Deep learning for medical image super resolution and segmentation

Chi-Hieu Pham

UNIVERSITE BRETAGNE LOIRE / MATHSTIC

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

THESE DE DOCTORAT DE

L'ÉCOLE NATIONALE SUPERIEURE MINES-TELECOM ATLANTIQUE
BRETAGNE PAYS DE LA LOIRE - IMT ATLANTIQUE
COMUE UNIVERSITE BRETAGNE LOIRE

ECOLE DOCTORALE N° 601
Mathématiques et Sciences et Technologies de l'Information et de la Communication
Spécialité : Signal - Image - Vision

Par

Chi-Hieu PHAM

Apprentissage profond pour la super-résolution et la segmentation d'images médicales

Thèse présentée et soutenue à Brest, le 20/12/2018
Unité de recherche : LaTIM Inserm U1101
Thèse N° : 2018IMTA0124

Rapporteurs avant soutenance :

Su RUAN	Professeure des universités, Université de Rouen
Yuemin ZHU	Directeur de recherche, CNRS

Composition du Jury :

Président :	Valérie BURDIN	Professeur, IMT Atlantique
Examinateurs :	Su RUAN	Professeure des universités, Université de Rouen
	Yuemin ZHU	Directeur de recherche, CNRS
	Vincent NOBLET	Ingénieur de recherche HDR, CNRS
	Ronan FABLET	Professeur, IMT Atlantique
	François ROUSSEAU	Professeur, IMT Atlantique
Dir. de thèse :	François ROUSSEAU	Professeur, IMT Atlantique

Contents

Acknowledgements

Part of this work was funded by ANR MAIA project, grant ANR-15-CE23-0009 of the French National Research Agency, INSERM and Institut Mines Télécom Atlantique (Chaire "Imagerie médicale en thérapie interventionnelle").

I am especially grateful to Prof. Su Ruan and Prof. Yuemin Zhu for reviewing my thesis. I appreciate their interest in my work as well as all of their insightful comments and suggestions. I would like to thank M. Vincent Noblet, Prof. Valérie Burdin and Prof. Ronan Fablet for taking part in my PhD defense.

In particular, I would especially like to express my special thanks of gratitude to my supervisor Prof. François Rousseau who gave me the golden opportunity to do this wonderful thesis, which leverage me to learn a great amount of knowledge and experience in the fields of computer vision, image processing as well as machine learning and deep learning. His patience, motivation, immense knowledge guided me in all the time of research. "Merci beaucoup François !"

I am really thankful to Prof. Valérie Burdin (IMT Atlantique) et Prof. Mireille Gareau (Université de Rennes 1) who have gave a chance to study in France. I am thankful to Prof. Jean-Louis Dillenseger (Université de Rennes 1) for guiding me during my master internship. I also would like to thank all my professors and my colleges from IMT Atlantique, LaTIM and "Domino Reading Group" as well as Ho Chi Minh University of Technology.

A very special thank you goes to my friends in "4 bis rue de Kérangoff", "Bờ zet hội"and "Hội cầu lông Brest"for the moment we share together.

Last but not the least, thanks to "Pa", "Mẹ", "Ông Bà", "Vợ, "Bi Bô"and my big family for always being with me through every step of my life.

Résumé étendu

La modélisation des images est un point clef important pour de nombreuses tâches en traitement d'images, comme la super-résolution, la segmentation ou la synthèse de texture. L'analyse et le traitement des différentes caractéristiques composant une image, nécessite la mise en place d'approches adaptatives locales. Dans ce contexte, la définition de représentations locales efficaces est dédiée pour une application visée. Les approches par l'apprentissage profond ont permis des avancées significatives en terme de performance de traitement. D'un côté, cette approche se fond sur la notion de relation entre les patches, et plus particulièrement en analysant automatiquement les caractéristiques des patches pour les agréger par la suite. D'un autre côté, les méthodes d'apprentissage profond reposent sur une hypothèse de fonctionnement des neurones biologiques d'une représentation de l'image par un ensemble de filtrage. Ces représentations ont été introduites dans des modèles a priori dans le cadre de résolution de problèmes inverses.

L'objectif de cette thèse est d'étudier le comportement de différentes représentations d'images, notamment par apprentissage profond, dans le contexte d'application en imagerie médicale. Le but est de développer une méthode unifiée efficace pour les applications visées que sont la super-résolution, la segmentation et la synthèse. La super-résolution est un processus

d'estimation d'une image haute-résolution à partir d'une ou plusieurs images basses-résolutions. Dans cette thèse, nous nous concentrons sur la super-résolution unique, c'est-à-dire que l'image haute-résolution (HR) est estimée par une image basse-résolution (LR) correspondante. Augmenter la résolution de l'image grâce à la super-résolution est la clé d'une compréhension plus précise de l'anatomie. L'application de la super-résolution permet d'obtenir des cartes de segmentation plus précises. Étant donné que deux bases de données qui contiennent les images différentes (par exemple, les images d'IRM et les images de CT), la synthèse est un processus d'estimation d'une image qui est présentée dans la base de données de cible à partir d'une image de la base de données de source. Parfois, certains contrastes tissulaires ne peuvent pas être acquis pendant la séance d'imagerie en raison du temps et des coûts élevés ou de l'absence d'appareils. Une solution possible est à utiliser des méthodes de synthèse d'images médicales pour générer les images avec le contraste différent qui est manquée dans le domaine à cible à partir de l'image du domaine donnée. L'objectif des images synthétiques est d'améliorer d'autres étapes du traitement automatique des images médicales telles que la segmentation, la super-résolution ou l'enregistrement. Dans cette thèse, nous proposons les réseaux neurones pour la super-résolution et la synthèse d'image médicale. Les résultats démontrent le potentiel de la méthode que nous proposons en ce qui concerne les applications médicales pratiques.

Un réseau de neurones convolutifs (en anglais CNN - Convolutional Neural Networks) est un type de réseau de neurones artificiels. Une architecture de réseau de neurones convolutifs est structurée par un ensemble de couches de traitement, particulièrement les couches convolutives et les fonctions d'activation. En outre, selon une application visée, nous pouvons

rajouter les autre éléments comme la couche de pooling, la couche entièrement connectée (fully connected) ou la couche convolutifs transposée. Le réseau de neurones convolutives, qui fut présenté il y a longtemps [LeCun et al., 1998] a vraiment reçu l'attention de la communauté de recherche à partir de 2012 par une méthode ayant gagné un challenge de classification dans une conférence de vision par ordinateur. Ce réseau appelé Alexnet [Krizhevsky et al., 2012] contient huit couches: cinq couches convolutives et trois couches entièrement connectées. Ensuite, les architectures de CNN sont devenues l'état de l'art pour de nombreuses tâches en traitement d'images comme la super-résolution [Dong et al., 2016a, Kim et al., 2016a], la segmentation [Kamnitsas et al., 2017, Ronneberger et al., 2015] ou la classification [He et al., 2016a, Simonyan and Zisserman, 2014]. Par la suite, plusieurs réseaux CNNs ont été améliorés afin d'augmenter leur performance pour la classification, par exemple, en augmentant le nombre de couches (e.g. VGGnet avec 19 couches [Simonyan and Zisserman, 2014], Resnet avec 152 couches [He et al., 2016a]), en concaténant les filtres en un bloc (e.g. GoogLenet [Szegedy et al., 2015]), ou par l'apprentissage résiduel en bloc (e.g. Resnet [He et al., 2016b]), ou en connectant tous les couches à forte densité (e.g. Densenet [Huang et al., 2017a]). Afin de détecter l'objet dans les images avec CNNs, nous pouvons attacher la boîte de délimitation parallèle à sa classification d'objet (R-CNN) [Girshick et al., 2014] et sa segmentation (Mask R-CNN) [He et al., 2017]. Plusieurs méthodes de CNNs ont été proposées pour la segmentation de l'imagerie médicale. On peut notamment citer U-Net [Ronneberger et al., 2015] et DeepMedic [Kamnitsas et al., 2017]. U-net a une forme de la lettre U avec les skip-connections entre les couches. DeepMedic combine deux réseaux de CNNs afin d'augmenter la performance de segmentation cérébrale: un chemin pour l'image original et une autre pour sa version de

basse-résolution.

Les architectures CNN sont devenues l'état de l'art en super-résolution (SR). Initialement, [Dong et al., 2016a] a proposé une architecture CNN à trois couches. La première couche convolutionnelle extrait implicitement un ensemble des caractéristiques pour l'image LR d'entrée, la deuxième couche représente non-linéairement des caractéristiques de l'image basse-resolution aux patches haute-résolution et la troisième couche reconstruit l'image HR à partir de ces représentations de patchs. Et puis, les caractéristiques suivantes ont été rapportées pour améliorer la performance SR tel que un réseau plus profond [Kim et al., 2016a], bloc résiduel [Ledig et al., 2017], couche de sous-pixel [Shi et al., 2016], fonction de coût perceptuelle (au lieu de fonctions de coût quadratiques moyennes)[Johnson et al., 2016, Ledig et al., 2017, Zhao et al., 2017], réseaux récurrents [Kim et al., 2016b], le réseau contradictoire générateur [Ledig et al., 2017], Très récemment, [Chen et al., 2018b] ont proposé une version 3D de densenet pour la SR des image IRM. Inspiré du travail de [Jog et al., 2016], [Zhao et al., 2018] a étudié la super-résolution automatique pour l'IRM en utilisant des réseaux résiduels profonds [Lim et al., 2017]. Récemment, un réseau plus profond avec 20 couches [Kim et al., 2016a] inspiré par VGGnet [Simonyan and Zisserman, 2014] est devenu une basé pour les méthodes suivantes [Timofte et al., 2017]. Cependant, en raison de la variété des méthodes proposées et du nombre de paramètres pour l'architecture des réseaux, il est actuellement difficile d'identifier les componants clés de l'architecture CNN pour obtenir des bonnes performances pour la SR et évaluer leur applicabilité dans le contexte de l'image IRM cérébrale 3D. De plus, l'extension des architectures CNN aux images 3D, en tenant compte des facteurs de mise à l'échelle anisotropes peut être intéressante pour

s'adresser aux nombreux paramètres d'acquisition clinique possibles, tandis que les architectures CNN classiques n'adressent qu'un facteur d'échelle prédéfini. La disponibilité de l'imagerie multimodale pose également la question sur la capacité des architectures de CNN à exploiter de telles données multimodales pour améliorer la SR d'une modalité donnée.

Ce manuscrit est rédigé en anglais et structuré en cinq chapitres et une annexe.

Le chapitre 1 correspond à une introduction générale où sont décrits le contexte, la motivation, l'objectif et la méthodologie de cette thèse.

Le chapitre 2 décrit notre méthode de super-résolution en imagerie cérébrale en utilisant les réseaux CNNs (convolutional neural networks). D'abord, nous allons passer une bibliographie qui contient différents types de méthodes de super-résolution tel que la méthode basée sur les modèles et celle basée sur l'apprentissage comprenant les réseaux de neurones convolutifs. Et puis, la méthode de super-résolution est consacrée à l'imagerie médicale. Ensuite, nous proposon l'application de la méthode de super-résolution basée sur CNNs aux images cérébrales d'IRM. Il s'agit de l'application des CNNs 3D afin d'obtenir la super-résolution à partir d'une seule image. Huit paramètres principales du réseau 3D sont étudiés en détail avec des expérimentations pour améliorer sa performance: méthodes d'optimisation, initialisation des poids, apprentissage résiduel, profondeur du réseau, taille du filtre, nombre de filtres, taille de patch d'apprentissage, nombre de sujets pour l'apprentissage. Pour exploiter la capacité du réseau, deux autres applications sont proposées. Le premier est à mélanger plusieurs facteurs échelles (par exemple, mis à échelle deux fois et trois fois par rapport d'une basse résolution) dans le même ensemble de données d'apprentissage. Le

réseau appris avec plusieur facteurs peut être appliqué pour des échelles arbitraires tandis que celui appris une facteur est seulement utilisé pour une résolution désirée. Le deuxième application vise à concaténer les images haute-résolution référence pondérée et l'image bass-résolution interpolée à l'entrée du réseau CNN. Par rapport à ces deux applications, la diversité des bases de données d'apprentissage est également abordée. Enfin, nous appliquons notre méthode de super-résolution aux images cérébrales d'IRM des nouveau-nés et puist segmenter les images de haute résolution obtenues afin d'évaluer la contribution de la méthode proposée. Nous montons les résultats visuelles. Les illustrations contribute que la super-résolution peut aider la segmentation d'image.

Le chapitre 3 décrit une approche pour une réalisation simultanée de la superrésolution et de la segmentation à partir d'une seule image. Elle est basée sur le réseau de neurones génératives contradictoires dit generative adversarial network (GAN). La superrésolution et la segmentation sont souvent effectuées de manière séparée comme la section dernière du chapitre 2. Dans cet chapitre, nous proposons réaliser ces deux opérations en même temps. L'application est focalisée seulement sur des images IRM néonatales du cerveau en T2 qui ont les résolutions basses, car les nouveaunés ne peuvent pas patienter allonger sur une machine d'acquisition dans plusieurs cases clinques. Les résultats de la super-résolution sont comparés avec la méthode proposée dans le chapitre précédent. Les images haute-resolution estimées semblent légèrement inférieurs en termes des métriques de qualité mais meilleurs visuellement. Concernant les résultats de la segmentation qui sont évalués par DICE, notre méthode montre les meilleurs résultats comparés avec deux méthodes de segmentation de littérature.

Le chapitre 4 introduit la synthèse d'images médicale. Une synthèse des

méthodes existantes basées principalement sur l'apprentissage et sur le réseau CNN est fait dans la première section. Ensuite, nous proposons deux approches basées respectivement sur le réseau CNNs et sur le réseau GAN. La première est directement appliquée du principe qui a été utilisé pour la super-résolution décrit dans le deuxième chapitre pour synthétiser des images couplées, c'est-à-dire, dans la base d'apprentissage, les deux séquences d'images sont toutes appairées (paired cross-modal synthesis). Les résultats rassemblent aux images vérité-terrain mais avec un peu de floue et de bruit. En plus, nous considérons le deuxième cas plus difficile: les deux séquences d'images sont toutes non appairées (unpaired cross-modal synthesis). Cette méthode de synthèse d'images est basée sur le réseau GAN. Afin de resoudre le problème de synthèse d'image non appariée, nous proposons utiliser trois fonctions de coût: adversarial loss, cycle consistency loss et total variation. Cependant, il reste la difficulté pour choisir un coefficient optimal de pondération pour total variation qui controle le compromis entre la réduction des artéfacts de haute fréquence et le flou induit. Malgré cela, nous croyons que notre nouvelle approche peut amener de nouvelles perspectives très intéressantes pour beaucoup d'applications.

Le chapitre 5 conclut notre thèse et ouvre des perspectives.

