium enhancement cmr. *Proceedings of International Society for Magnetic Resonance in Medicine (ISMRM).*, 2020.

[31] Jin Zhu, Duo Wang, Zhongzhao Teng, and Pietro Lio. A multi-pathway 3d dilated convolutional neural network for brain tumor segmentation. *Proceedings of the International MICCAI BraTS Challenge*, pages 342–347, 2017.

[32] Duo Wang, Rui Zhang, Jin Zhu, Zhongzhao Teng, Yuan Huang, Filippo Spiga, Michael Hong-Fei Du, Jonathan H Gillard, Qingsheng Lu, and Pietro Liò. Neural network fusion: a novel ct-mr aortic aneurysm image segmentation method. In *Medical Imaging 2018: Image Processing*, volume 10574, pages 542–549. SPIE, 2018.

[33] Jun Lv, Jin Zhu, and Guang Yang. Which gan? a comparative study of generative adversarial network-based fast mri reconstruction. *Philosophical Transactions of the Royal Society A*, 379(2200):20200203, 2021.

[34] Guang Yang, Jun Lv, Yutong Chen, Jiahao Huang, and Jin Zhu. Generative adversarial networks (gan) powered fast magnetic resonance imaging–mini review, comparison and perspectives. *arXiv preprint arXiv:2105.01800*, 2021.

[35] KR1442 Chowdhary. Natural language processing. *Fundamentals of artificial intelligence*, pages 603–649, 2020.

[36] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[37] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[38] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. Deep learning (vol. 1, no. 2). *Cambridge: MIT Press. Retrieved December*, 21:2020, 2016.

[39] Wojciech Samek, Grégoire Montavon, Sebastian Lapuschkin, Christopher J Anders, and Klaus-Robert Müller. Explaining deep neural networks and beyond: A review of methods and applications. *Proceedings of the IEEE*, 109(3):247–278, 2021.

[40] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015.

[41] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[42] Wenjie Luo, Yujia Li, Raquel Urtasun, and Richard Zemel. Understanding the effective receptive field in deep convolutional neural networks. *Advances in neural information processing systems*, 29, 2016.

[43] Yasaman Bahri, Jonathan Kadmon, Jeffrey Pennington, Sam S Schoenholz, Jascha Sohl-Dickstein, and Surya Ganguli. Statistical mechanics of deep learning. *Annual Review of Condensed Matter Physics*, 11(1), 2020.

[44] Soufiane Hayou, Arnaud Doucet, and Judith Rousseau. On the impact of the activation function on deep neural networks training. In *International conference on machine learning*, pages 2672–2680. PMLR, 2019.

[45] Jianli Feng and Shengnan Lu. Performance analysis of various activation functions in artificial neural networks. *Journal of physics: conference series*, 1237(2):022030, 2019.

[46] Yann LeCun, Bernhard Boser, John Denker, Donnie Henderson, Richard Howard, Wayne Hubbard, and Lawrence Jackel. Handwritten digit recognition with a back-propagation network. *Advances in neural information processing systems*, 2, 1989.

[47] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Icml*, 2010.

[48] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

[49] Bing Xu, Naiyan Wang, Tianqi Chen, and Mu Li. Empirical evaluation of rectified activations in convolutional network. *arXiv preprint arXiv:1505.00853*, 2015.

[50] Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). *arXiv preprint arXiv:1511.07289*, 2015.

[51] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

[52] Ian Goodfellow, David Warde-Farley, Mehdi Mirza, Aaron Courville, and Yoshua Bengio. Maxout networks. In *International conference on machine learning*, pages 1319–1327. PMLR, 2013.

[53] Forest Agostinelli, Matthew Hoffman, Peter Sadowski, and Pierre Baldi. Learning activation functions to improve deep neural networks. *arXiv preprint arXiv:1412.6830*, 2014.

[54] Min Lin, Qiang Chen, and Shuicheng Yan. Network in network. *arXiv preprint arXiv:1312.4400*, 2013.

[55] Sebastian Ruder. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, 2016.

[56] Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010: 19th International Conference on Computational StatisticsParis France, August 22-27, 2010 Keynote, Invited and Contributed Papers*, pages 177–186. Springer, 2010.

[57] Richard S Sutton. Two problems with backpropagation and other steepest-descent learning procedures for networks. In *Proc. of Eightth Annual Conference of the Cognitive Science Society*, pages 823–831, 1986.

[58] Ning Qian. On the momentum term in gradient descent learning algorithms. *Neural networks*, 12(1):145–151, 1999.

[59] Yurii Evgen'evich Nesterov. A method of solving a convex programming problem with convergence rate $o\left(\frac{1}{k^2}\right)$. *Doklady Akademii Nauk*, 269(3):543–547, 1983.

[60] Matthew D Zeiler. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012.

[61] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[62] Timothy Dozat. Incorporating nesterov momentum into adam, 2016.

[63] Xin He, Kaiyong Zhao, and Xiaowen Chu. Automl: A survey of the state-of-the-art. *Knowledge-Based Systems*, 212:106622, 2021.

[64] Radwa Elshawi, Mohamed Maher, and Sherif Sakr. Automated machine learning: State-of-the-art and open challenges. *arXiv preprint arXiv:1906.02287*, 2019.

[65] Roger Grosse. Lecture 15: Exploding and vanishing gradients. *University of Toronto Computer Science*, 2017.

[66] Meenal V Narkhede, Prashant P Bartakke, and Mukul S Sutaone. A review on weight initialization strategies for neural networks. *Artificial intelligence review*, 55(1):291–322, 2022.

[67] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, 2020.

[68] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022.

[69] Chaoning Zhang, Chenshuang Zhang, Junha Song, John Seon Keun Yi, Kang Zhang, and In So Kweon. A survey on masked autoencoder for self-supervised learning in vision and beyond. *arXiv preprint arXiv:2208.00173*, 2022.

[70] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, 2015.

[71] Lei Huang, Jie Qin, Yi Zhou, Fan Zhu, Li Liu, and Ling Shao. Normalization techniques in training dnns: Methodology, analysis and application. *arXiv preprint arXiv:2009.12836*, 2020.

[72] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[73] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[74] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016.

[75] Tim Salimans and Durk P Kingma. Weight normalization: A simple reparameterization to accelerate training of deep neural networks. *Advances in neural information processing systems*, 29, 2016.

[76] Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

[77] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pages 1501–1510, 2017.

[78] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2337–2346, 2019.

[79] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019.

[80] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020.

[81] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems*, 34:852–863, 2021.

[82] Weihao Xia, Yulun Zhang, Yujiu Yang, Jing-Hao Xue, Bolei Zhou, and Ming-Hsuan Yang. Gan inversion: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[83] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

[84] Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48, 2019.

[85] Phillip Chlap, Hang Min, Nym Vandenberg, Jason Dowling, Lois Holloway, and Annette Haworth. A review of medical image data augmentation techniques for deep learning applications. *Journal of Medical Imaging and Radiation Oncology*, 65(5):545–563, 2021.

[86] Xiangtao Kong, Xina Liu, Jinjin Gu, Yu Qiao, and Chao Dong. Reflash dropout in image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6002–6012, 2022.

[87] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2014.