Acronyms

2D	Two dimensions
3D	Three dimensions
3T	3 Tesla Machine
Adam	Adaptive Moment Estimation
BN	Batch Normalization
CNN	Convolutional Neural Network
CSF	Cerebrospinal fluid
Conv	Convolution
dB	decibel
FLAIR	Fluid-Attenuated Inversion Recovery
FOV	Field Of View
GAN	Generative Adversarial Network
GC	Gradient Clipping
GM	Gray matter
GPU	Graphics Processing Unit
HR	High Resolution
InsNorm	Instance Normalization
ILR	Interpolated Low Resolution
LR	Low Resolution
LRTV	Low Rank Total Variation
MAE	Mean Absolute Error
MPRAGE	Magnetization-Prepared Rapid Gradient-Echo

MRI Magnetic Resonance Imaging

MSE Mean Squared Error

NAG Nesterov's Accelerated Gradient

Net Network

NMU Non-local Means Upsampling

PCA Principal Component Analysis

PSNR Peak Signal to Noise Ratio

ReLU Rectified Linear Unit

RMSProp Root-Mean-Square Propagation

SGD Stochastic Gradient Descent

SNR Signal to Noise Ratio

SSIM Structural Similarity index

SR Super Resolution

T1w T1-weighted

T2w T2-weighted

TSE Turbo-spin Echo

WM White matter

Chapter 1

Introduction

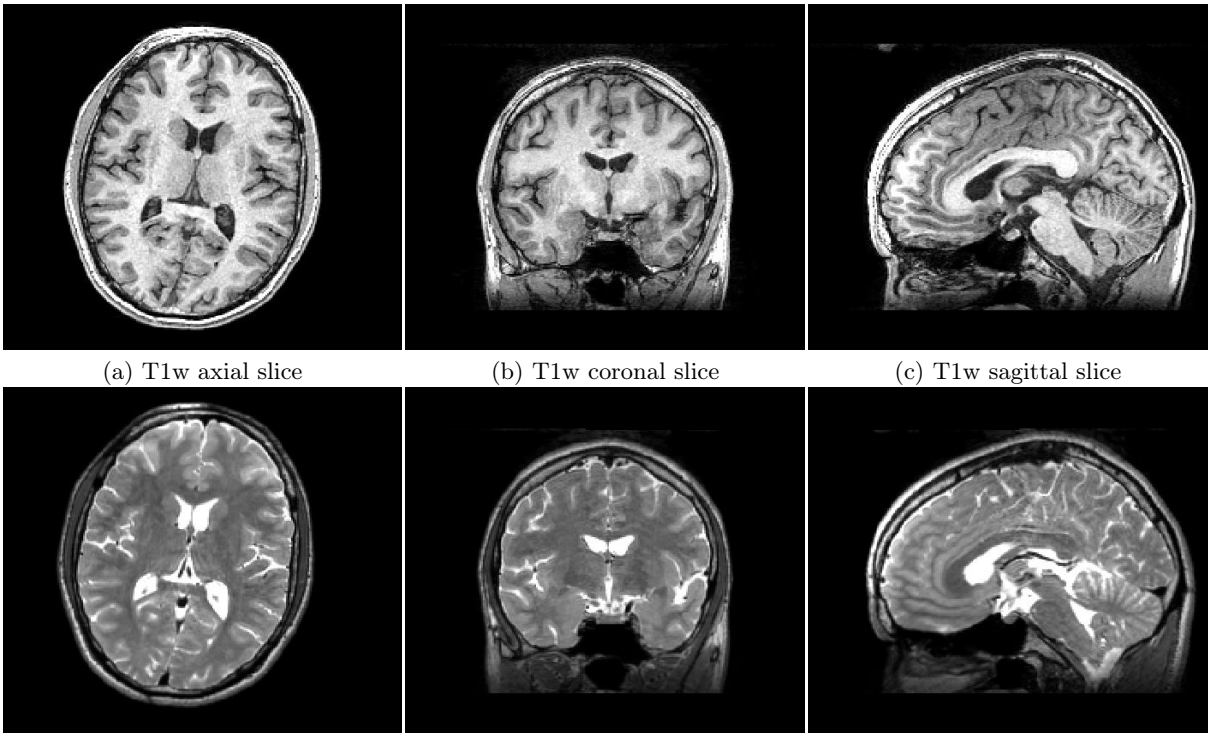
Contents

1.1 Context and motivation

MRI is a medical imaging technique used to visualize the anatomy and the physiological processes of the body. MRI scanners are based on the interaction of a nuclear spin with an external magnetic field. The rotation of a particle around some axis as an intrinsic form of angular momentum is called spin. An MRI scanner forms a strong magnetic field around the area of a subject to be imaged. The protons of hydrogen atoms from biological organisms are excited by a radio frequency (RF) pulse and then emit energy in the form of RF signal when returning to the original state. By applying different types of the sequence of RF pulses, different types of modality are created. Two important terms of the acquisition process are repetition time (TR) and time to echo (TE). TR denotes the period between successive pulse sequences at the same slice, TE denotes the period between the emission of the RF pulse and the reception of the echo signal. Common anatomical MRI sequences are T1-weighted (T1w) images, T2-weighted (T2w) images and fluid-attenuated inversion recovery (FLAIR). T1w, T2w MRIs and FLAIR are generated by using respectively short TE and TR, long TR and TE, and very long TR and TE times. A modality shows up different physical properties of tissue, that induces different contrasts between them. Thus, each modality has a specific range of applications in medical diagnosis.

1.1.1 Medical single image super-resolution: an approach to generate high-resolution images

Acquisition time of MRI data and signal-to-noise ratio are two parameters that drive the choice of an appropriate image resolution for a given study. The accuracy of further analysis such as brain morphometry can be highly dependent on image resolution. A typical image resolution of a current MRI is desired for greater than or equal to $1mm$. However, imaging with desired resolutions costs of low signal to noise ratio and long scan time. For example, MR images with an isotropic resolution of $1mm$ and $0.7mm$ are respectively shown in Figure 1.1 and Figure 1.2. Visually, the image with higher resolution of $0.7mm$ visualizes better the anatomy of a brain. In addition, the objective of medical imaging systems is aimed at increasing the resolution to create true isotropic 3D imaging. Isotropic 3-D MRI images with high resolution are a key role for visualization of 3D volumes and for early medical diagnosis. Nevertheless, in many clinical cases, radiology procedures do not allow to achieve possibly isotropic resolutions such as neonatal brain scan. Figure 1.3 shows an example of real anisotropic 3-D images. These reasons raise a question that finding a post-processing method which can augment the resolution of low-resolution images or enhance them to achieve an isotropic high-resolution images.

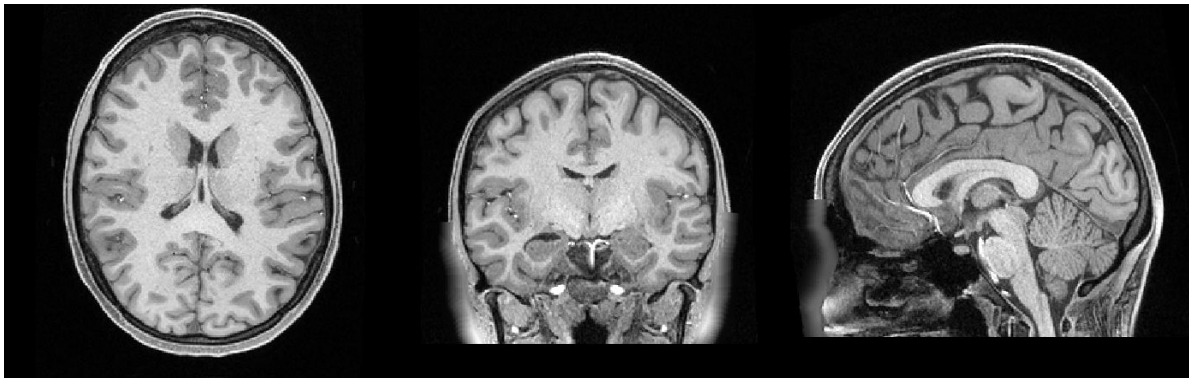


(a) T1w axial slice (b) T1w coronal slice (c) T1w sagittal slice

(d) T2w axial slice (e) T2w coronal slice (f) T2w sagittal slice

FIGURE 1.1: Adult brain MRI (Subject: 01011-t1w of the dataset NAMIC). The voxel size of the images is $1 \times 1 \times 1$mm.

Super-Resolution (SR) aims to enhance the image resolution using single or multiple data acquisitions [Milanfar, 2010]. Increasing image resolution through super-resolution is a key to more accurate understanding of the anatomy [Greenspan, 2008]. The applications of super-resolution have been shown that applying super-resolution techniques leads to more accurate



(a) Axial slice (b) Coronal slice (c) Sagittal slice

FIGURE 1.2: Adult brain MRI (Subject: 100307 of the dataset HCP100). The voxel size of the images is $0.7 \times 0.7 \times 0.7$mm.

segmentation maps of brain MRI data [Rueda et al., 2013] or cardiac data [Oktay et al., 2016].

Recently, a series of papers suggested the successful application of deep learning, leading to state-of-the-art results in many practically tasks of computer vision [Dong et al., 2014, Kim et al., 2016a, Krizhevsky et al., 2012, Shi et al., 2016] and medical image processing [Charron et al., 2018, Chen et al., 2018b, Kamnitsas et al., 2017, Meyer et al., 2018, Oktay et al., 2016, Pham et al., 2017a, Ronneberger et al., 2015]. In this thesis, the architectures of convolutional neural networks are investigated for MRI super-resolution. The performance of a given architecture depends on several parameters such as the filter size, the number of filters, the number of layers, etc. Understanding how these parameters affect the reconstruction of the HR image with respect to the considered application setting (e.g., number of training samples, image size, scaling factor) is a key issue, which remains poorly explored. For instance, regarding the number of layers, it is commonly believed that the deeper the better [Kim et al., 2016a, Simonyan and Zisserman, 2014]. However, adding layers increases the number of parameters and can lead to overfitting. Previous works [Dong et al., 2016a, Oktay et al., 2016], have shown that "a deeper structure does not always lead to better results" [Dong et al., 2016a].

Specifically focusing on MRI data, the specific objectives of this study are:

- Are 2D or 3D networks relevant to brain MRI SR ?
- the evaluation and understanding of the effect of key elements of CNN for brain MRI SR
- How can networks handle different scaling factors ?
- Investigating multimodality-guided SR using CNN ?
- Can a pre-trained model apply to different data ?
- How do networks apply to a real data ?

- Does the application of super-resolution improve automatic segmentation algorithms ?
- Two steps for the 3D isotropic segmentation of anisotropic MRI images are: increasing the image resolution using interpolation techniques or SR and then isotropic image segmentation. Do we have a method for simultaneous super-resolution and segmentation ?

1.1.2 Medical image cross-modal synthesis: an approach to generate synthesized images

The pulse sequences in the acquisition process influence strongly the performance of MRI analysis algorithms. Medical image analysis techniques, which optimally learned with data from one specific modality, could not apply to data of a different modality because each modality expresses particular tissue contrast of the body anatomy. For example, neonatal brain T2w MRIs are appropriate to reconstruct brain surface while the T1w scans lack sufficient tissue contrast [Leroy et al., 2011]. Sometimes, certain tissue contrasts may not be acquired during the imaging session because of time-consuming, expensive cost or lacking of devices. One possible solution is to use medical image cross-modal synthesis methods to generate the missing subject-specific scans in the desired target domain from the given source image domain. The objective of synthetic images is to improve other automatic medical image processing steps such as segmentation, super-resolution or registration.

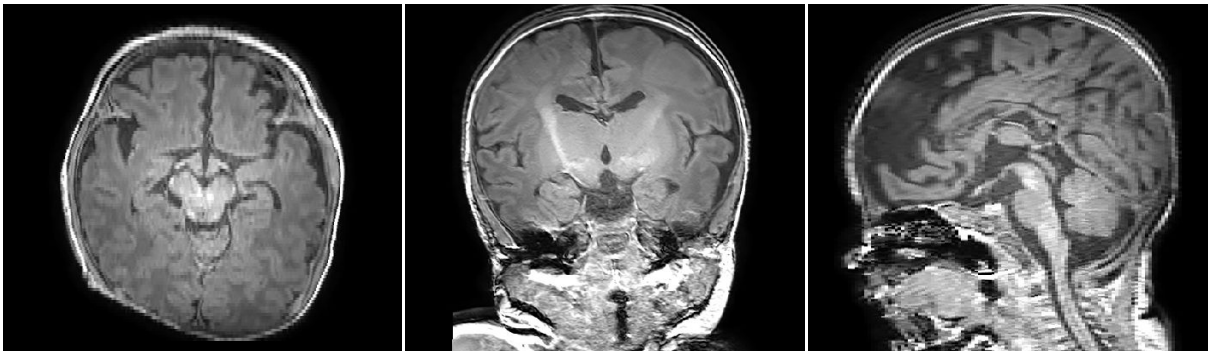
In this thesis, convolutional neural networks are applied to cross-modal synthesis in the context of supervised learning. In addition, an attempt to apply generative adversarial networks for unpaired cross-modal synthesis brain MRI is described. The specific objectives of this study are:

- Can CNN-based methods be applied to solve cross-modal synthesis problem ?
- Is there a method which can generate the synthetic image of a specific subject given unpaired training dataset ?

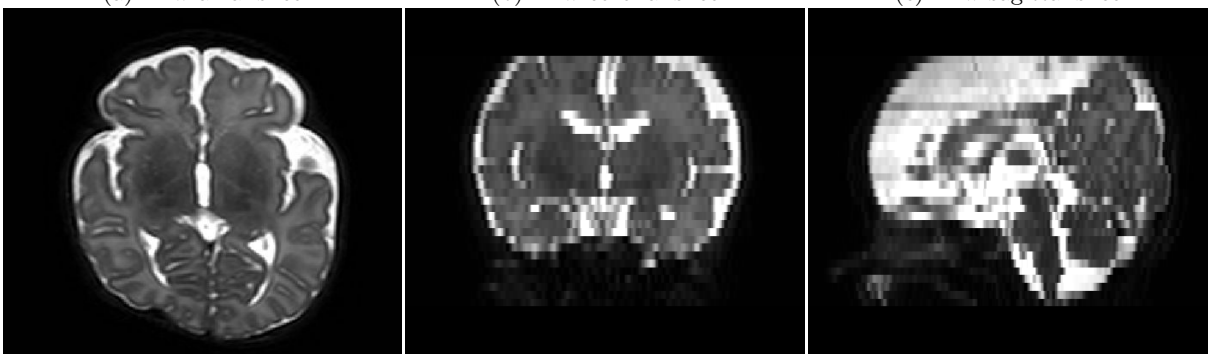
1.2 Thesis overview

In this thesis, our motivation is dedicated to studying the behaviours of different image representations and developing a method for super-resolution, cross-modal synthesis and segmentation of medical imaging.

Chapter 2 introduces single image super-resolution. Firstly, single image super-resolution is first modelled by the image acquisition process. Several methods for super-resolution of natural images are discussed from model-based to learning-based approaches. Since the observation model is assumed unknown or hard-defined, "blind super-resolution" is then



(a) T1w axial slice (b) T1w coronal slice (c) T1w sagittal slice



(d) T2w axial slice (e) T2w coronal slice (f) T2w sagittal slice

FIGURE 1.3: Neonatal brain MRI (Subject: S00007 of the dataset MAIA). The voxel sizes of the T1w image and the T2 image are respectively about 0.2679 $\times$ 0.2679 $\times$ 1.2mm and $0.4464 \times 0.4464 \times 3$mm.

mentioned to estimate the point spread function of the acquisition process. However, many example-based super-resolution methods rely on an external database. When a training set is not available, zero-shot learning super-resolution algorithms are proposed to exploit the internal cross-scale patches internally within the testing image. After the review of 2D natural super-resolution methods, the applications of learning-based and model-based super-resolution in medical imaging are reviewed. Secondly, chapter 2 introduces convolutional neural networks approaches for brain MRI super-resolution. Experiments demonstrate the need of 3D networks for 3D data generation instead of 2D networks due to the ability of 3D representations of pre-trained filter set of 3D layers. Next, performance analysis of the network architecture with respect to various algorithmic design choices such as: optimization methods, weight initialization, residual learning, the depth of networks, filter size, number of filters, training patch size and training subject number. A multi-scale training approach is then proposed to handle arbitrary magnification factors. Moreover, the convolutional networks are extended to leverage information of multimodal input for improved SR reconstructions. In addition, two datasets are used to verify the transferable ability of the pre-trained networks. Furthermore, our method is applied to low-resolution in-vivo neonatal brain MR images so as demonstrates the qualitative performance.

Chapter 3 introduces an approach to simultaneous super-resolution and segmentation using a generative adversarial network. Generative adversarial networks have been investigated to

estimate realistic super-resolved images and efficient semantic segmentation. However, super-resolution and segmentation are usually processed separately. Firsly, an end-to-end generative adversarial network for simultaneous high-resolution reconstruction and segmentation of brain MRI data is proposed. This joint approach is first assessed on the simulated low-resolution images of the high-resolution neonatal dataset. Then, the learned model is used to enhance and segment real clinical low-resolution images.

Chapter 4 introduces cross-modal medical image synthesis. Two main approaches of medical image synthesis are summed up relied on the property of training dataset: paired and unpaired images. Next, two approaches for brain MRI synthesis are proposed. The first approach applies the most performing convolutional neural networks in Chapter 2 for paired dataset. The second approach lies on the application of generative adversarial networks for unpaired image synthesis.

Chapter 5 concludes the thesis and draws future works.

Appendix A brings up a brief introduction of deep learning. This part consists of the definition of a neural network, the different architectures of neural networks such as convolutional neural networks, activation functions, residual networks and densely connected networks. Next, an application of convolutional neural networks to style transfer and generative adversarial networks are described in detail. The last section introduces the optimization methods of neural networks.

Chapter 2

Brain MRI super-resolution using 3D convolutional neural networks

Contents

2.1 Introduction to single image super-resolution

If you are a fan of fiction films, you sometimes watch the scenes in which the main characters use a computer to verify a video surveillance to track a crime and then they say: *"Hold on. Use an enhancement program. Zoom in right here !"*. Welcome to the *"Let's enhance !"* club. It is no longer with fiction but a very active research areas nowadays: super-resolution reconstruction. Many methods have been proposed for super-resolution [Borman and Stevenson, 1998, Park et al., 2003] since the first work by [Tsai and Huang, 1984].

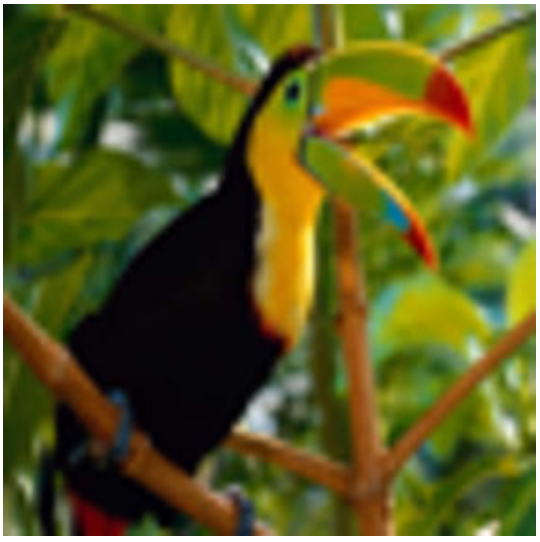
A digital image is composed of elements called pixels. Image spatial resolution, which refers to line pairs per unit distance or pixels (dots) per unit distance, describes the details contained in an image [Gonzalez and Woods, 2006]. For example, a two-dimensional (2D) image with the resolution of 0.1×0.1 mm^2 has 5 line pairs per unit distance (mm) for each direction. High-resolution image can improve the quality of image for human interpretation and machine perception due to the representation of more details. However, imaging acquisition device, which consists of imaging sensors, or imaging acquisition procedure (e.g. the purpose of users) can limit the image resolution. Theoretically, the higher density of the sensors in a digital imaging device may induce higher resolution image. In fact, it is not easy to increase the number of the sensors on a fixed area of the device because of the increase in cost of products and the limitations of current integrated circuit. The post-processing approaches as super-resolution (SR) can overcome physical constraints and also improve the image resolution.

Super-resolution is the process of estimating high-resolution (HR) images from one or several low-resolution (LR) images. The unknown HR image can be reconstructed by multi-image super-resolution methods using several interrelated LR images involved with a determined equation set (e.g. linear constraints) [Milanfar, 2010]. In this work, we focus on single image super-resolution (SR) that estimates the HR image from one corresponding LR image. A closely related method with SR to address this problem is to use the single-image interpolation [Hou and Andrews, 1978, Thévenaz et al., 2000] as a weighted average of the LR pixels y_j:

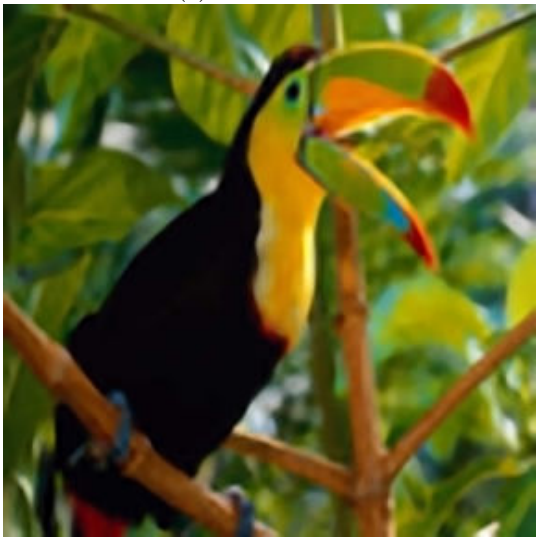
$$\begin{cases} x_i = y_j & i = j \\ x_i = \frac{1}{M} \sum_{i=1}^{M} w_{ij} y_j & otherwise \end{cases} \tag{2.1}$$

where x_i is HR pixels, the weights w are calculated as a function that changes over the distance between the new pixel and M LR ones. The result of the single-image interpolation approach is too smooth because there is no additional information that compensates for the lost of high-frequency components [Milanfar, 2010]. An example of SR results is illustrated in Figure 2.1. The most common up-sampling method, which is image interpolation in Figure 2.1 (b), shows a blurred reconstruction, while the SR start-of-the-art methods such as A+ [Timofte et al., 2014] and SRCNN [Dong et al., 2016a] preserve edges and provide higher visual quality.

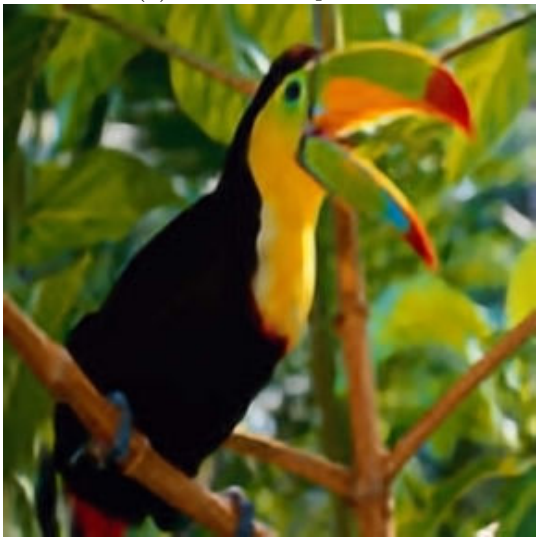
(a) Ground truth



(b) Bicubic interpolation



(c) A+ [Timofte et al., 2014]



(d) SRCNN [Dong et al., 2016a]

FIGURE 2.1: The examples of single SR methods for a LR image of dataset Set5. LR image "bird" is reconstructed using the following methods: (b) bicubic interpolation, (c) A+ [Timofte et al., 2014], (d) SRCNN [Dong et al., 2016a] using the available code from authors.

2.1.1 Image observation model

The image acquisition device can be affected by various factors such as: digital sampling, the relative motion of scene and the camera, optical blur, decimation and noise. Mathematically, let $\mathbf{X}$ and $\mathbf{Y}$ denote the desired HR and the observed LR image, the acquisition process can be modelled as follows:

$$\mathbf{Y} = H\mathbf{X} + N = D_{\downarrow}BF\mathbf{X} + N \tag{2.2}$$

where $\mathbf{Y} \in \mathbb{R}^n$, $\mathbf{X} \in \mathbb{R}^m$, $H \in \mathbb{R}^{m \times n}$ is the observation matrix ($m > n$) and N denotes an additive noise. $D_{\downarrow}$ represents the downsampling operator, B is the blur matrix and F encodes the motion information. B is also called the point spread function (PSF). The purpose of SR methods is to estimate $\mathbf{X}$ from the observations $\mathbf{Y}$. SR is an ill-posed inverse problem where there may be many solutions (i.e. not unique) for one observed input, expressing that the dimension of the observed data always is less than those of the latent HR image. In fact, the observation matrix may be unknown due to the complexity of real imaging systems. Even if the matrix is known, SR is still ill-posed. Thus, many solutions from two main categories: model-based and learning-based methods that can be proposed for this problem. In the next sections, we will introduce some basic techniques proposed in the literature.

2.1.2 Model-based methods

Given an observation model as Equation (2.2), the SR image can be estimated by minimizing a least-square cost function as:

$$\widehat{\mathbf{X}} = \underset{\mathbf{X}}{\operatorname{argmin}}\, \phi(\mathbf{X}, \mathbf{Y}) = \underset{\mathbf{X}}{\operatorname{argmin}}\, \|\mathbf{Y} - H\mathbf{X}\|^2. \tag{2.3}$$

where $\phi(\mathbf{X}, \mathbf{Y})$ denotes the fidelity term. The linear least squares method gives the solution of this equation as:

$$\widehat{\mathbf{X}} = (H^T H)^{-1} H^T \mathbf{Y} \tag{2.4}$$

However, there are many possible solutions since H is ill-conditioned. Based on the observation model, the iterative back-projection (IBP) method [Irani and Peleg, 1991] proposes to calculate the residual between a simulated LR image with the LR observation $\mathbf{Y}$ and then sum the reconstruction error back to the estimated HR image $\hat{\mathbf{X}}$ as:

$$\begin{cases} \hat{\mathbf{X}}^0 = S^{\uparrow}\mathbf{Y} \\ \hat{\mathbf{X}}^{t+1} = \hat{\mathbf{X}}^t + S^{\uparrow}(H\hat{\mathbf{X}}^t - \mathbf{Y}) \end{cases} \tag{2.5}$$

where t is the current iteration, $S^{\uparrow}$ is a upscaling operation (e.g. nearest-neighbor interpolation). The contrast along edges is better recovered than interpolation method. However, the IBP technique which depends the initialized results, is highly sensitive to noise and outliers.

In these cases, the result may contain high frequency artifacts because of ignoring the visual complexity of the ill-posed problem [Milanfar, 2010, Rousseau et al., 2010c]. Thus, this limitations raise the importance of regularizations. A regularizer can be added into the cost function to stabilize the problem as:

$$\widehat{\mathbf{X}} = \underset{\mathbf{X}}{\operatorname{argmin}} \|\mathbf{Y} - H\mathbf{X}\|^2 + \lambda \mathcal{R}(\mathbf{X}) \tag{2.6}$$

where λ is a global weight and $\mathcal{R}(\mathbf{X})$ is a regularization term. The most commonly used method for regularization of ill-posed problems is Tikhonov regularization as:

$$\mathcal{R}(\mathbf{X}) = \sum_p \int_\Omega C_p \mid \mathbf{X}^{(p)} \mid^2 \tag{2.7}$$

where C_p is a positive parameter, Ω is the searching zone and $\mathbf{X}^{(p)}$ denotes the p^{th} order derivative of $\mathbf{X}$. Another regularization is total variation as:

$$\mathcal{R}(\mathbf{X}) = \sum_p \int_\Omega C_p \mid \mathbf{X}^{(p)} \mid \tag{2.8}$$

These approaches assume smooth regions of natural images separated by sharp edges. [Sun et al., 2008] propose gradient profile prior which is fitted by a general exponential generalized Gaussian distribution as:

$$\mathcal{R}(\mathbf{X}) = \frac{\lambda\alpha(\lambda)}{2\sigma\Gamma(\frac{1}{\lambda})} exp\left\{ - \left[\alpha(\lambda) \frac{\mathbf{X}}{\sigma} \right]^\lambda \right\} \tag{2.9}$$

where Γ denotes Gamma function and $\alpha(\lambda) = \sqrt{\Gamma(\frac{3}{\lambda}) / \Gamma(\frac{1}{\lambda})}$ denotes the scaling factor which makes the second moment of the distribution equal to σ^2. λ is the trade-off parameter. [Kim and Kwon, 2010, Tappen et al., 2003] propose natural image prior as Markov random field model:

$$Pr(\mathbf{x} \mid \mathbf{y}) = \prod_s \phi(\mathbf{x}_s) \prod_r \phi(\mathbf{x}_r, \mathbf{y}) \tag{2.10}$$

where $\phi()$ can be a function (e.g. ℓ_2-norm). $\mathbf{x}$ and $\mathbf{y}$ denote the HR and LR patches. Each regularizer assumes a specific image model as data distribution. The minimization of the equation (2.6) with different regularizations on $\mathbf{X}$ usually leads to different solutions. The choice of image prior is crucial for solving the SR problem. In addition, adding prior knowledge on the image solution (such as piecewise smooth image) may lead to unrealistic solution. The work in [Efrat et al., 2013] investigates that an accurate estimate of the PSF is more influenced than a sophisticated prior. Thus, the parameterized prior of the model-based methods is inadequate for the general solution of the SR problem, that requires an approach can learn locally the prior by samples.