[88] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

[89] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

[90] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[91] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[92] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015.

[93] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4*, pages 3–11. Springer, 2018.

[94] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018.

[95] Huimin Huang, Lanfen Lin, Ruofeng Tong, Hongjie Hu, Qiaowei Zhang, Yutaro Iwamoto, Xianhua Han, Yen-Wei Chen, and Jian Wu. Unet 3+: A full-scale connected unet for medical image segmentation. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1055–1059. IEEE, 2020.

[96] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. *arXiv preprint arXiv:2105.05537*, 2021.

[97] Junyoung Park, Donghwi Hwang, Kyeong Yun Kim, Seung Kwan Kang, Yu Kyeong Kim, and Jae Sung Lee. Computed tomography super-resolution using deep convolutional neural network. *Physics in Medicine & Biology*, 63(14):145011, 2018.

[98] Chi-Mao Fan, Tsung-Jung Liu, and Kuan-Hsien Liu. Sunet: Swin transformer unet for image denoising. *arXiv preprint arXiv:2202.14009*, 2022.

[99] Pengju Liu, Hongzhi Zhang, Kai Zhang, Liang Lin, and Wangmeng Zuo. Multi-level wavelet-cnn for image restoration. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 773–782, 2018.

[100] Nahian Siddique, Sidike Paheding, Colin P Elkin, and Vijay Devabhaktuni. U-net and its variants for medical image segmentation: A review of theory and applications. *Ieee Access*, 9:82031–82057, 2021.

[101] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014.

[102] Antonia Creswell, Tom White, Vincent Dumoulin, Kai Arulkumaran, Biswa Sengupta, and Anil A Bharath. Generative adversarial networks: An overview. *IEEE signal processing magazine*, 35(1):53–65, 2018.

[103] Jie Gui, Zhenan Sun, Yonggang Wen, Dacheng Tao, and Jieping Ye. A review on generative adversarial networks: Algorithms, theory, and applications. *IEEE transactions on knowledge and data engineering*, 2021.

[104] Xin Yi, Ekta Walia, and Paul Babyn. Generative adversarial network in medical imaging: A review. *Medical image analysis*, 58:101552, 2019.

[105] Salome Kazeminia, Christoph Baur, Arjan Kuijper, Bram van Ginneken, Nassir Navab, Shadi Albarqouni, and Anirban Mukhopadhyay. Gans for medical image analysis. *Artificial Intelligence in Medicine*, 109:101938, 2020.

[106] Alexia Jolicoeur-Martineau. The relativistic discriminator: a key element missing from standard gan. *arXiv preprint arXiv:1807.00734*, 2018.

[107] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *Advances in neural information processing systems*, 29, 2016.

[108] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan, 2017.

[109] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016.

[110] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[111] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 5907–5915, 2017.

[112] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[113] Hao Tang, Dan Xu, Nicu Sebe, Yanzhi Wang, Jason J Corso, and Yan Yan. Multi-channel attention selection gan with cascaded semantic guidance for cross-view image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2417–2426, 2019.

[114] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network, 2017.

[115] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *International conference on machine learning*, pages 1060–1069. PMLR, 2016.

[116] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[117] Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, et al. A survey on visual transformer. *arXiv preprint arXiv:2012.12556*, 2(4), 2020.

[118] Emerald U Henry, Onyeka Emebob, and Conrad Asotie Omonhinmin. Vision transformers in medical imaging: A review. *arXiv preprint arXiv:2211.10043*, 2022.

[119] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021.

[120] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[121] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 213–229. Springer, 2020.

[122] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020.

[123] Minghang Zheng, Peng Gao, Renrui Zhang, Kunchang Li, Xiaogang Wang, Hongsheng Li, and Hao Dong. End-to-end object detection with adaptive clustering transformer. *arXiv preprint arXiv:2011.09315*, 2020.

[124] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021.

[125] Yifan Jiang, Shiyu Chang, and Zhangyang Wang. Transgan: Two pure transformers can make one strong gan, and that can scale up. *Advances in Neural Information Processing Systems*, 34:14745–14758, 2021.

[126] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021.

[127] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12299–12310, 2021.

[128] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[129] Linwei Yue, Huanfeng Shen, Jie Li, Qiangqiang Yuan, Hongyan Zhang, and Liangpei Zhang. Image super-resolution: The techniques, applications, and future. *Signal processing*, 128:389–408, 2016.

[130] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):295–307, 2015.

[131] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1646–1654. IEEE, 2016.

[132] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2472–2481. IEEE, 2018.

[133] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution, 2017.

[134] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0. Springer Science+Business Media, 2018.

[135] Xuecai Hu, Haoyuan Mu, Xiangyu Zhang, Zilei Wang, Tieniu Tan, and Jian Sun. Meta-sr: A magnification-arbitrary network for super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1575–1584. IEEE, 2019.

[136] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks, 2018.

[137] Ben Niu, Weilei Wen, Wenqi Ren, Xiangde Zhang, Lianping Yang, Shuzhen Wang, Kaihao Zhang, Xiaochun Cao, and Haifeng Shen. Single image super-resolution via a holistic attention network. In *European conference on computer vision*, pages 191–207. Springer, 2020.

[138] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1833–1844, 2021.

[139] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution, 2016.

[140] Kai Zhang, Shuhang Gu, and Radu Timofte. Ntire 2020 challenge on perceptual extreme super-resolution: Methods and results. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 492–493, 2020.

[141] Andreas Lugmayr, Martin Danelljan, and Radu Timofte. Ntire 2020 challenge on real-world image super-resolution: Methods and results. In *Proceedings of the*

IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pages 494–495, 2020.

[142] Qiong Wu, Chunxiao Fan, Yong Li, Yang Li, and Jiahao Hu. A novel perceptual loss function for single image super-resolution. *Multimedia Tools and Applications*, 79:21265–21278, 2020.

[143] Shane D Sims. Frequency domain-based perceptual loss for super resolution. In *2020 IEEE 30th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE, 2020.

[144] Yan Zhang, Weihong Li, Weiguo Gong, Zixu Wang, and Jingxi Sun. An improved boundary-aware perceptual loss for building extraction from vhr images. *Remote Sensing*, 12(7):1195, 2020.

[145] Sean Bell, Paul Upchurch, Noah Snavely, and Kavita Bala. Material recognition in the wild with the materials in context database, 2015.

[146] Mohammad Saeed Rad, Behzad Bozorgtabar, Urs-Viktor Marti, Max Basler, Hazim Kemal Ekenel, and Jean-Philippe Thiran. Srobb: Targeted perceptual loss for single image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2710–2719, 2019.

[147] Kui Fu, Jiansheng Peng, Hanxiao Zhang, Xiaoliang Wang, and Frank Jiang. Image super-resolution based on generative adversarial networks: a brief review. *Computers, Materials & Continua*, 64(3):1977–1997, 2020.

[148] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks, 2020.

[149] Leonid I Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: nonlinear phenomena*, 60(1-4):259–268, 1992.

[150] Sébastien Tourbier, Xavier Bresson, Patric Hagmann, Jean-Philippe Thiran, Reto Meuli, and Meritxell Bach Cuadra. An efficient total variation algorithm for super-resolution in fetal brain mri with adaptive regularization. *NeuroImage*, 118:584–597, 2015.