2.1.3 Learning-based methods

2.1.3.1 Learning methods for SR

Another approach is to find out the relationship of HR images and corresponding LR images by assuming available external data. Given a set of extracted patch pairs $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{N}$, a method is to learn a direct mapping [Freeman et al., 2002] from the LR patches $\mathbf{y}_i$ to the HR patches $\mathbf{x}_i$, connected by the observation model as Equation (2.2):

$$\mathbf{y}_i = H\mathbf{x}_i + N \tag{2.11}$$

The relationship between these training pairs is denoted as a mapping $\phi(\mathbf{x}_i, \mathbf{y}_i)$. The HR patches $\hat{\mathbf{x}}$ of a testing LR patch $\mathbf{y}$ are reconstructed based on Markov random field by counting the neighbour searching zone and the trained mapping as:

$$P(\hat{\mathbf{x}} \mid \mathbf{y}) = \frac{1}{Z} \prod_{m,n \in \Omega_I} \theta_{mn}(\hat{\mathbf{x}}_m, \hat{\mathbf{x}}_n) \prod_{m \in \Omega_I} \phi(\hat{\mathbf{x}}_m, \mathbf{y}_m) \tag{2.12}$$

where Z is a normalization constant, Ω_I denotes the image space and the node matrix $\theta_{mn}(\hat{\mathbf{x}}_m, \hat{\mathbf{x}}_n)$ is calculated as:

$$\theta_{mn}(\hat{\mathbf{x}}_m, \hat{\mathbf{x}}_n) = \exp\left\{ -\frac{\sum_j (\hat{\mathbf{x}}_{m,j} - \hat{\mathbf{x}}_{n,j})^2}{2\sigma} \right\} \tag{2.13}$$

where j denotes the pixel of patches and σ is a noise parameter. This method is impacted by the patch size. Small patches infer the mapping very fragile but larger patches need large training images. The assumption of two corresponding manifolds of paired patches called neighbour embedding for SR [Chang et al., 2004], which can be used to decrease the amount of training pairs thanks to nearest neighbours search. The method estimates an HR patch $\mathbf{x}$ from k-nearest neighbours Ω_k in the training set of LR testing patches $\mathbf{y}$:

$$\begin{aligned} \hat{\alpha}_i = \underset{\alpha_i}{\operatorname{argmin}} \|\mathbf{y} - \sum_{\mathbf{y}_i \in \Omega_k} \alpha_i \mathbf{y}_i\| \quad s.t. \sum_{\mathbf{y}_i \in \Omega_k} \alpha_i = 1 \\ \hat{\mathbf{x}} = \sum_{\mathbf{y}_i \in \Omega_k} \hat{\alpha}_i \mathbf{x}_i \end{aligned} \tag{2.14}$$

One disadvantage of this method is difficult to choose an effective number of k, for example, a large k can lead to overfitting. An effective method is based on the assumption of an over-complete dictionary and searching for sparse representation which can combine linearly the atoms of the dictionary (called the sparse coding method). For the SR problem, the sparse-coding-based method [Yang et al., 2008] proposes to train dictionaries between HR patches and LR patches. The objective is to find the coefficients α as:

$$\hat{\alpha} = \underset{\alpha}{\operatorname{argmin}} \|\alpha\|_1 \, s.t. \, \|FD_l\alpha - F\mathbf{y}\|_2^2 \leq \epsilon \tag{2.15}$$

where the LR dictionary D_l and the HR dictionary D_h consist of training LR and correspond HR patches respectively. F denotes feature extractions as follows:

$$\begin{cases} F_1 = [-1, 0, 1] \\ F_2 = F_1^T \\ F_3 = [1, 0, -2, 0, 1] \\ F_4 = F_3^T \end{cases} \tag{2.16}$$

where T denotes transpose. After finding the optimal coefficients, the HR patches are estimated as $\hat{\mathbf{x}} = D_h\hat{\alpha}$. An improved version of this work [Yang et al., 2010] proposes to train joint dictionaries D_l and D_h instead of one single constraint on the LR dictionary to enforce the similarity of the representation of image pairs as:

$$\hat{\alpha} = \underset{\alpha}{\operatorname{argmin}} \|\alpha\|_1 \; s.t. \; \|FD_l\alpha - F\mathbf{y}\|_2^2 \leq \epsilon_1, \|PD_h\alpha - \mathbf{m}\|_2^2 \leq \epsilon_2 \tag{2.17}$$

where P denotes the overlapped patch extraction and $\mathbf{m}$ denotes overlapped reconstructed HR values. The sparse representation method for SR is extended in many works by different training approaches and dimensionality reduction [Zeyde et al., 2012], anchored neighborhood embedding [Timofte et al., 2013, 2014] or network-based approximation [Wang et al., 2015]. A sparse-coding-based network for SR is proposed in [Wang et al., 2015] (SCN) by using the learned iterative shrinkage and thresholding algorithm (LISTA)[Gregor and LeCun, 2010]. The method SCN approximates the coefficients by using a multi-layer network Φ as: $\alpha = \Phi(\mathbf{y}, W)$, where W denotes network parameters. The HR dictionary D_h and network parameters is optimized by minimizing the loss function as:

$$\mathcal{L}(W, D_h) = \sum_i \|D_h\Phi(\mathbf{y}_i, W) - \mathbf{x}_i\|_2 \tag{2.18}$$

where the training pairs $(\mathbf{y}_i, \mathbf{x}_i)$. While the sparse-coding method in [Yang et al., 2010] proposes to use first- and second-order derivatives per one image dimension as the feature (i.e. 4 operators). The method in [Gu et al., 2015] decomposes the whole image into several features by learned convolutional filters (more than 4 as in [Yang et al., 2010]) and then uses the sparse representation to match the LR-HR patches of each feature. In order to accelerate the speed of sparse representations (searching the coefficients), anchored neighborhood regression (ANR) [Timofte et al., 2013] proposes to use nearest neighbours of dictionaries. The ANR method replaces ℓ_1-norm by ℓ_2-norm in Equation 2.15 in order to take advantage of a least squares regression as:

$$\min_{\alpha} \|FN_l\alpha - F\mathbf{y}\|_2^2 + \lambda \|\alpha\|_2 \tag{2.19}$$

where N_l, which corresponds to local neighbourhood of LR dictionary D_l, can be computed as in the case neighbour embedding [Chang et al., 2004]. The solution of α is now given by:

$$\alpha = (N_l^T N_l + \lambda I)^{-1} N_l^T F\mathbf{y} \tag{2.20}$$

where I denotes identity matrix and λ is a constant. The testing HR patches is then calculated through the neighborhood N_h of the HR dictionary as:

$$\hat{\mathbf{x}} = N_h\alpha = N_h(N_l^T N_l + \lambda I)^{-1} N_l^T F\mathbf{y} = P_j F\mathbf{y} = P_j \mathbf{y}_F \tag{2.21}$$

Here, P_j is called stored projection matrix and $\mathbf{y}_F$ denotes the input feature. The work of [Timofte et al., 2014] (A+), which develops ANR, proposes to finding K training samples which have the same cluster with the input patch $\mathbf{y}$ instead of the nearest neighbors of LR space, leading to more accurate results and faster estimation. In order to cluster LR patches before the dictionary training, A+ adopts the method of regressions in [Yang and Yang, 2013] as:

$$\hat{\mathbf{x}} = \hat{C}_k \begin{bmatrix} \mathbf{y} \\ 1 \end{bmatrix}, \hat{C}_k = \underset{C_k}{\operatorname{argmin}} \left\| \mathbf{y}_k - C_k \begin{bmatrix} \mathbf{x}_k \\ \mathbf{1} \end{bmatrix} \right\|_2^2 \tag{2.22}$$

where the transformation matrix C_k is found by training patch pairs $(\mathbf{y}_k, \mathbf{x}_k)$ for k^{th} cluster and $\mathbf{1}$ denotes the vector filled with ones. However, each group of patches learns a single regressor where the estimation hardly yields satisfactory results. [Dai et al., 2015] proposes to optimize jointly regressors as:

$$\hat{C}_k, \hat{P}_j = \underset{C_k, P_j}{\operatorname{argmax}} \sum_{k=1}^{K} \sum_{j=1}^{M} C_{k,j} \|\mathbf{x}_k - P_j F\mathbf{y}_k\|^2 \tag{2.23}$$

Then, the expectation–maximization (EM) algorithm is used to optimize the cost function. Several approaches have been investigated to improve the sparse-coding-based method by analysing the sensitivity of principal components such as dictionary size, augmentation of data or combining other techniques [Timofte et al., 2016].

Other family of learning-based approaches relies on clustering the patches by feature extracting and then matching HR-LR patches by random forest [Huang et al., 2015b, Salvador and Perez-Pellitero, 2015, Schulter et al., 2015]. Instead of implicitly figuring out patch regressions based on the dictionaries as Equation (2.17), the method in [Schulter et al., 2015] (RFL) proposes to use a random forest regressor:

$$\hat{\mathbf{x}} = P_j(\mathbf{y}_F)\mathbf{y}_F = T(\mathbf{y}_F) \tag{2.24}$$

where $P_j(\mathbf{y}_F)$ denotes locally linear regressions and T is tree ensembles. By averaging the linear model Tr of each tree j, the estimated HR image is modelled as:

$$\hat{\mathbf{x}} = \frac{1}{J} \sum_{j=1}^{J} Tr^{(j)}(\mathbf{y}_F) \tag{2.25}$$

where r represents the leaf in tree $Tr^{(j)}$ aligned to the feature input $\mathbf{y}_F$. RFL uses 4 filters for the feature extractions as ANR and A+ (shown in Equation (2.16)). The work in [Huang et al., 2015b] finds that there are four main edge-based patterns in which patches are grouped. Then, four random forest can be used to train the linear regressions of each pattern class. For

fast inference and adaptively feature extraction, Local Naive Bayes framework is propose for random-forest-based SR in [Salvador and Perez-Pellitero, 2015]. The optimal regressor from tree $Tr^{(j)}$ for a patch $\hat{\mathbf{x}}$ is estimated by Naive Bayes derivation as:

$$Tr_i^{(j^\star)} = \underset{Tr_i^{(j)}}{\operatorname{argmax}}\, p(Tr_i^{(j)} \mid \hat{\mathbf{x}}) = \underset{Tr_i^{(j)}}{\operatorname{argmax}} \log p(\hat{\mathbf{x}} \mid Tr_i^{(j)}) \tag{2.26}$$

Assuming that we have M features on which the clusters are grouped as $\mathbf{y}_{Fi}(1 \leq i \leq M)$, feature independence results in the log likelihoods as:

$$Tr_i^{(j^\star)} = \underset{Tr_i^{(j)}}{\operatorname{argmax}} \sum_{i}^{\log_2(M)} \log p(\mathbf{y}_{Fi} \mid Tr_i^{(j)}) \tag{2.27}$$

However, these approaches depend crucially on the feature extractions based on pre-defined filters. In addition, because of patch regressions, these methods need optimally global optimization when applying on a testing image, that takes computation costs for each patch reconstruction. In the next section, the methods, which use convolutional neural networks, attempt to learn implicitly necessary features in the networks.

Another approach for SR problem defines matrix H^{-1} as a combination of a restoration matrix $R \in \mathbb{R}^{m \times m}$ and a upscaling interpolation operator $S^{\uparrow} : \mathbb{R}^n \to \mathbb{R}^m$ with respect to the interpolated LR (ILR) image $\mathbf{Z} \in \mathbb{R}^m$ ($\mathbf{Z} = S^{\uparrow}\mathbf{Y}$). Given a set of HR images $\mathbf{X}_i$ and their corresponding LR images $\mathbf{Y}_i$, the restoration operator R can be estimated by minimizing the following loss function:

$$\widehat{R} = \underset{R}{\operatorname{argmin}} \sum_{i}^{k} \|\mathbf{X}_i - R(S^{\uparrow}\mathbf{Y}_i)\|^2 = \underset{R}{\operatorname{argmin}} \sum_{i}^{k} \|\mathbf{X}_i - R(\mathbf{Z}_i)\|^2 \tag{2.28}$$

Once $\widehat{R}$ is estimated, given a LR image $\mathbf{Y}$, the computation of an HR image $\mathbf{X}$ is straightforward: $\mathbf{X} = \widehat{R}(S^{\uparrow}\mathbf{Y})$. In order to model the restoration operation R, the first deep learning method (SRCNN) proposes to use 3 convolutional layers [Dong et al., 2014] for LR feature representation, LR-HR feature matching and image reconstruction. This method does not require any feature descriptions and outperforms the previous hand-crafted methods. The first convolutional layer called R_1 implicitly extracts a set of feature maps for the input LR image as:

$$R_1(\mathbf{Z}) = \max(0, W_1 * \mathbf{Z} + B_1) \tag{2.29}$$

where W_1 and B_1 represent the filters and biases respectively, and "$*$" denotes the convolution operation. A rectified linear unit (ReLU) is applied on the filter responses. The second layer maps these feature maps nonlinearly to HR patch representations:

$$R_2(\mathbf{Z}) = \max(0, W_{L-1} * R_1(\mathbf{Z}) + B_2) \tag{2.30}$$

Finally, the third layer reconstruct the HR image from these patch representations:

$$R_L(\mathbf{Z}) = W_L * R_2(\mathbf{Z}) + B_L \tag{2.31}$$

where L denotes the number of weight layers of networks (i.e. $L = 3$ with SRCNN). In order to optimize the network, SRCNN uses the stochastic gradient descent with momentum algorithm. However, SRCNN attempts to add more than 4 weighted layers but deeper models give lower performance. An illustration of SRCNN is shown in Figure 2.2. After SRCNN, deep learning methods have become a dramatic leap in the SR problem. Several studies have further investigated CNN-based architectures for image SR. An increased depth of the network [Kim et al., 2016a] (VDSR) is proposed up to 20 fully convolutional layers, that rewrites Equation 2.30 as:

$$R_j(\mathbf{Z}) = \max(0, W_j * R_{j-1}(\mathbf{Z}) + B_j) \quad j \in [2, L-1] \tag{2.32}$$

In this case, L is equal to 20. The networks of VDSR are proposed to learn the mapping from the interpolation LR images to the residual between the interpolation LR images and the corresponding HR images as:

$$\widehat{R} = \underset{R}{\operatorname{argmin}} \sum_{i}^{k} \|(\mathbf{X}_i - \mathbf{Z}_i) - R(\mathbf{Z}_i)\|^2 \tag{2.33}$$