[151] Leon Gatys, Alexander S Ecker, and Matthias Bethge. Texture synthesis using convolutional neural networks. *Advances in neural information processing systems*, 28, 2015.

[152] Mehdi SM Sajjadi, Bernhard Scholkopf, and Michael Hirsch. Enhancenet: Single image super-resolution through automated texture synthesis. In *Proceedings of the IEEE international conference on computer vision*, pages 4491–4500, 2017.

[153] Yassir Saquil, Kwang In Kim, and Peter Hall. Ranking cgans: Subjective control over semantic image attributes. *arXiv preprint arXiv:1804.04082*, 2018.

[154] Wenlong Zhang, Yihao Liu, Chao Dong, and Yu Qiao. Ranksrgan: Generative adversarial networks with ranker for image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3096–3105, 2019.

[155] Hang Zhao, Orazio Gallo, Iuri Frosio, and Jan Kautz. Loss functions for image restoration with neural networks. *IEEE Transactions on computational imaging*, 3(1):47–57, 2016.

[156] Damon M Chandler. Seven challenges in image quality assessment: past, present, and future research. *International Scholarly Research Notices*, 2013, 2013.

[157] Guangtao Zhai and Xiongkuo Min. Perceptual image quality assessment: a survey. *Science China Information Sciences*, 63(11):1–52, 2020.

[158] Zhou Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, April 2004.

[159] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, volume 2, pages 1398–1402. Ieee, 2003.

[160] Chaofeng Li and Alan C Bovik. Content-partitioned structural similarity index for image quality assessment. *Signal Processing: Image Communication*, 25(7):517–526, 2010.

[161] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *J. Mach. Learn. Res.*, 23(47):1–33, 2022.

[162] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

[163] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.

[164] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[165] Guang Yang, Xiahai Zhuang, Habib Khan, Shouvik Haldar, Eva Nyktari, Lei Li, Ricardo Wage, Xujiong Ye, Greg Slabaugh, Raad Mohiaddin, et al. Fully automatic segmentation and objective assessment of atrial scars for long-standing persistent atrial fibrillation patients using late gadolinium-enhanced mri. *Medical physics*, 45(4):1562–1576, 2018.

[166] Maximilian Seitzer, Guang Yang, Jo Schlemper, Ozan Oktay, Tobias Würfl, Vincent Christlein, Tom Wong, Raad Mohiaddin, David Firmin, Jennifer Keegan, et al. Adversarial and perceptual refinement for compressed sensing mri reconstruction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 232–240. Springer, 2018.

[167] Jongyoo Kim and Sanghoon Lee. Deep learning of human visual sensitivity in image quality assessment framework. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1676–1684, 2017.

[168] Kede Ma, Wentao Liu, Tongliang Liu, Zhou Wang, and Dacheng Tao. dipiq: Blind image quality assessment by learning-to-rank discriminable image pairs. *IEEE Transactions on Image Processing*, 26(8):3951–3964, 2017.

[169] Kede Ma, Wentao Liu, Kai Zhang, Zhengfang Duanmu, Zhou Wang, and Wangmeng Zuo. End-to-end blind image quality assessment using deep neural networks. *IEEE Transactions on Image Processing*, 27(3):1202–1213, 2017.

[170] Hossein Talebi and Peyman Milanfar. Nima: Neural image assessment. *IEEE transactions on image processing*, 27(8):3998–4011, 2018.

[171] Xialei Liu, Joost Van De Weijer, and Andrew D Bagdanov. Rankiqa: Learning from rankings for no-reference image quality assessment. In *Proceedings of the IEEE international conference on computer vision*, pages 1040–1049, 2017.

[172] Shane Barratt and Rishi Sharma. A note on the inception score. *arXiv preprint arXiv:1801.01973*, 2018.

[173] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision, 2015.

[174] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

[175] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium, 2018.

[176] Hamid R Sheikh, Alan C Bovik, and Gustavo De Veciana. An information fidelity criterion for image quality assessment using natural scene statistics. *IEEE Transactions on image processing*, 14(12):2117–2128, 2005.

[177] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012.

[178] N Venkatanath, D Praneeth, Maruthi Chandrasekhar Bh, Sumohana S Channappayya, and Swarup S Medasani. Blind image quality evaluation using perception based features. In *2015 twenty first national conference on communications (NCC)*, pages 1–6. IEEE, 2015.

[179] Chao Ma, Chih-Yuan Yang, Xiaokang Yang, and Ming-Hsuan Yang. Learning a no-reference quality metric for single-image super-resolution. *Computer Vision and Image Understanding*, 158:1–16, 2017.

[180] Yulun Zhang, Kai Zhang, Zheng Chen, Yawei Li, Radu Timofte, Junpei Zhang, Kexin Zhang, Rui Peng, Yanbiao Ma, Licheng Jia, et al. Ntire 2023 challenge on image super-resolution (x4): Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1864–1883, 2023.

[181] Yawei Li, Yulun Zhang, Radu Timofte, Luc Van Gool, Lei Yu, Youwei Li, Xinpeng Li, Ting Jiang, Qi Wu, Mingyan Han, et al. Ntire 2023 challenge on efficient super-resolution: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1921–1959, 2023.

[182] Matthew D Zeiler, Dilip Krishnan, Graham W Taylor, and Rob Fergus. Deconvolutional networks. In *2010 IEEE Computer Society Conference on computer vision and pattern recognition*, pages 2528–2535. IEEE, 2010.

[183] Jin Yamanaka, Shigesumi Kuwashima, and Takio Kurita. Fast and accurate image super resolution by deep cnn with skip connection and network in network. In *Neural Information Processing: 24th International Conference, ICONIP 2017, Guangzhou,*

China, November 14-18, 2017, Proceedings, Part II 24, pages 217–225. Springer, 2017.

[184] Chenyu You, Guang Li, Yi Zhang, Xiaoliu Zhang, Hongming Shan, Mengzhou Li, Shenghong Ju, Zhen Zhao, Zhuiyang Zhang, Wenxiang Cong, et al. Ct super-resolution gan constrained by the identical, residual, and cycle learning ensemble (gan-circle). *IEEE transactions on medical imaging*, 39(1):188–203, 2019.

[185] Feilong Cao, Kaixuan Yao, and Jiye Liang. Deconvolutional neural network for image super-resolution. *Neural Networks*, 132:394–404, 2020.

[186] Md Rifat Arefin, Vincent Michalski, Pierre-Luc St-Charles, Alfredo Kalaitzis, Sookyung Kim, Samira E Kahou, and Yoshua Bengio. Multi-image super-resolution for remote sensing using deep recurrent networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 206–207, 2020.

[187] Shuhang Gu, Wangmeng Zuo, Qi Xie, Deyu Meng, Xiangchu Feng, and Lei Zhang. Convolutional sparse coding for image super-resolution. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1823–1831, 2015.

[188] Augustus Odena, Vincent Dumoulin, and Chris Olah. Deconvolution and checkerboard artifacts. *Distill*, 1(10):e3, 2016.

[189] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P. Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network, 2016.

[190] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5728–5739, 2022.

[191] Fuzhi Yang, Huan Yang, Jianlong Fu, Hongtao Lu, and Baining Guo. Learning texture transformer network for image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5791–5800, 2020.

[192] Hongyang Gao, Hao Yuan, Zhengyang Wang, and Shuiwang Ji. Pixel transposed convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 42(5):1218–1227, 2019.

[193] Andrew Aitken, Christian Ledig, Lucas Theis, Jose Caballero, Zehan Wang, and Wenzhe Shi. Checkerboard artifact free sub-pixel convolution: A note on sub-pixel convolution, resize convolution and convolution resize. *arXiv preprint arXiv:1707.02937*, 2017.

[194] Christiane Lemke, Marcin Budka, and Bogdan Gabrys. Metalearning: a survey of trends and technologies. *Artificial intelligence review*, 44(1):117–130, 2015.

[195] Timothy Hospedales, Antreas Antoniou, Paul Micaelli, and Amos Storkey. Meta-learning in neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(9):5149–5169, 2021.

[196] Juncheng Li, Faming Fang, Jiaqian Li, Kangfu Mei, and Guixu Zhang. Mdcn: Multi-scale dense cross network for image super-resolution. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(7):2547–2561, 2020.

[197] Zhisheng Lu, Juncheng Li, Hong Liu, Chaoyan Huang, Linlin Zhang, and Tieyong Zeng. Transformer for single image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 457–466, 2022.

[198] Haoying Li, Yifan Yang, Meng Chang, Shiqi Chen, Huajun Feng, Zhihai Xu, Qi Li, and Yueting Chen. Srdiff: Single image super-resolution with diffusion probabilistic models. *Neurocomputing*, 479:47–59, 2022.

[199] Guangwei Gao, Zhengxue Wang, Juncheng Li, Wenjie Li, Yi Yu, and Tieyong Zeng. Lightweight bimodal network for single-image super-resolution via symmetric cnn and recursive transformer. *arXiv preprint arXiv:2204.13286*, 2022.

[200] Ying Tai, Jian Yang, Xiaoming Liu, and Chunyan Xu. Memnet: A persistent memory network for image restoration. In *Proceedings of the IEEE international conference on computer vision*, pages 4539–4547, 2017.

[201] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Deeply-recursive convolutional network for image super-resolution. *CoRR*, abs/1511.04491, 2015.

[202] Ying Tai, Jian Yang, and Xiaoming Liu. Image super-resolution via deep recursive residual network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3147–3155, 2017.

[203] Juncheng Li, Faming Fang, Kangfu Mei, and Guixu Zhang. Multi-scale residual network for image super-resolution. In *Proceedings of the European conference on computer vision (ECCV)*, pages 517–532, 2018.

[204] Zheng Hui, Xiumei Wang, and Xinbo Gao. Fast and accurate single image super-resolution via information distillation network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 723–731, 2018.

[205] Wei Han, Shiyu Chang, Ding Liu, Mo Yu, Michael Witbrock, and Thomas S Huang. Image super-resolution via dual-state recurrent networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1654–1663, 2018.

[206] Haoyu Ren, Mostafa El-Khamy, and Jungwon Lee. Image super resolution based on fusing multiple convolution neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 54–61, 2017.

[207] Armin Mehri, Parichehr B Ardakani, and Angel D Sappa. Mprnet: Multi-path residual network for lightweight image super resolution. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2704–2713, 2021.

[208] Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang. Second-order attention network for single image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11065–11074, 2019.

[209] Defu Qiu, Yuhu Cheng, and Xuesong Wang. Dual u-net residual networks for cardiac magnetic resonance images super-resolution. *Computer Methods and Programs in Biomedicine*, 218:106707, 2022.

[210] Assaf Shocher, Nadav Cohen, and Michal Irani. "zero-shot" super-resolution using deep internal learning, 2017.

[211] Kai Zhang, Luc Van Gool, and Radu Timofte. Deep unfolding network for image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3217–3226, 2020.

[212] Muhammad Haris, Greg Shakhnarovich, and Norimichi Ukita. Deep back-projection networks for super-resolution, 2018.

[213] Zhen Li, Jinglei Yang, Zheng Liu, Xiaomin Yang, Gwanggil Jeon, and Wei Wu. Feedback network for image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3867–3876, 2019.

[214] Muhammad Haris, Gregory Shakhnarovich, and Norimichi Ukita. Recurrent back-projection network for video super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3897–3906, 2019.

[215] Chao Dong, Chen Change Loy, and Xiaoou Tang. Accelerating the super-resolution convolutional neural network. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 391–407. Springer, 2016.

[216] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 624–632, 2017.

[217] Yifan Wang, Federico Perazzi, Brian McWilliams, Alexander Sorkine-Hornung, Olga Sorkine-Hornung, and Christopher Schroers. A fully progressive approach to single-image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 864–873, 2018.

[218] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Fast and accurate image super-resolution with deep laplacian pyramid networks. *IEEE transactions on pattern analysis and machine intelligence*, 41(11):2599–2613, 2018.

[219] Hang Zhao, Orazio Gallo, Iuri Frosio, and Jan Kautz. Loss functions for neural networks for image processing. *arXiv preprint arXiv:1511.08861*, 2015.

[220] Tong Tong, Gen Li, Xiejie Liu, and Qinquan Gao. Image super-resolution using dense skip connections. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4799–4807. IEEE, 2017.

[221] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. *Proceedings of the AAAI conference on artificial intelligence*, 31(1), 2017.

[222] Meng-Hao Guo, Tian-Xing Xu, Jiang-Jiang Liu, Zheng-Ning Liu, Peng-Tao Jiang, Tai-Jiang Mu, Song-Hai Zhang, Ralph R Martin, Ming-Ming Cheng, and Shi-Min Hu. Attention mechanisms in computer vision: A survey. *Computational Visual Media*, 8(3):331–368, 2022.

[223] Hongyu Zhu, Chao Xie, Yeqi Fei, and Huanjie Tao. Attention mechanisms in cnn-based single image super-resolution: A brief review and a new perspective. *Electronics*, 10(10):1187, 2021.

[224] Saeed Anwar and Nick Barnes. Densely residual laplacian super-resolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3):1192–1204, 2020.

[225] Yiqun Mei, Yuchen Fan, Yuqian Zhou, Lichao Huang, Thomas S Huang, and Honghui Shi. Image super-resolution with cross-scale non-local attention and exhaustive self-exemplars mining. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5690–5699, 2020.

[226] Yiqun Mei, Yuchen Fan, and Yuqian Zhou. Image super-resolution with non-local sparse attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3517–3526, 2021.

[227] Jae-Seok Choi and Munchurl Kim. A deep convolutional neural network with selection units for super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 154–160, 2017.

[228] Jie Liu, Wenjie Zhang, Yuting Tang, Jie Tang, and Gangshan Wu. Residual feature aggregation network for image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2359–2368, 2020.

[229] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17683–17693, 2022.

[230] Wenbin Zou, Tian Ye, Weixin Zheng, Yunchen Zhang, Liang Chen, and Yi Wu. Self-calibrated efficient transformer for lightweight super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 930–939, 2022.