Due to residual learning, effective weight initialization and gradient-clipping optimization scheme, VDSR can build more layers than SRCNN, leading to more accurate performance. Recursive neural networks are first proposed in [Kim et al., 2016b] (DRCN). This network replaces the mapping function as the series of convolutional layers in Equation 2.32 by the recursive convolutional layers as:

$$R_2(H) = (g \circ g \circ ...\circ)g(H) = g^D(H) \tag{2.34}$$

where $\circ$ denotes a function composition and g^D denotes the D-fold product of g. Assuming $H_0 = R_1(Z)$, a recurrent relation g as:

$$H_d = g(H_{d-1}) = \max(0, W * H_{d-1} + b) \tag{2.35}$$

DRCN can improve performance by increasing recursion depth, that does not add new parameters for additional convolution layers. A common point of these methods is that they use interpolated images as the input of the networks. The use of interpolation operator consumes of memory (i.e. larger weights storage of each filter per layer). A new layer called sub-pixel layer proposed in [Shi et al., 2016] or a deconvolution layer in [Dong et al., 2016b], inside which the LR image is upscaled, allows the networks independent of interpolation techniques as:

$$\widehat{R} = \underset{R}{\operatorname{argmin}} \sum_{i}^{k} \|\mathbf{X}_i - R(\mathbf{Y}_i)\|^2 \tag{2.36}$$

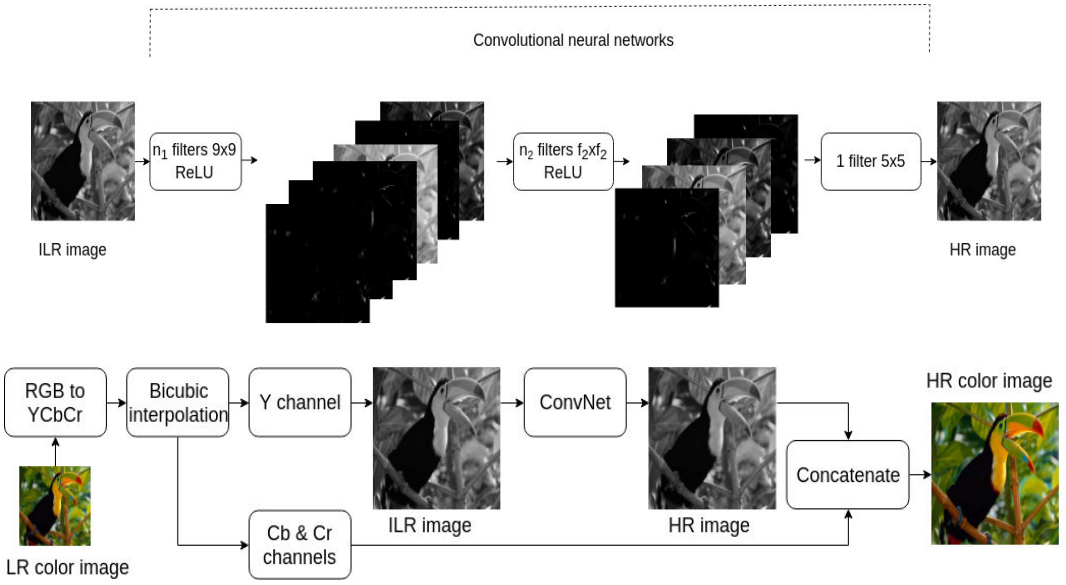


FIGURE 2.2: Pipeline of the method SRCNN [Dong et al., 2016a].

These layers are proposed to be attached at the end of the networks:

$$R_L(\mathbf{Y}) = W_L * S^{\uparrow} R_{L-1}(\mathbf{Y}) \tag{2.37}$$

Instead of learning one scale factor, laplacian pyramid networks in [Lai et al., 2017] propose to train simultaneous several factors through a set of progressive upscaling layers. A network with more than 16 residual blocks (a block consists of two convolutional layers with batch normalization, ReLU and skip connection) is proposed in [Ledig et al., 2017] (SRResnet). The recursive blocks and the residual blocks are then combined in the work of [Tai et al., 2017] to build more layers but still maintain the efficiency. Although, the deeper networks (more than 20 weight layers) such as SRResnet have very accurate quantitative metrics, the methods give less perceptual reconstructions [Ledig et al., 2017]. The investigation of other effective functions instead of mean squared error-based cost functions has been proposed such as ℓ_1-norm loss [Zhao et al., 2017], Charbonnier loss [Lai et al., 2017], perceptual loss [Johnson et al., 2016, Ledig et al., 2017]. The objective function in Equation 2.36 can be rewritten as a ℓ_1-norm:

$$\widehat{R} = \underset{R}{\operatorname{argmin}} \sum_{i}^{k} \|\mathbf{X}_i - R(\mathbf{Y}_i)\| \tag{2.38}$$

or a robust Charbonnier loss function as:

$$\widehat{R} = \underset{R}{\operatorname{argmin}} \sum_{i}^{k} \sqrt{\|\mathbf{X}_i - R(\mathbf{Y}_i)\|^2 + \epsilon_\rho^2} \tag{2.39}$$

where ϵ_ρ is set to $1e-3$. [Johnson et al., 2016] propose to train the restoration network R to generate the output $R(\mathbf{Y})$ which has the perceptual content of the HR image $\mathbf{X}$ based on

the perceptual loss as:

$$\mathcal{L}_{perceptual}(\mathbf{X}, R(\mathbf{Y})) = \sum_k (F_k^l(R(\mathbf{Y})) - F_k^l(\mathbf{X}))^2 \tag{2.40}$$

where F^l is the feature maps of the l^{th} layer of a pre-trained network (e.g. VGG-net [Simonyan and Zisserman, 2014]). Generative adversarial networks [Ledig et al., 2017] (SRGAN) improve the idea of the perceptual loss by adding an adversarial loss. SRGAN consists of two networks: a network called the generator R generates super-resolved images and another discriminates the generated images and the true ones as the discriminator D. The adversarial objective of SRGAN can be described as:

$$\mathcal{L}_{adversarial} = \min_R \max_D \mathbb{E}_{\mathbf{X}\sim\mathbb{P}_\mathbf{X}}[logD(\mathbf{X})] + \mathbb{E}_{\mathbf{Y}\sim\mathbb{P}_\mathbf{Y}}[log(1 - D(R(\mathbf{Y})))] \tag{2.41}$$

The total objective of SRGAN is formulated as the weighted losses:

$$\mathcal{L}_{SRGAN} = \mathcal{L}_{perceptual} + 10^{-3}\mathcal{L}_{adversarial} \tag{2.42}$$

We have reviewed several learning-based image super-resolution algorithms using statistical approaches, sparse coding, random forest and CNNs. One of the main benefits of learning-based methods is the potential for non linear representation between HR and LR pairs. Furthermore, the methods are capable to learn a substantial amount of regressors which observes and generalizes relationships inside the data. However, not all the information in the training set may be relevant to observed LR images. The feature extractions are crucial to some learning-based methods such as sparse coding or random forest. Since the first success of SRCNN, the number of convolutional neural networks based methods are numerous, thanks to the ability of feature learning inside the networks and the support of GPU computational power (that reduces the training time). On the other hand, the CNN-based techniques are lacking of mathematical theories because they can optimize considered non-convex functions. But we can not deny that SR methods using CNNs work extremely well.

2.1.3.2 Blind super-resolution

Most methods assume a known PSF of imaging systems for the observation model. Then, the models are trained based on this assumption. However, the pre-trained model significantly decreases the quality of results when applying to real LR images acquired with a different PSF. Blind super-resolution methods attempt to estimate the appropriate PSF of the observed LR image instead of using a pre-defined kernel. In order to estimate the PSF, these approaches assume stochastic reconstruction steps initialized by a random PSF for an optimal reconstruction. Every patch $\mathbf{y}_i$ extracted from the LR image ($i = 1, ..., M$) can be expressed from the observation model as:

$$\mathbf{y}_i = H\mathbf{x}_i + N \tag{2.43}$$

where $\mathbf{x}_i$ denotes the patches of the HR image, H is the observation matrix. Some methods [He et al., 2009, Wang et al., 2005] propose to simultaneously estimate the HR image and the PSF parameter using the a joint maximum a posteriori (MAP) of probabilistic combination models p as:

$$p(\mathbf{x}_i, h \mid \mathbf{y}_i) \propto p(\mathbf{y}_i \mid \mathbf{x}_i, h)p(\mathbf{x}_i)p(h) \tag{2.44}$$

where $p(\mathbf{x}_i)$ and $p(h)$ are prior terms and $p(\mathbf{y}_i \mid \mathbf{x}_i, h)$ is the data likelihood. Assuming that the term N of Equation 2.43 stands for a white Gaussian noise with a zero-mean and the standard deviation of σ, the data likelihood can be expressed by image formation model as:

$$p(\mathbf{y}_i \mid \mathbf{x}_i, h) = \prod_{i=1}^{M} \frac{1}{N} \sum_{i=1}^{M} \exp\left\{ -\frac{\|\mathbf{y}_i - H(h)_i \mathbf{x}_i\|^2}{2\sigma^2} \right\} \tag{2.45}$$

Here, the HR image prior $p(\mathbf{x}_i)$ can be computed by the learning methods [Freeman et al., 2002] and the PSF prior $p(h)$ can be assumed to be a uniform distribution over a pre-defined range because of no prior knowledge on it [Wang et al., 2005]. $H(h)_i$ is the estimated observation model during the generation of $\mathbf{y}_i$. However, these assumptions may lead to inaccurate estimation [Michaeli and Irani, 2013] because the methods attempt to estimate simultaneously the prior of the HR image $\mathbf{x}_i$ and the kernel h. Instead, [Michaeli and Irani, 2013] only computes the MAP estimate of the kernel h:

$$\begin{aligned} \hat{h} &= \underset{h}{\operatorname{argmax}}\, p(h) \prod_{i=1}^{M} p(\mathbf{y}_i \mid h) \\ &= \underset{h}{\operatorname{argmax}}\, p(h) \prod_{i=1}^{M} \int_{\mathbf{x}_i} p(\mathbf{y}_i \mid \mathbf{x}_i, h) p(\mathbf{x}_i) d\mathbf{x}_i \end{aligned} \tag{2.46}$$

where $p(\mathbf{x}_i)$ is a prior term. Similarly, we can express the estimation as :

$$\hat{h} = \underset{h}{\operatorname{argmax}}\, p(h) \prod_{i=1}^{M} \int_{\mathbf{x}_i} \exp\left\{ -\frac{\|\mathbf{y}_i - H\mathbf{x}_i\|^2}{2\sigma^2} \right\} p(\mathbf{x}_i) d\mathbf{x}_i \tag{2.47}$$

Given N HR training patches $\mathbf{x}_i$, the prior term can be approximated by empirical mean as:

$$\hat{h} = \underset{h}{\operatorname{argmax}}\, p(h) \prod_{i=1}^{M} \frac{1}{N} \sum_{j=1}^{N} \exp\left\{ -\frac{\|\mathbf{y}_i - H\mathbf{x}_j\|^2}{2\sigma^2} \right\} \tag{2.48}$$

where $p(h)$ is a nonparametric prior. This is in contrast to [Wang et al., 2005] which assumes a parametric prior. [Michaeli and Irani, 2013] emphasize that the term $H\mathbf{x}_j$ can be equivalently written as $\mathbf{X}_j h$ because of the dependence of Equation (2.48) on h, where $\mathbf{X}_j$ is a matrix corresponding to convolution with $\mathbf{x}_i$ and a down-sampling operator. Equation (2.48) can be solved by taking the *log* as:

$$\hat{h} = \underset{h}{\operatorname{argmin}} \frac{1}{2}\|Ch\| - \sum_{i=1}^{M} log \left(\sum_{j=1}^{N} \exp\left\{ -\frac{\|\mathbf{y}_i - \mathbf{X}_j h\|^2}{2\sigma^2} \right\} \right) \tag{2.49}$$

where C can be a chosen matrix to penalize for non-smooth kernels. The blind SR methods can be used to approximate the real PSF of observed LR images using principled MAP estimations. The use of blind SR based on learning methods is very potential for real applications such as enhancing the historical image. However, the current algorithms for blind SR are based on several assumptions, that may reduce the generalization of the observation model.

2.1.3.3 Zero-shot learning

If external training datasets are not available, one approach called zero-shot learning proposes to exploit the similarity of patches inside the image. Assuming that the observation model in Equation 2.2 with noise free as:

$$\mathbf{Y} = D_{\downarrow} B \mathbf{X} \tag{2.50}$$

The method in [Glasner et al., 2009] attempts to find the HR of a LR image by exploiting cross-scale patch redundancy called internal examples. A set of several downscaled versions from the LR one $\mathbf{Y}$ can be generated as $I_{-i} = D_{\downarrow}^{\times -i} B\mathbf{Y}$. The strategy first finds the nearest neighbours of a patch $\mathbf{y}$ in the LR image from several downscaled versions and then copies to upscaled versions I_i. Then, the method combines these upscaled versions to reconstruct the HR image by the multi-image methods as [Milanfar, 2010]. Instead of 2D transformation as in [Glasner et al., 2009] (i.e. translation), [Huang et al., 2015a] propose a transform matrix to find the self-similarity between internal recurrence of patches inside the testing image. The method in [Shocher et al., 2018] exploits the kernel estimation in [Michaeli and Irani, 2013] and the powerful representation of CNN-based technique for training the internal example patches by assuming the testing LR image as HR patches and its lower-resolution versions as LR cross-scale patches.

Zero-shot learning is used to overcome difficulties where the external dataset is lacking. In addition, these methods are very useful for LR images which contain redundant patches. However, since one LR image patch can construct several HR image patches, zero-shot learning may ignore details which are missed in the testing image.

2.1.4 Applications of super-resolution in medical imaging

Previously, the techniques for 2D natural images have been reviewed. However, photo-realistic images can not model specific 3D organs or the human body. In addition, medical imaging modalities are very diverse. Each modality has specific features which can be used to medical image analysis. Thus, SR methods for specific medical imaging are also studied. Besides, a set of different 2D images could only represent the slices of 3D architectures, not connections in 3D space, that raises the need of 3D models for 3D medical images. The study focused on medical imaging supports better for other practical applications.

Moreover, higher resolution medical image is the key to early detection of abnormalities or pathologies. One of the tasks of medical imaging is to increase and to extent the possible resolution so as achieve true isotropic 3-D images. In practice, the maximal sampling frequency of the imaging device detectors limits the captured range of radio frequencies from the imaged object. A solution to increase resolution is to reduce detectors size, however, this increases the noise, thus reduces SNR. Increasing image resolution through super-resolution is a key to better understanding of the anatomy [Greenspan, 2008]. Medical image SR can be used to improve the performance of image segmentation and image registration methods. A better quality of an image can result more accurate segmentation and registration. Previous works have shown that applying super-resolution techniques leads to more accurate segmentation maps of brain MRI data [Jog et al., 2016, Rueda et al., 2013] or cardiac data [Oktay et al., 2016].

The use of SR techniques has been studied in the context of medical analysis, specially of brain images: anatomical MRI [Luo et al., 2017, Manjón et al., 2010a,b, Rousseau, 2008, Rousseau et al., 2010a,b, Rueda et al., 2013, Shi et al., 2015], diffusion MRI [Fogtmann et al., 2014, Poot et al., 2013, Scherrer et al., 2012, Steenkiste et al., 2016], spectroscopy MRI [Jain et al., 2017], quantitative T_1 mapping [Ramos-Llordén et al., 2017, Van Steenkiste et al., 2017], fusion of orthogonal scans of moving subjects [Gholipour et al., 2010, Jia et al., 2017, Kainz et al., 2015, Rousseau et al., 2010c]. In the next sections, we will focus on two families of medical image SR: model-based methods and learning-based methods.