[231] Jinsheng Fang, Hanjiang Lin, Xinyu Chen, and Kun Zeng. A hybrid network of cnn and transformer for lightweight image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1103–1112, 2022.

[232] Garas Gendy, Guanghui He, and Nabil Sabor. Lightweight image super-resolution based on deep learning: State-of-the-art and future directions. *Information Fusion*, 94:284–310, 2023.

[233] Junjun Jiang, Chenyang Wang, Xianming Liu, and Jiayi Ma. Deep learning-based face super-resolution: A survey. *ACM Computing Surveys (CSUR)*, 55(1):1–36, 2021.

[234] Y Li, Bruno Sixou, and F Peyrin. A review of the deep learning methods for medical images super resolution problems. *IRBM*, 2020.

[235] R Mark Henkelman. Erratum: Measurement of signal intensities in the presence of noise in mr images [med. phys. 12, 232 (1985)]. *Medical physics*, 13(4):544–544, 1986.

[236] Hákon Gudbjartsson and Samuel Patz. The rician distribution of noisy mri data. *Magnetic resonance in medicine*, 34(6):910–914, 1995.

[237] Qiaoqiao Ding, Yong Long, Xiaoqun Zhang, and Jeffrey A Fessler. Statistical image reconstruction using mixed poisson-gaussian noise model for x-ray ct. *arXiv preprint arXiv:1801.09533*, 2018.

[238] Pierre Gravel, Gilles Beaudoin, and Jacques A De Guise. A method for modeling noise in medical images. *IEEE Transactions on medical imaging*, 23(10):1221–1232, 2004.

[239] Thanh-Trung Nguyen, Dinh-Hoan Trinh, and Nguyen Linh-Trung. An efficient example-based method for ct image denoising based on frequency decomposition and sparse representation. In *2016 International Conference on Advanced Technologies for Communications (ATC)*, pages 293–296. IEEE, 2016.

[240] Jithin Saji Isaac and Ramesh Kulkarni. Super resolution techniques for medical image processing. In *2015 International Conference on Technologies for Sustainable Development (ICTSD)*, pages 1–6. IEEE, 2015.

[241] Sharon Peled and Yehezkel Yeshurun. Superresolution in mri: application to human white matter fiber tract visualization by diffusion tensor imaging. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 45(1):29–35, 2001.

[242] Hayit Greenspan, G Oz, N Kiryati, and SLBG Peled. Mri inter-slice reconstruction using super-resolution. *Magnetic resonance imaging*, 20(5):437–446, 2002.

[243] Richard Z Shilling, Trevor Q Robbie, Timothée Bailloeul, Klaus Mewes, Russell M Mersereau, and Marijn E Brummer. A super-resolution framework for 3-d high-resolution and high-contrast imaging using 2-d multislice mri. *IEEE transactions on medical imaging*, 28(5):633–644, 2008.

[244] François Rousseau. Brain hallucination. In *Computer Vision–ECCV 2008: 10th European Conference on Computer Vision, Marseille, France, October 12-18, 2008, Proceedings, Part I 10*, pages 497–508. Springer, 2008.

[245] José V Manjón, Pierrick Coupé, Antonio Buades, D Louis Collins, and Montserrat Robles. Mri superresolution using self-similarity and image priors. *Journal of Biomedical Imaging*, 2010:1–11, 2010.

[246] Chi-Hieu Pham, Aurélien Ducournau, Ronan Fablet, and François Rousseau. Brain mri super-resolution using deep 3d convolutional networks. In *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, pages 197–200. IEEE, 2017.

[247] Ozan Oktay, Wenjia Bai, Matthew Lee, Ricardo Guerrero, Konstantinos Kamnitsas, Jose Caballero, Antonio de Marvao, Stuart Cook, Declan O'Regan, and Daniel Rueckert. Multi-input cardiac image super-resolution using convolutional neural networks. In *Medical Image Computing and Computer-Assisted Intervention-MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part III 19*, pages 246–254. Springer, 2016.

[248] Steven McDonagh, Benjamin Hou, Amir Alansary, Ozan Oktay, Konstantinos Kamnitsas, Mary Rutherford, Jo V Hajnal, and Bernhard Kainz. Context-sensitive super-resolution for fast fetal magnetic resonance imaging. In *Molecular Imaging, Reconstruction and Analysis of Moving Body Organs, and Stroke Imaging and Treatment: Fifth International Workshop, CMMI 2017, Second International Workshop, RAMBO 2017, and First International Workshop, SWITCH 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 14, 2017, Proceedings 5*, pages 116–126. Springer, 2017.

[249] Jun Shi, Zheng Li, Shihui Ying, Chaofeng Wang, Qingping Liu, Qi Zhang, and Pingkun Yan. Mr image super-resolution via wide residual networks with fixed skip connection. *IEEE journal of biomedical and health informatics*, 23(3):1129–1140, 2018.

[250] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

[251] Jinglong Du, Zhongshi He, Lulu Wang, Ali Gholipour, Zexun Zhou, Dingding Chen, and Yuanyuan Jia. Super-resolution reconstruction of single anisotropic 3d mr images using residual convolutional neural network. *Neurocomputing*, 392:209–220, 2020.

[252] José V Manjón, Pierrick Coupé, Antonio Buades, Vladimir Fonov, D Louis Collins, and Montserrat Robles. Non-local mri upsampling. *Medical image analysis*, 14(6):784–792, 2010.

[253] Yuhua Chen, Yibin Xie, Zhengwei Zhou, Feng Shi, Anthony G Christodoulou, and Debiao Li. Brain mri super resolution using 3d deep densely connected neural networks. In *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, pages 739–742. IEEE, 2018.

[254] Yinhao Li, Yutaro Iwamoto, Lanfen Lin, Rui Xu, Ruofeng Tong, and Yen-Wei Chen. Volumenet: a lightweight parallel network for super-resolution of mr and ct volumetric data. *IEEE Transactions on Image Processing*, 30:4840–4854, 2021.

[255] Xiaole Zhao, Yulun Zhang, Tao Zhang, and Xueming Zou. Channel splitting network for single mr image super-resolution. *IEEE transactions on image processing*, 28(11):5649–5662, 2019.

[256] Liming Zhao, Jingdong Wang, Xi Li, Zhuowen Tu, and Wenjun Zeng. Deep convolutional neural networks with merge-and-run mappings. *arXiv preprint arXiv:1611.07718*, 2016.

[257] Defu Qiu, Yuhu Cheng, and Xuesong Wang. Progressive u-net residual network for computed tomography images super-resolution in the screening of covid-19. *Journal of Radiation Research and Applied Sciences*, 14(1):369–379, 2021.

[258] Ozan Oktay, Enzo Ferrante, Konstantinos Kamnitsas, Mattias Heinrich, Wenjia Bai, Jose Caballero, Stuart A Cook, Antonio De Marvao, Timothy Dawes, Declan P O'Regan, et al. Anatomically constrained neural networks (acnns): application to cardiac image enhancement and segmentation. *IEEE transactions on medical imaging*, 37(2):384–395, 2017.

[259] Mayur Bhandary, J Patricio Reyes, Eylul Ertay, and Aman Panda. Double u-net for super-resolution and segmentation of live cell images. *arXiv preprint arXiv:2212.02028*, 2022.

[260] Yuhua Chen, Feng Shi, Anthony G. Christodoulou, Zhengwei Zhou, Yibin Xie, and Debiao Li. Efficient and accurate mri super-resolution using a generative adversarial network and 3d multi-level densely connected network, 2018.

[261] Irina Sánchez and Verónica Vilaplana. Brain mri super-resolution using 3d generative adversarial networks. *arXiv preprint arXiv:1812.11440*, 2018.

[262] Yuchong Gu, Zitao Zeng, Haibin Chen, Jun Wei, Yaqin Zhang, Binghui Chen, Yingqin Li, Yujuan Qin, Qing Xie, Zhuoren Jiang, et al. Medsrgan: medical images super-resolution using generative adversarial networks. *Multimedia Tools and Applications*, 79:21815–21840, 2020.

[263] Mingfeng Jiang, Minghao Zhi, Liying Wei, Xiaocheng Yang, Jucheng Zhang, Yongming Li, Pin Wang, Jiahao Huang, and Guang Yang. Fa-gan: Fused attentive generative adversarial networks for mri image super-resolution. *Computerized Medical Imaging and Graphics*, 92:101969, 2021.

[264] Senrong You, Baiying Lei, Shuqiang Wang, Charles K Chui, Albert C Cheung, Yong Liu, Min Gan, Guocheng Wu, and Yanyan Shen. Fine perceptive gans for brain mr image super-resolution in wavelet domain. *IEEE transactions on neural networks and learning systems*, 2022.

[265] Erick Costa de Farias, Christian Di Noia, Changhee Han, Evis Sala, Mauro Castelli, and Leonardo Rundo. Impact of gan-based lesion-focused medical image super-resolution on the robustness of radiomic features. *Scientific reports*, 11(1):21361, 2021.

[266] Xin Jiang, Mingzhe Liu, Feixiang Zhao, Xianghe Liu, and Helen Zhou. A novel super-resolution ct image reconstruction via semi-supervised generative adversarial network. *Neural Computing and Applications*, 32:14563–14578, 2020.

[267] Qing Lyu, Chenyu You, Hongming Shan, and Ge Wang. Super-resolution mri through deep learning. *arXiv preprint arXiv:1810.06776*, 2018.

[268] Faezehsadat Shahidi. Breast cancer histopathology image super-resolution using wide-attention gan with improved wasserstein gradient penalty and perceptual loss. *IEEE Access*, 9:32795–32809, 2021.

[269] Qingsong Yang, Pingkun Yan, Yanbo Zhang, Hengyong Yu, Yongyi Shi, Xuanqin Mou, Mannudeep K Kalra, Yi Zhang, Ling Sun, and Ge Wang. Low-dose ct image denoising using a generative adversarial network with wasserstein distance and perceptual loss. *IEEE transactions on medical imaging*, 37(6):1348–1357, 2018.

[270] Qing Lyu, Hongming Shan, Cole Steber, Corbin Helis, Chris Whitlow, Michael Chan, and Ge Wang. Multi-contrast super-resolution mri through a progressive network. *IEEE transactions on medical imaging*, 39(9):2738–2749, 2020.

[271] Se-In Jang, Tinsu Pan, Gary Ye Li, Junyu Chen, Quanzheng Li, and Kuang Gong. Pet image denoising based on transformer: evaluations on datasets of multiple tracers, 2022.

[272] Pengfei Guo, Yiqun Mei, Jinyuan Zhou, Shanshan Jiang, and Vishal M Patel. Reconformer: Accelerated mri reconstruction using recurrent transformer. *arXiv preprint arXiv:2201.09376*, 2022.

[273] Muralikrishna Puttaguntaa, Ravi Subbanb, and Nelson Kennedy Babu Cc. Swinir transformer applied for medical image super-resolution. *Procedia Computer Science*, 204:907–913, 2022.

[274] Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4791–4800, 2021.

[275] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1905–1914, 2021.

[276] Yucheng Tang, Dong Yang, Wenqi Li, Holger R Roth, Bennett Landman, Daguang Xu, Vishwesh Nath, and Ali Hatamizadeh. Self-supervised pre-training of swin transformers for 3d medical image analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20730–20740, 2022.

[277] Jiahao Huang, Yingying Fang, Yinzhe Wu, Huanjun Wu, Zhifan Gao, Yang Li, Javier Del Ser, Jun Xia, and Guang Yang. Swin transformer for fast mri. *Neurocomputing*, 493:281–304, 2022.

[278] Guangyuan Li, Jun Lv, Yapeng Tian, Qi Dou, Chengyan Wang, Chenliang Xu, and Jing Qin. Transformer-empowered multi-scale contextual matching and aggregation for multi-contrast mri super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20636–20645, 2022.

[279] M Angulakshmi and GG Lakshmi Priya. Automated brain tumour segmentation techniques—a review. *International Journal of Imaging Systems and Technology*, 27(1):66–77, 2017.

[280] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, and et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE Transactions on Medical Imaging*, 34(10):1993–2024, Oct 2015.

[281] Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin Kirby, and et al. Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. *Scientific Data*, 4, 09 2017.

[282] Spyridon Bakas, Mauricio Reyes, András Jakab, Stefan Bauer, Markus Rempfler, Alessandro Crimi, and et al. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. *CoRR*, abs/1811.02629, 2018.

[283] Dana C Peters, John V Wylie, Thomas H Hauser, Kraig V Kissinger, René M Botnar, Vidal Essebag, Mark E Josephson, and Warren J Manning. Detection of pulmonary vein and left atrial scar after catheter ablation with three-dimensional

navigator-gated delayed enhancement mr imaging: initial experience. *Radiology*, 243(3):690–695, 2007.

[284] Christopher McGann, Nazem Akoum, Amit Patel, Eugene Kholmovski, Patricia Revelo, Kavitha Damal, Brent Wilson, Josh Cates, Alexis Harrison, Ravi Ranjan, et al. Atrial fibrillation ablation outcome is predicted by left atrial remodeling on mri. *Circulation: Arrhythmia and Electrophysiology*, 7(1):23–30, 2014.

[285] James L Harrison, Christian Sohns, Nick W Linton, Rashed Karim, Steven E Williams, Kawal S Rhode, Jaswinder Gill, Michael Cooklin, C Aldo Rinaldi, Matthew Wright, et al. Repeat left atrial catheter ablation: cardiac magnetic resonance prediction of endocardial voltage and gaps in ablation lesion sets. *Circulation: Arrhythmia and Electrophysiology*, 8(2):270–278, 2015.

[286] Dana C Peters, John V Wylie, Thomas H Hauser, Reza Nezafat, Yuchi Han, Jeong Joo Woo, Jason Taclas, Kraig V Kissinger, Beth Goddu, Mark E Josephson, et al. Recurrence of atrial fibrillation correlates with the extent of post-procedural late gadolinium enhancement: a pilot study. *JACC: Cardiovascular Imaging*, 2(3):308–316, 2009.

[287] Robert S Oakes, Troy J Badger, Eugene G Kholmovski, Nazem Akoum, Nathan S Burgon, Eric N Fish, Joshua JE Blauer, Swati N Rao, Edward VR DiBella, Nathan M Segerson, et al. Detection and quantification of left atrial structural remodeling with delayed-enhancement magnetic resonance imaging in patients with atrial fibrillation. *Circulation*, 119(13):1758–1767, 2009.