2.1.4.1 Applications of model-based methods

The non-local mean upsampling [Manjón et al., 2010b] (NMU) method performs first high quality reconstructed image via the iteration patch-based filtering as:

$$\hat{x}^{t+1} = \frac{1}{C} \sum_{\forall k \in \Omega} w(\hat{x}^t, \hat{x}_k^t) \hat{x}_k^t \tag{2.51}$$

where $\hat{x}^t$ is the voxel of the reconstructed HR image at the current iteration t, C is a constant and Ω is the searching zone. The initialized image is supposed as $\mathbf{X}^0 = S \uparrow \mathbf{Y}$. The weighted coefficient w is calculated based on the non-local mean (NLM) filter [Coupé et al., 2008] as:

$$w(\hat{x}^t, \hat{x}_k^t) = \begin{cases} e^{\frac{|\hat{\mathbf{x}}^t - \hat{\mathbf{x}}_k^t|^2}{h^2}} & if \mid \mu^t - \mu_k^t \mid < 3h/\sqrt{N} \\ 0 & otherwise \end{cases} \tag{2.52}$$

where μ is the average of 3D patches $\mathbf{x}$ around the voxel x, h denotes a filtering parameter and N is the number of voxel in the 3D patch. The second step of NMU exploits the IBP method [Irani and Peleg, 1991] for ensuring consistency between the observation model and the estimated high resolution.

The medical image SR problem could be also solved by minimizing a objective function with a regularization term as Equation 2.6. The objective function with a ℓ_2-norm regularization, which is proposed in [Gholipour et al., 2010, Rousseau et al., 2010c], can be written as:

$$\widehat{\mathbf{X}} = \underset{\mathbf{X}}{\operatorname{argmin}} \|\mathbf{Y} - H\mathbf{X}\|^2 + \lambda \|C\mathbf{X}\|_2^2 \tag{2.53}$$

where C is a positive definite matrix. A combination of low-rank regularization [Liu et al., 2013] and total-variation regularization [Rudin et al., 1992] proposed in [Shi et al., 2015] (LRTV) transforms the SR problem as:

$$\widehat{\mathbf{X}} = \underset{\mathbf{X}}{\operatorname{argmin}} \|\mathbf{Y} - H\mathbf{X}\|^2 + \lambda_{rank} Rank(\mathbf{X}) + \lambda_{tv} TV(\mathbf{X}) \tag{2.54}$$

where *Rank* is the weighted sum of trace norms of all slices along each dimension of an image and *TV* (total-variation) denotes the integral of the absolute gradients of data. The regularizer *Rank* takes advantages of the similarity between the slices in different directions, that can not happen in the 2D image cases. However, these priors assume that the image is too smooth leading to lack of the details of the true image representation.

2.1.4.2 Applications of learning-based methods

The learning-based methods can not only find implicitly the parameters of prior energy function via examples but also define a specific regularization expression. The work in [Rousseau and Studholme, 2013] extends the NMU method for capturing more information of a training dataset. Given a training dataset which consists of paired HR-LR images $\mathbb{D} = \{(\mathbf{X}_i, \mathbf{Y}_i) \mid i = 1, ..., N\}$, we can reconstruct the HR image $\mathbf{X}$ of the testing LR image $\mathbf{Y}$ as

$$\hat{\mathbf{X}}(x) = \frac{\sum_{i=1}^{N} \sum_{\forall k \in \Omega} w_i(x, x_k) \mathbf{X}_i(x_k)}{\sum_{i=1}^{N} \sum_{\forall k \in \Omega} w_i(x, x_k)} \tag{2.55}$$

where x is the current voxel with the neighbour searching zone Ω and $w_i(x, x_k)$, which denotes the weighted coefficients, is calculated by the similarity between $\mathbf{Y}$ and each LR sample from the external set as:

$$w_i(x, x_k) = \begin{cases} e^{\frac{|\mathbf{x} - \mathbf{x}_{k,i}|^2}{h^2}} & if \mid \mu - \mu_k \mid < 3h/\sqrt{N} \\ 0 & otherwise \end{cases} \tag{2.56}$$

where the parameters of this equation are similar to Equation 2.52. This method also needs a correction step in order to improve the robustness. Instead of using IBP as in [Manjón et al., 2010b], the 3D patches $\mathbf{x}$ of the HR reconstruction are calibrated by the HR samples as:

$$\hat{\mathbf{x}}(x) = \sum_{i=1}^{N} \sum_{\forall k \in \Omega} w_i(x, x_k) \mathbf{x}_i(x_k) \tag{2.57}$$

The extension of sparse representation methods [Yang et al., 2010, Zeyde et al., 2012] is proposed in [Rueda et al., 2013] by using multi-scale Sobel filters. In this work, the authors demonstrated the importance of 3D feature detectors within brain MRI data. The filter set, which is proposed to analyse multi-scale edges of interpolated testing LR images, consists of 2 high-frequency filters with the patch size of 3 and 5 for each direction. Then, the HR image is reconstructed by finding in the LR-HR sparse dictionaries and is then corrected by the IBP method as [Irani and Peleg, 1991].

Recently, 3D convolutional neural networks for MRI SR, which have been investigated in [Pham et al., 2017a], learn the feature representation automatically inside the networks. We will discuss this approach in the Section 2.2. Later, [Chen et al., 2018b] proposed a 3D version of densely connected networks (DenseNet) [Huang et al., 2017a] for brain MRI SR. Before DenseNet, the residual networks in [He et al., 2016a] (ResNet) achieved the most performance in image classification. ResNet can build up to 1000 convolution layers thanks to the residual blocks [He et al., 2016b], that is impossible to the previous networks [Simonyan and Zisserman, 2014, Szegedy et al., 2015]. However, ResNet takes a lot of memory training. Densely connected networks [Huang et al., 2017a] can achieve performance as good as deep networks (e.g. ResNet [He et al., 2016a]) but reduces memory training thanks to feature concatenations through all layers. Assuming that the external dataset is not available, inspired by the work of [Jog et al., 2016], [Zhao et al., 2018] investigated self super-resolution for MRI using enhanced deep residual networks [Lim et al., 2017]. [Zhao et al., 2018] relies on the fact that a LR anisotropic 3D image has a in-plane high resolution (e.g. axial slice). Then, the LR image is interpolated to generate an interpolated isotropic image as a HR reference image. A simulated LR image is then generated from the HR reference. The deep network in [Lim et al., 2017] is trained with the patches of simulated pairs and finally applied to the original LR image.

2.1.5 Evaluation

For quantitative comparison, the peak signal to noise ratio (PSNR) in decibels (dB) and Structural Similarity Index (SSIM) [Wang et al., 2004] are commonly used to evaluate the performance of image reconstruction algorithms. Given a dynamic range d, the PSNR is defined as:

$$PSNR = 10 \log_{10}(\frac{d^2}{MSE}) \tag{2.58}$$

where the mean squared error (MSE) is defined as:

$$MSE = \sum_{i \epsilon \Omega} (\mathbf{X}(i) - \hat{\mathbf{X}}(i))^2 \tag{2.59}$$

where $\hat{\mathbf{X}}$ is the reconstructed image with respected to the ground truth $\mathbf{X}$, Ω is the number of pixels or voxels of images.

SSIM is used for measuring the image quality based on perceived similarity. SSIM is calculated as:

$$SSIM = \frac{(2\mu_X \mu_{\hat{X}} + c_1)(2\sigma_{X\hat{X}} + c_2)}{(\mu_X^2 + \mu_{\hat{X}}^2 + c_1)(\sigma_X^2 + \sigma_{\hat{X}}^2 + c_2)} \tag{2.60}$$

where μ_X, σ_X^2 are respectively the average and the variance of window X of image $\mathbf{X}$, similarly for $\hat{\mathbf{X}}$ and $\sigma_{X\hat{X}}$ denotes the covariance of these windows. c_1 and c_2 are two constants.

However, in some cases, a higher PSNR or SSIM does not indicate that the reconstruction is of higher quality because they do not correlate with human assessment of visual quality [Johnson et al., 2016, Ledig et al., 2017, Wang and Bovik, 2009]. PSNR and SSIM rely only differences between pixels which may not describe the high-level human visual perception via feature representation. Thus, when comparing methods, the need of qualitative results should be shown to have a general assessment.

2.1.6 Discussion

Major advances in the domains of computer vision indicated the ability of SR methods. The most popular approach is based on solving the observation model. In order to constraint the ill-posedness of model-based methods, the adding prior can bound the conditions of the solution. The capacity of non linear representations which is used in the learning-based methods helps to capture the relationships of low-resolution images and high-resolution ones. In contrast to other learning methods which strictly depend on the feature extractions, convolutional neural networks with implicitly feature representation have become the state-of-the-art models for SR. However, SR algorithms may face the fact that the point spread function of observation model is not always ideally. Thus, several methods attempt to solve SR which does not assume a fixed blurring function (blind SR). In addition, many techniques exploit the redundant information of internal patches to increase the solution of low-resolution images.

The methods of two dimensional natural image SR, which have also mentioned, can be expanded to 3D images. However, the medical image SR can not be viewed inseparable from 2D photo-like techniques. In application of SR in medical imaging, we have introduced two main categories: model-based and learning-based methods. The techniques based on the observation models, which depend on the assumptions of image priors, do not need to collect other external data. However, they can lead to too smooth results due to crucial prior. In order to exploit the missing information which can provided by a training set, the learning-based methods can be used. Several learning-based techniques require feature extractions that can

reduce the information of image representation. In addition, the better learning algorithms will help us to better performance, augment the capacity of feature learning storage or faster convergence. Convolutional neural networks is one of the methods which do not depend on feature extraction because they can learn representation filters implicitly inside their layers.

2.2 Learning-based single super-resolution using convolutional neural networks

CNN architectures have become the state-of-the-art for image SR. However, due to the variety of the proposed methods and the high number of parameters for the networks architecture design, it is currently difficult to identify the key elements of CNN architecture to achieve good performance for image SR and assess their applicability in the context of 3D brain MRI. In addition the extension of CNN architectures to 3D images, taking into account floating and possibly anisotropic scaling factors may be of interest to address the wide range of possible clinical acquisition settings, whereas classical CNN architectures only address a predefined (integer) scaling factor. The availability of multimodal imaging setting also questions the ability of CNN architectures to benefit from such multimodal data to improve the SR of a given modality.

First of all, our work verifies the need of fitting data and network parameters for 3D brain MRI. Then, this work presents a comprehensive review of deep 3D convolutional neural networks, and associated key elements, for brain MRI SR. Following [Timofte et al., 2016], who have experimentally showed several ways to improve SR techniques from a baseline architecture, we study the impact of eight key elements on the performance of convolutional neural networks for 3D brain MRI SR. We demonstrate empirically that residual learning associated with appropriate optimization methods can significantly reduce the time of the training step and fast convergence can be achieved in 3D SR context. Overall, we report better performance when learning deeper fully 3D convolution neural networks and using larger filters. Interestingly, we demonstrate that a single network can handle multiple arbitrary scale factors efficiently, for example, from $2 \times 2 \times 2$ mm to $2 \times 2 \times 1$ mm or $1 \times 1 \times 1$ mm, by learning multiscale residuals from spline-interpolated image. We also report significant improvement using a multimodal architecture, where a HR reference image can guide the CNN-based SR of a given MRI volume.

Recall that single image SR is a typically ill-posed inverse problem that can be stated according to the following linear formulation:

$$\mathbf{Y} = H\mathbf{X} + N = D_{\downarrow}B\mathbf{X} + N \tag{2.61}$$

where $\mathbf{Y} \in \mathbb{R}^n$ is the LR observed image, $\mathbf{X} \in \mathbb{R}^m$ is the HR image, $H \in \mathbb{R}^{m \times n}$ is the observation matrix ($m > n$) and N denotes an additive noise. $D_{\downarrow}$ represents the downsampling operator and B is the PSF. In a learning-based context where a set of image pairs $(\mathbf{X}_i, \mathbf{Y}_i)$

is available, the objective is to learn the mapping H^{-1} from the LR images $\mathbf{Y}_i$ to the HR images $\mathbf{X}_i$, leading to the following formulation:

$$\widehat{\mathbf{X}} = \arg\min_{\mathbf{X}} \|\mathbf{X} - H^{-1}\mathbf{Y}\|^2. \tag{2.62}$$

In this setting, the matrix H^{-1} can be modeled as a combination of a restoration matrix $F \in \mathbb{R}^{m\times m}$ and an upscaling interpolation operator $S^{\uparrow} : \mathbb{R}^n \to \mathbb{R}^m$. Given a set of K HR images $\mathbf{X}_i$ and their corresponding LR images $\mathbf{Y}_i$, the restoration operator F can be estimated as follows:

$$\widehat{F} = \arg\min_{F} \sum_{i=1}^{K} \|\mathbf{X}_i - F(S^{\uparrow}\mathbf{Y}_i)\|^2 = \arg\min_{F} \sum_{i=1}^{K} \|\mathbf{X}_i - F(\mathbf{Z}_i)\|^2 \tag{2.63}$$

where $\mathbf{Z} \in \mathbb{R}^m$ is the interpolated LR (ILR) version of $\mathbf{Y}$ (*i.e.* $\mathbf{Z} = S^{\uparrow}\mathbf{Y}$). F is then a mapping from the ILR image space to the HR image space.

2.2.1 Methodology

2.2.1.1 Restoration by convolutional neural networks : 2D or 3D models for 3D data ?

The restoration matrix F corresponds to the mapping from $\mathbf{Z}$ to $\mathbf{X}$. In SRCNN [Dong et al., 2016a], this mapping is decomposed into three operations, described as follows:

$$\begin{cases} F_1(\mathbf{Z}) = \max(0, W_1 * \mathbf{Z} + B_1) \\ F_2(\mathbf{Z}) = \max(0, W_2 * F_1(\mathbf{Z}) + B_2) \\ F_3(\mathbf{Z}) = W_3 * F_2(\mathbf{Z}) + B_3 \end{cases} \tag{2.64}$$

where:

- W_i and B_i are the convolution parameters to learn, where $i \in \{1, 2, 3\}$. W_i corresponds to n_i convolution filters of support $c \times f_i \times f_i \times f_i$, where c is the number of channels in the input of layer i, f_i and n_i are respectively the spatial size of the filters and the number of filters of layer i,
- $\max(0, \cdot)$ refers to a ReLU applied to the filter responses.