[288] Jennifer Keegan, Permi Jhooti, Sonya V Babu-Narayan, Peter Drivas, Sabine Ernst, and David N Firmin. Improved respiratory efficiency of 3d late gadolinium enhancement imaging using the continuously adaptive windowing strategy (claws). *Magnetic resonance in medicine*, 71(3):1064–1074, 2014.

[289] Jennifer Keegan, Peter D Gatehouse, Shouvik Haldar, Ricardo Wage, Sonya V Babu-Narayan, and David N Firmin. Dynamic inversion time for improved 3d late gadolinium enhancement imaging in patients with atrial fibrillation. *Magnetic resonance in medicine*, 73(2):646–654, 2015.

[290] Choukri Mekkaoui, Timothy G Reese, Marcel P Jackowski, Himanshu Bhat, and David E Sosnovik. Diffusion mri in the heart. *NMR in Biomedicine*, 30(3):e3426, 2017.

[291] Sonia Nielles-Vallespin, Zohya Khalique, Pedro F Ferreira, Ranil de Silva, Andrew D Scott, Philip Kilner, Laura-Ann McGill, Archontis Giannakidis, Peter D Gatehouse,

Daniel Ennis, et al. Assessment of myocardial microstructural dynamics by in vivo diffusion tensor cardiac magnetic resonance. *Journal of the American College of Cardiology*, 69(6):661–676, 2017.

[292] Sonia Nielles-Vallespin, Choukri Mekkaoui, Peter Gatehouse, Timothy G Reese, Jennifer Keegan, Pedro F Ferreira, Steve Collins, Peter Speier, Thorsten Feiweier, Ranil De Silva, et al. In vivo diffusion tensor mri of the human heart: reproducibility of breath-hold and navigator-based approaches. *Magnetic resonance in medicine*, 70(2):454–465, 2013.

[293] Margarita Gorodezky, Andrew D Scott, Pedro F Ferreira, Sonia Nielles-Vallespin, Dudley J Pennell, and David N Firmin. Diffusion tensor cardiovascular magnetic resonance with a spiral trajectory: An in vivo comparison of echo planar and spiral stimulated echo sequences. *Magnetic resonance in medicine*, 80(2):648–654, 2018.

[294] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.

[295] Abien Fred Agarap. Deep learning using rectified linear units (relu). *arXiv preprint arXiv:1803.08375*, 2018.

[296] Jacques Froment. Parameter-free fast pixelwise non-local means denoising. *Image Processing On Line*, 4:300–326, 2014.

[297] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017.

[298] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. *Advances in neural information processing systems*, 29, 2016.

[299] Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE transactions on pattern analysis and machine intelligence*, 41(9):2251–2265, 2018.

[300] Yogesh Balaji, Swami Sankaranarayanan, and Rama Chellappa. Metareg: Towards domain generalization using meta-regularization. *Advances in neural information processing systems*, 31, 2018.

[301] David Ha, Andrew M. Dai, and Quoc V. Le. Hypernetworks. *CoRR*, abs/1609.09106, 2016.

[302] Luca Bertinetto, João F Henriques, Jack Valmadre, Philip Torr, and Andrea Vedaldi. Learning feed-forward one-shot learners. *Advances in neural information processing systems*, 29, 2016.

[303] Sameer Khan and Suet-Peng Yong. A comparison of deep learning and hand crafted features in medical image modality classification. In *2016 3rd International Conference on Computer and Information Sciences (ICCOINS)*, pages 633–638. IEEE, 2016.

[304] Daniel S Marcus, Tracy H Wang, Jamie Parker, John G Csernansky, John C Morris, and Randy L Buckner. Open access series of imaging studies (oasis): cross-sectional mri data in young, middle aged, nondemented, and demented older adults. *Journal of cognitive neuroscience*, 19(9):1498–1507, 2007.

[305] Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester, et al. Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved? *IEEE transactions on medical imaging*, 37(11):2514–2525, 2018.

[306] P An, S Xu, SA Harmon, EB Turkbey, TH Sanford, A Amalou, M Kassin, N Varble, M Blain, V Anderson, F Patella, G Carrafiello, BT Turkbey, and BJ Wood. Ct images in covid-19 [data set]. *The Cancer Imaging Archive.*, 2020.

[307] Matthew Brett, Christopher J. Markiewicz, Michael Hanke, Marc-Alexandre Côté, Ben Cipollini, Paul McCarthy, Dorota Jarecka, Christopher P. Cheng, Yaroslav O. Halchenko, Michiel Cottaar, Eric Larson, Satrajit Ghosh, Demian Wassermann, Stephan Gerhard, Gregory R. Lee, Hao-Ting Wang, Erik Kastman, Jakub Kaczmarzyk, Roberto Guidotti, Jonathan Daniel, Or Duek, Ariel Rokem, Cindee Madison, Dimitri Papadopoulos Orfanos, Anibal Sólon, Brendan Moloney, Félix C. Morency, Mathias Goncalves, Zvi Baratz, Ross Markello, Cameron Riddell, Christopher Burns, Jarrod Millman, Alexandre Gramfort, Jaakko Leppäkangas, Jasper J.F. van den Bosch, Robert D. Vincent, Henry Braun, Krish Subramaniam, Andrew Van, Krzysztof J. Gorgolewski, Pradeep Reddy Raamana, Julian Klug, B. Nolan Nichols, Eric M. Baker, Soichi Hayashi, Basile Pinsard, Christian Haselgrove, Mark Hymers, Oscar Esteban, Serge Koudoro, Fernando Pérez-García, Jérôme Dockès, Nikolaas N. Oosterhof, Bago Amirbekian, Ian Nimmo-Smith, Ly Nguyen, Samir Reddigari, Samuel St-Jean, Egor Panfilov, Eleftherios Garyfallidis, Gael Varoquaux, Jon Haitz Legarreta, Kevin S. Hahn, Lea Waller, Oliver P. Hinds, Bennet Fauber, Jacob Roberts, Jean-Baptiste Poline, Jon Stutters, Kesshi Jordan, Matthew Cieslak, Miguel Estevan Moreno, Tomáš Hrnčiar, Valentin Haenel, Yannick Schwartz,

Benjamin C Darwin, Bertrand Thirion, Carl Gauthier, Igor Solovey, Ivan Gonzalez, Jath Palasubramaniam, Justin Lecher, Katrin Leinweber, Konstantinos Raktivan, Markéta Calábková, Peter Fischer, Philippe Gervais, Syam Gadde, Thomas Ballinger, Thomas Roos, Venkateswara Reddy Reddam, and freec84. nipy/nibabel: 5.0.0, January 2023.

[308] Andrew L Maas, Awni Y Hannun, Andrew Y Ng, et al. Rectifier nonlinearities improve neural network acoustic models. *Proc. icml*, 30(1):3, 2013.

[309] Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun 2018.

[310] Pramod Gaur, Karl McCreadie, Ram Bilas Pachori, Hui Wang, and Girijesh Prasad. Tangent space features-based transfer learning classification model for two-class motor imagery brain–computer interface. *International journal of neural systems*, 29(10):1950025, 2019.