Each of these operations is designed using one layer of the neural network. The first step, called F_1, extracts overlapping patches of the LR image and computes a set of feature maps. F_1 is similar to a popular strategy in image restoration by representing patches by a set of pre-trained bases (such as PCA or DCT). In SRCNN, this step is performed by convolving the image by a set of learned filters. The second operation, F_2, which is mathematically very close to F_1, is a non-linear mapping from the LR feature maps to HR feature maps. Finally, the third operation, F_3, is a convolutional layer corresponding to the image reconstruction.

W_3 can be seen as the projection of HR feature maps onto the image domain and then patches averaging.

SRCNN has been originally designed for 2D natural image processing. In [Dong et al., 2016a], W_1, W_2 and W_3 consist of n_1 filters with 2D patch size $f_1 \times f_1$, n_2 filters with patch size $f_2 \times f_2$ and one filter with patch size $f_3 \times f_3$ respectively. In order to apply this restoration operator called F_{2D}, we propose first a straightforward strategy consisting in averaging restored versions of 3D ILR images $\mathbf{Z}_{3D}$ for each direction to estimate a 3D HR image $\widehat{\mathbf{X}}_{3D}$:

$$\widehat{\mathbf{X}}_{3D} = \overline{F_{2D}^{axial}(\mathbf{Z}_{3D}) + F_{2D}^{coronal}(\mathbf{Z}_{3D}) + F_{2D}^{sagittal}(\mathbf{Z}_{3D})} \tag{2.65}$$

Using this strategy, it is possible to apply the model learned with natural images [Dong et al., 2016a] (called here SRCNNF-Nat). In addition, a network is trained with a dedicated learning image dataset (called SRCNNF-Brain).

In addition, we investigate the use of a 3D network which consists of n_1 filters with voxel size $f_1 \times f_1 \times f_1$, n_2 filters with voxel size $f_2 \times f_2 \times f_2$ and one filter with voxel size $f_3 \times f_3 \times f_3$ in Section 2.2.2.2. The 3D HR image is then computed as follows: $\widehat{\mathbf{X}}_{3D} = F_{3D}(\mathbf{Z}_{3D})$.

2.2.1.2 Restoration by 3D residual-learning convolutional neural networks

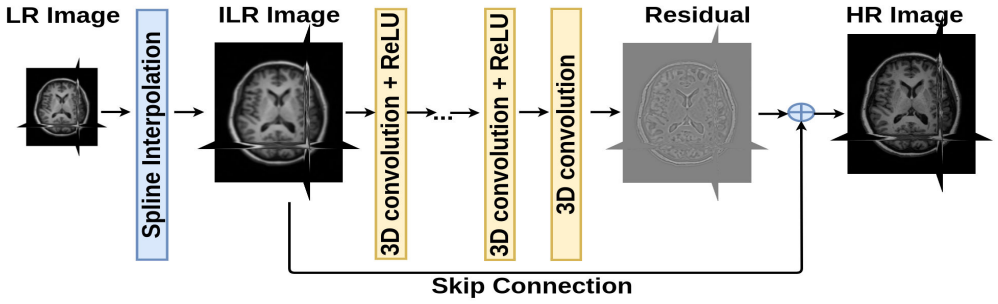


FIGURE 2.3: 3D residual-learning convolutional neural networks for single brain MRI super-resolution.

Instead of learning the mapping directly from the LR space to the HR one, it might be easier to estimate a mapping from the LR space to the missing high-frequency components, also called the residual between HR and LR data: $\mathbf{R} = \mathbf{X} - \mathbf{Z}$ or equivalently $\mathbf{X} = \mathbf{Z} + \mathbf{R}$. This approach can be modeled by a skip connection in the network. In such a residual-based modeling, one typically assumes that $\mathbf{R}$ is a function of $\mathbf{Z}$. The computation of HR data is then expressed as follows: $\mathbf{X} = \mathbf{Z} + F(\mathbf{Z})$ where F can be learned using the following equation:

$$\widehat{F} = \arg\min_{F} \sum_{i=1}^{K} \|(\mathbf{X}_i - \mathbf{Z}_i) - F(\mathbf{Z}_i)\|^2. \tag{2.66}$$

Following [Kim et al., 2016a], mapping F from $\mathbf{Z}$ to $(\mathbf{X} - \mathbf{Z})$ is decomposed into nonlinear operations corresponding to the combination of convolution-based and rectified linear unit (ReLU) layers. The baseline deeper architecture used in this work can be described as follows:

$$\begin{cases} F_1(\mathbf{Z}) = \max(0, W_1 * \mathbf{Z} + B_1) \\ F_i(\mathbf{Z}) = \max(0, W_i * F_{i-1}(\mathbf{Z}) + B_i) \quad for \quad 1 < i < L \\ F_L(\mathbf{Z}) = W_L * F_{L-1}(\mathbf{Z}) + B_L \end{cases} \tag{2.67}$$

where L is the number of layers. This network architecture is depicted in Figure 2.3. Please note that, for instance, the SRCNN model [Dong et al., 2016a] corresponds to a specific parameterization of this baseline architecture ($f_1 = 9$, $f_2 = 1$, $f_3 = 5$, $n_1 = 64$, $n_2 = 32$ and with no skip connection).

2.2.2 Experimental setting

2.2.2.1 MRI dataset and LR simulation

To evaluate SR performances of CNN-based architectures, we have used two MRI datasets: the Kirby 21 dataset and the NAMIC Brain Multimodality dataset.

The Kirby 21 dataset [Landman et al., 2011] consists of MRI scans of twenty-one healthy volunteers with no history of neurological conditions. Magnetization prepared gradient echo (MPRAGE, T1-weighted) scans were acquired using a 3-T MR scanner (Achieva, Philips Healthcare, The Netherlands) with a $1.0 \times 1.0 \times 1.2$ mm^3 resolution over an FOV of $240 \times 204 \times 256$ mm acquired in the sagittal plane. Flair data were acquired using $1.1 \times 1.1 \times 1.1$ mm^3 resolution over an FOV of $242 \times 180 \times 200$ mm acquired in the sagittal plane. The T2-weighted volumes were acquired using a 3D multi-shot turbo-spin echo (TSE) with a TSE factor of 100 with over an FOV of $200 \times 242 \times 180$ mm including a sagittal slice thickness of 1 mm.

MR images of NAMIC Brain Multimodality [1] dataset have been acquired using a 3T GE at BWH in Boston, MA. An 8 Channel coil was used in order to perform parallel imaging using ASSET (Array Spatial Sensitivity Encoding techniques, GE) with a SENSE-factor (speed-up) of 2. The structural MRI acquisition protocol included two MRI pulse sequences. The first results in contiguous spoiled gradient-recalled acquisition (fastSPGR) with the following parameters; TR=7.4ms, TE=3ms, TI=600, 10 degree flip angle, 25.6cm^2 field of view, matrix=256×256. The voxel dimensions are $1 \times 1 \times 1mm^3$. The second- XETA (eXtended Echo Train Acquisition) produces a series of contiguous T2-weighted images (TR=2500ms, TE=80ms, 25.6 cm^2 field of view, 1 mm slice thickness). Voxel dimensions are $1 \times 1 \times 1mm^3$.

As in [Shi et al., 2015] and [Rueda et al., 2013], LR images have been generated from a Gaussian blur and a down-sampling by isotropic scaling factors. In the training phase, a set of patches of training images is randomly extracted. The training dataset comprises 10 subjects (3200 patches $25 \times 25 \times 25$ per subject randomly sampled) and the testing dataset is composed of 5 subjects. During the testing step, the network is applied on the whole images. The peak signal-to-noise ratio (PSNR) in decibels (dB) is used to evaluate the SR results with

[1] NAMIC : http://hdl.handle.net/1926/1687

respect to the original HR images. No denoising or bias correction algorithms were applied to the data. Image intensity has been normalized between 0 and 1. The following figures are drawn based on the average PSNR over all test images.

2.2.2.2 Results with respect to 2D and 3D networks

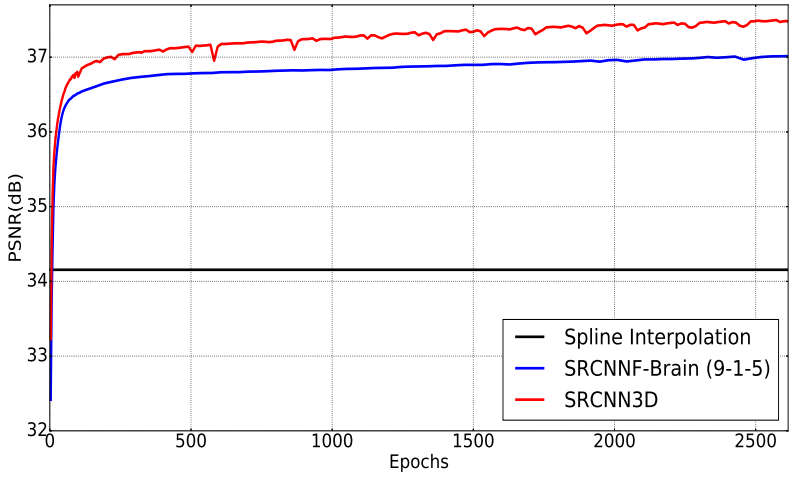


FIGURE 2.4: The evolution of the mean PSNR of SRCNNF-Brain and SRCNN3D with respect to the number of epochs ©[2017] IEEE.

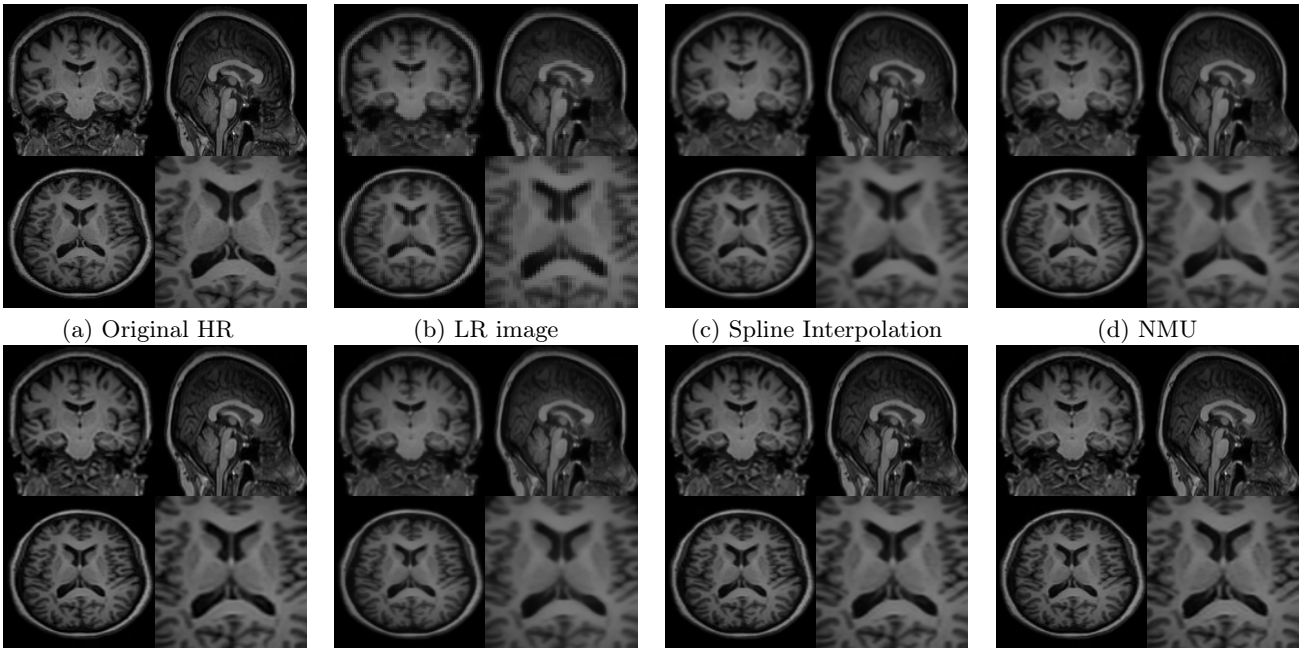
First, we studied the impact of the number of epochs used for training for both SRCNNF-Brain and SRCNN3D networks (see Figure 2.4). A strong improvement with respect to spline interpolation can be noted with few epochs (less than 500). Then, the mean PSNR increases slowly to reach substantial improvements around 2500 epochs. SRCNN3D seems to lead to better performances than SRCNNF-Brain no matter what the number of epochs used.

Table 3.1 provides a summary of quantitative evaluation within isotropic scale factor 2 for the following methods: cubic spline interpolation, non-local means upsampling (NMU) [Manjón et al., 2010b], Low-rank total variation (LRTV) [Shi et al., 2015], SRCNNF-Nat [Dong et al., 2016a], SRCNNF-Brain and SRCNN3D. The reported mean gain tends to show that CNN-based approaches achieve better performance than spline interpolation, NMU or LRTV. For NMU and LRTV, we used the code provided by the authors. Our experiments show that the use of CNN-based approaches can lead to significant improvement over spline interpolation. More specifically, it can be seen that training the networks using specific data provides better results than using models trained over natural images. Moreover, the use of a 3D CNN-based model achieves better performance than averaging 2D model outputs. Figure 2.5 shows examples of reconstructed 3D images obtained from all the compared techniques. Visually, HR estimation of SRCNN3D best preserves contours and has the best contrast compared with the results of other methods.

Our experiments shows that better performance can be achieved by learning model parameters on adequate data. 3D SR models for 3D data outperforms 2D counterparts thanks to the fact

Image	Cubic Spline		NMU		LRTV		SRCNNF-Nat		SRCNNF-Brain		SRCNN3D	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
KKI2009-01-MPRAGE	33.42	0.9234	33.60	0.9303	34.29	0.9458	34.47	0.9366	36.16	0.9616	36.61	0.9656
KKI2009-02-MPRAGE	31.27	0.9402	31.37	0.9465	31.88	0.9597	32.40	0.9500	34.19	0.9700	34.60	0.9727
KKI2009-03-MPRAGE	35.88	0.9541	36.19	0.9596	36.88	0.9688	37.11	0.9608	38.93	0.9783	39.57	0.9808
KKI2009-04-MPRAGE	34.49	0.9441	34.73	0.9499	35.48	0.9617	35.59	0.9526	37.43	0.9728	37.91	0.9756
KKI2009-05-MPRAGE	35.72	0.9392	36.08	0.9458	36.86	0.9583	36.72	0.9482	38.40	0.9695	38.88	0.9728
Mean	34.16	0.9402	34.40	0.9464	35.08	0.9585	35.26	0.9496	37.02	0.9704	37.51	0.9735
Standard deviation	1.90	0.0111	2.00	0.0106	2.09	0.0083	1.90	0.0088	1.90	0.0060	1.97	0.0053
Gain	-	-	0.24	0.0063	0.92	0.0183	1.10	0.0094	2.87	0.0302	**3.36**	**0.0333**

TABLE 2.1: The results of PSNR/SSIM for isotropic scale factor $\times 2$ with the gain between compared methods and spline interpolation ©[2017] IEEE.