[311] Andrés Ortiz, Jorge Munilla, Francisco J Martínez-Murcia, Juan M Górriz, and Javier Ramírez. Empirical functional pca for 3d image feature extraction through fractal sampling. *International journal of neural systems*, 29(02):1850040, 2019.

[312] Xiuhui Wang and Wei Qi Yan. Human gait recognition based on frame-by-frame gait energy images and convolutional long short-term memory. *International journal of neural systems*, 30(01):1950027, 2020.

[313] Mohammad Hesam Hesamian, Wenjing Jia, Xiangjian He, and Paul Kennedy. Deep learning techniques for medical image segmentation: achievements and challenges. *Journal of digital imaging*, 32(4):582–596, 2019.

[314] Xintao Wang, Chao Dong, and Ying Shan. Repsr: Training efficient vgg-style super-resolution networks with structural re-parameterization and batch normalization. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 2556–2564, 2022.

[315] Matthias S Treder, Ryan Codrai, and Kamen A Tsvetanov. Quality assessment of anatomical mri images from generative adversarial networks: Human assessment and image quality metrics. *Journal of Neuroscience Methods*, 374:109579, 2022.

[316] Octavio Martinez Manzanera, Sanne K Meles, Klaus L Leenders, Remco J Renken, Marco Pagani, Dario Arnaldi, Flavio Nobili, Jose Obeso, Maria Rodriguez Oroz, Silvia Morbelli, et al. Scaled subprofile modeling and convolutional neural networks for the identification of parkinson's disease in 3d nuclear imaging data. *International journal of neural systems*, 29(09):1950010, 2019.

[317] Kelei He, Chen Gan, Zhuoyuan Li, Islem Rekik, Zihao Yin, Wen Ji, Yang Gao, Qian Wang, Junfeng Zhang, and Dinggang Shen. Transformers in medical image analysis: A review. *arXiv preprint arXiv:2202.12165*, 2022.

[318] Xin He, Yong Zhou, Jiaqi Zhao, Di Zhang, Rui Yao, and Yong Xue. Swin transformer embedding unet for remote sensing image semantic segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–15, 2022.

[319] Zhou Wang and Alan C Bovik. Mean squared error: Love it or leave it? a new look at signal fidelity measures. *IEEE signal processing magazine*, 26(1):98–117, 2009.

[320] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. IEEE, 2016.

[321] Jun Ma, Jianan Chen, Matthew Ng, Rui Huang, Yu Li, Chen Li, Xiaoping Yang, and Anne L Martel. Loss odyssey in medical image segmentation. *Medical Image Analysis*, 71:102035, 2021.

[322] Jun Ma, Yixin Wang, Xingle An, Cheng Ge, Ziqi Yu, Jianan Chen, Qiongjie Zhu, Guoqiang Dong, Jian He, Zhiqiang He, et al. Toward data-efficient learning: A benchmark for covid-19 ct lung and infection segmentation. *Medical physics*, 48(3):1197–1210, 2021.

[323] Pavel Iakubovskii. Segmentation models pytorch. `https://github.com/qubvel/segmentation_models.pytorch`, 2019.

[324] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11976–11986, 2022.

[325] Xiaohan Ding, Xiangyu Zhang, Jungong Han, and Guiguang Ding. Scaling up your kernels to 31x31: Revisiting large kernel design in cnns. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11963–11975, 2022.

[326] Xue Ying. An overview of overfitting and its solutions. *Journal of physics: Conference series*, 1168(2):022022, 2019.

[327] Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. Dynamicvit: Efficient vision transformers with dynamic token sparsification. *Advances in neural information processing systems*, 34:13937–13949, 2021.

[328] Yehui Tang, Kai Han, Yunhe Wang, Chang Xu, Jianyuan Guo, Chao Xu, and Dacheng Tao. Patch slimming for efficient vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12165–12174, 2022.

[329] Jelmer M Wolterink, Anna M Dinkla, Mark HF Savenije, Peter R Seevinck, Cornelis AT van den Berg, and Ivana Išgum. Deep mr to ct synthesis using unpaired data. In *International workshop on simulation and synthesis in medical imaging*, pages 14–23. Springer, 2017.

[330] Shengke Xue, Zhaowei Cheng, Guangxu Han, Chaoliang Sun, Ke Fang, Yingchao Liu, Jian Cheng, Xinyu Jin, and Ruiliang Bai. 2d probabilistic undersampling pattern optimization for mr image reconstruction. *Medical Image Analysis*, 77:102346, 2022.

[331] Runze Han, Craig K Jones, J Lee, Pengwei Wu, Prasad Vagdargi, Ali Uneri, Patrick A Helm, M Luciano, William S Anderson, and Jeffrey H Siewerdsen. Deformable mr-ct image registration using an unsupervised, dual-channel network for neurosurgical guidance. *Medical image analysis*, 75:102292, 2022.

[332] Clément Playout, Renaud Duval, Marie Carole Boucher, and Farida Cheriet. Focused attention in transformers for interpretable classification of retinal images. *Medical Image Analysis*, 82:102608, 2022.

[333] He Zhao, Huiqi Li, Sebastian Maurer-Stroh, and Li Cheng. Synthesizing retinal and neuronal images with generative adversarial nets. *Medical image analysis*, 49:14–26, 2018.

[334] Xiaohan Ding, Xiangyu Zhang, Jungong Han, and Guiguang Ding. Scaling up your kernels to 31x31: Revisiting large kernel design in cnns. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11963–11975, 2022.

[335] Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VII*, pages 17–33. Springer, 2022.

[336] Yasin Almalioglu, Kutsev Bengisu Ozyoruk, Abdulkadir Gokce, Kagan Incetan, Guliz Irem Gokceler, Muhammed Ali Simsek, Kivanc Ararat, Richard J Chen, Nicholas J Durr, Faisal Mahmood, et al. Endol2h: deep super-resolution for capsule endoscopy. *IEEE Transactions on Medical Imaging*, 39(12):4297–4309, 2020.

[337] Mehdi SM Sajjadi, Raviteja Vemulapalli, and Matthew Brown. Frame-recurrent video super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6626–6634, 2018.

[338] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*, 2022.

[339] Zhe Gan, Linjie Li, Chunyuan Li, Lijuan Wang, Zicheng Liu, Jianfeng Gao, et al. Vision-language pre-training: Basics, recent advances, and future trends. *Foundations and Trends® in Computer Graphics and Vision*, 14(3–4):163–352, 2022.

[340] Hao Li, Jinguo Zhu, Xiaohu Jiang, Xizhou Zhu, Hongsheng Li, Chun Yuan, Xiaohua Wang, Yu Qiao, Xiaogang Wang, Wenhai Wang, et al. Uni-perceiver v2: A generalist model for large-scale vision and vision-language tasks. *arXiv preprint arXiv:2211.09808*, 2022.

[341] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[342] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2085–2094, 2021.

[343] Ming Ding, Wendi Zheng, Wenyi Hong, and Jie Tang. Cogview2: Faster and better text-to-image generation via hierarchical transformers. *arXiv preprint arXiv:2204.14217*, 2022.

[344] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.