(a) Original HR (b) LR image (c) Spline Interpolation (d) NMU

(e) LRTV (f) SRCNNF-Nat (g) SRCNNF-Brain (h) SRCNN3D

FIGURE 2.5: Illustration of SR results (KKI2009-02-MPRAGE) with isotropic voxel upsampling (scale factor is 2). LR data (b) with voxel size $2.4 \times 2 \times 2mm^3$ is up sampled to size $1.2 \times 1 \times 1mm^3$ ©[2017] IEEE.

that 3D architecture directly learns the 3D structure of MRI volumetric images. In the next sections, we will focus on improving the performance of 3D networks based on the sensitivity analysis of baseline 3D architectures.

2.2.2.3 Baseline and benchmarked for 3D architectures

The network architecture that is used as a baseline approach in this study is illustrated in Figure 2.3. The baseline network is a 10 blocks (convolution+ReLU) network with the following parameters: 64 convolution filters of size $(3 \times 3 \times 3)$ at each layer, mean squared error (MSE) as loss function, weight initialization by [He et al., 2015] (MSRA filler), Adam (adaptive moment estimation) method for optimization [Kingma and Ba, 2015], 20 epochs on Nvidia GPU and using Caffe package [Jia et al., 2014], batch size of 64, learning rate set to 0.0001, no regularization or drop out has been used. The learning rate multipliers of weights and biases are respectively 1 and 0.1. For benchmarking purposes, we consider

two other state-of-the-art SR models: low-rank total variation (LRTV) [Shi et al., 2015] and SRCNN3D [Pham et al., 2017a]. SRCNN3D [Pham et al., 2017a], which is an extension in 3D of the method described in [Dong et al., 2016a], has 3 convolutional layers with the size of 9^3, 1^3 and 5^3 respectively. The layers of SRCNN3D consist respectively of 64 filters, 32 filters and one filter.

The next sections present the impact of the key parameters studied in this work: optimization method, weight initialization, residual-based model, network depth, filter size, filter number, training patch size and size of training dataset.

2.2.2.4 Optimization method

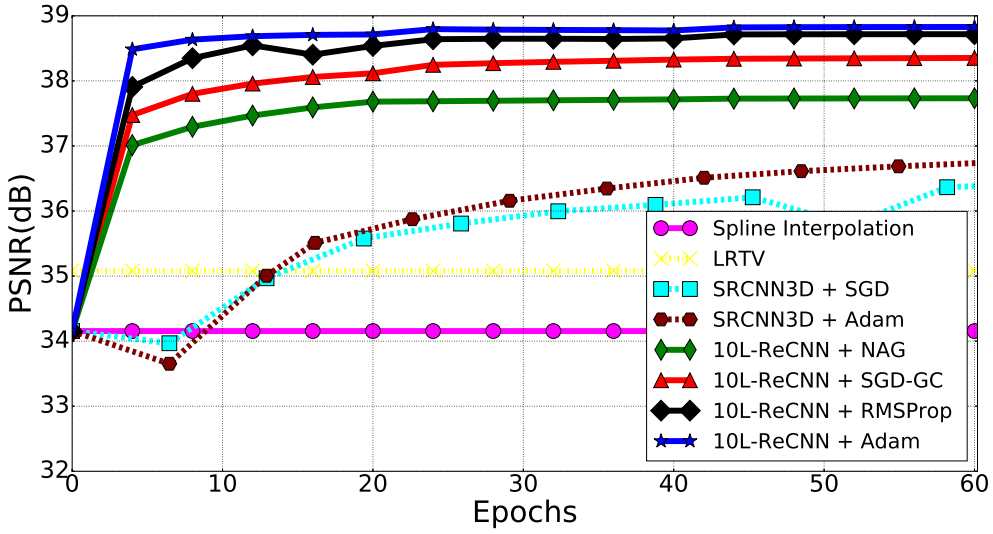


FIGURE 2.6: Impact of the optimization methods onto SR performance: SGD-GC, NAG, RMSProp and Adam optimisation of a 10L-ReCNN (10-layer residual-learning network with $f = 3$ and $n = 64$). We used Kirby 21 for training and testing with isotropic scaling factor $\times 2$. The initial learning rates of SGC-GC, NAG, RMSProp and Adam are set respectively to 0.1, 0.0001, 0.0001 and 0.0001. These learning rates are decreased by a factor of 10 every 20 epochs. The momentum of these methods, except RMSProp, is set to 0.9. All optimization methods use the same weight initialization described in [He et al., 2015].

Given a training dataset which consists of pairs of LR and HR images, network parameters are estimated by minimizing the objective function using optimization algorithms. These algorithms play a very important role in training neural networks. The more efficient and effective optimization strategies lead to faster convergence and better performance. More precisely, during the training step, the estimation of the restoration operator F corresponds to the minimization of the objective function $\mathcal{L}$ in Equation (2.66) over network parameters $\boldsymbol{\theta} = \{W_i, B_i\}_{i=1,\ldots,L}$.

Most optimization methods for CNNs are based on gradient descent. A classic method applies a mini-batch stochastic gradient descent with momentum (SGD) [LeCun et al., 1998] as used in [Dong et al., 2016a, Pham et al., 2017a]. However, the use of fixed momentum

causes numerical instabilities around the minimum. Nesterov's accelerated gradient (NAG) [Nesterov, 1983] was proposed to cope with this issued.

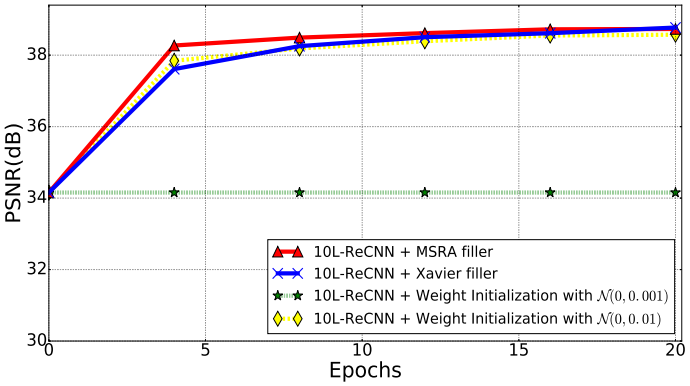
The use of small learning rates induces slow convergence. By contrast, high learning rates may lead to exploding gradients [Bengio et al., 1994, Glorot and Bengio, 2010]. In order to address this issue, [Kim et al., 2016a] proposed the stochastic gradient descent method with an adjustable gradient clipping (SGD-GC) [Pascanu et al., 2013] to achieve an optimization with high learning rates (e.g. $\alpha = 0.1$). The predefined range over which gradient clipping is applied may still cause SGD-GC not to converge quickly or make difficult the tuning of a global learning rate. Recently, methods have been proposed to address this issue through an automatic adaption of the learning rate for each parameter to be learned. RMSProp (root-mean-square propagation) [Tieleman and Hinton, 2012] and Adam (adaptive moment estimation) [Kingma and Ba, 2015] are the two most popular models in this category.

The results of four optimization methods (NAG, SGD-GC, RMSProp and Adam) for the baseline network are illustrated in Figure 2.6. Firstly, regardless the method used, the baseline network shows better performance than LRTV [Shi et al., 2015] and SRCNN3D [Pham et al., 2017a]. Secondly, it can be observed that the baseline network can converge very fast and stably. Concretely, the proposed optimization scheme needs only 20 epochs with small learning rate of 0.0001 to converge while the SRCNN3D shown in Figure 2.4 takes 2500 epochs. Finally, in these experiments, the most efficient and effective optimization method is Adam as regards both PSNR metric and convergence speed. Hence, in the next sections, we use Adam method with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ to train our networks with 20 epochs.

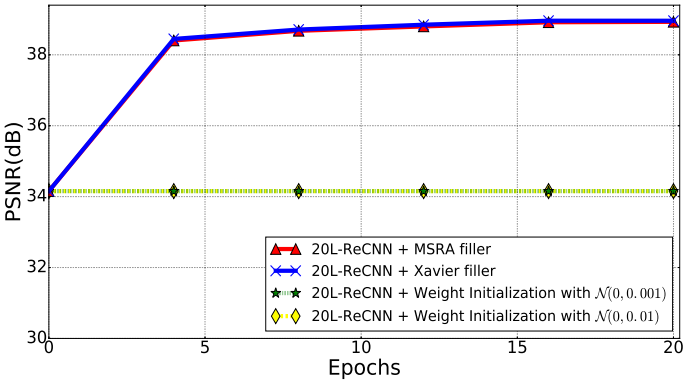
2.2.2.5 Weight initialization

The optimization algorithms for training a CNN are typically initialized randomly. Inappropriate initialization can lead to long time convergence or even divergence. Several studies [Dong et al., 2016a, Oktay et al., 2016, Pham et al., 2017a] used a normal distribution $\mathcal{N}(0, 0.001)$ to initialize the weights of convolutional filters. However, because of too small initial weights, the optimizer can be stuck into a local minimum especially when building deeper networks. Both [Dong et al., 2016a] concluded that deeper networks do not lead to better performance, and [Oktay et al., 2016] confirmed that the addition of extra convolutional layers to the 7-layer model is found to be ineffective. Uniform distribution $\mathcal{U}(-\sqrt{3/(nf^3)}, \sqrt{3/(nf^3)})$ (called Xavier filler) [Glorot and Bengio, 2010] was also proposed to initialize the weights of deeper networks. In order to add more layers to networks, [He et al., 2015] suggested an initial training stage by sampling from the normal distribution $\mathcal{N}(0, \sqrt{2/(nf^3)})$ (called here Microsoft Research Asia - MSRA filler).

Overall, we evaluate here the weight initialization schemes described in [Glorot and Bengio, 2010] and [He et al., 2015], a normal distribution $\mathcal{N}(0, 0.001)$ as in [Dong et al., 2016a, Oktay et al., 2016] and a normal distribution $\mathcal{N}(0, 0.01)$ for the considered SR architecture. Experiments with a deeper architecture were also performed, more precisely for a 20-layer



(a) : 10-layer residual-learning networks (10L-ReCNN)



(b) : 20-layer residual-learning networks (20L-ReCNN)

FIGURE 2.7: Weight Initialization Scheme vs Performance (residual-learning networks with the same filter numbers $n = 64$ and filter size $f = 3$ using Adam optimization and tested with isotropic scaling factor $\times 2$ using Kirby 21 for training and testing, 32000 patches with size 25^3 for training).

architecture, which is the deepest architecture that could be implemented for the considered experimental setup due to GPU memory setting. As shown in Figure 2.7, the initialization with normal distributions $\mathcal{N}(0, 0.001)$ failed to make the training of both 10-layer and 20-layer residual-learning networks converge. In addition, the 20-layer network also does not converge when initialized with normal distributions $\mathcal{N}(0, 0.01)$. By contrast, MSRA and Xavier filler schemes make the networks converge and reach similar reconstruction performance. For the rest of this chapter, we use MSRA weight filler as initialization scheme.

2.2.2.6 Residual learning

The CNN methods in [Dong et al., 2016a,b, Shi et al., 2016] use the LR image as input and outputs the HR one. We refer to such approach as a non-residual learning. Within these approaches, low-frequency features are propagated through the layers of networks, which may increase the representation of redundant features in each layer and in turn the computational efficiency of the training stage. By contrast, one may consider residual learning or normalized HR patch prediction as pointed out by several learning-based SR methods [Kim et al., 2016a, Timofte et al., 2013, 2014, Zeyde et al., 2012]. When considering CNN methods, one may design a network which predicts the residual between the HR image and the output of the first

transposed convolutional layer [Oktay et al., 2016]. Using residual blocks, a CNN architecture may implicitly embed residual learning while still predicting the HR image [Ledig et al., 2017].

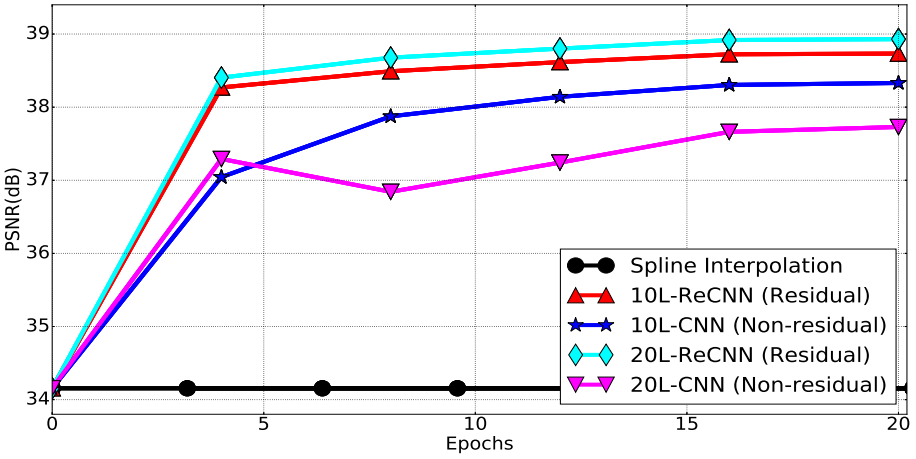


FIGURE 2.8: Non-residual-learning vs Residual-learning networks with the same $n = 64$ and $f^3 = 3^3$ and the depths of 10 and 20 (called here 10L-CNN vs 10L-ReCNN and 20L-CNN vs 20L-ReCNN) over 20 training epochs using Adam optimization with the same training strategy and tested with isotropic scale factor $\times 2$ using Kirby 21 for training and testing.

Here, we perform a comparative evaluation of non-residual learning vs. residual learning strategies. Figure 2.8 depicts PSNR values and convergence speed of residual vs non-residual network structures with 10 and 20 convolutional layers. The residual-learning networks converge faster than the non-residual-learning ones. In addition, residual learning leads to improvements in PSNR (+0.4dB for 10 layers and +1.2dB for 20 layers). It might be noted that these experiments do not support the common statement that the deeper, the better for CNNs. Here, the use of additional layers is only beneficial when using residual modeling. Deeper architectures even lower the reconstruction performance with non-residual learning.

2.2.2.7 Depth, filter size and number of filters

As shown by the previous experiment, the link between network depth and performance remains unclear. Besides, it is hard to train deeper networks because gradient computation can be unstable when adding layers [Glorot and Bengio, 2010]. For instance, [Oktay et al., 2016] tested extra convolutional layers to a 7-layer model but achieved negligible performance improvement. As mentioned above, SRCNN [Dong et al., 2016a] was also tested with deeper architectures but no improvement was reported. However, [Kim et al., 2016a] argue that the performance of CNNs for SR could be improved by increasing the depth of network compared to neural network architectures in [Dong et al., 2016a, Oktay et al., 2016].

The previous section supports that deeper architectures may be beneficial when considering a residual learning. We further evaluate here the reconstruction performance as a function of the number of layers. Results are reported in Figure 2.9. They stress that increasing network depth with residual learning improves the quality of the estimated HR image (e.g. +1.6dB increasing of the depth from 3 to 20 or +0.5dB increasing of the depth from 7 to 20).

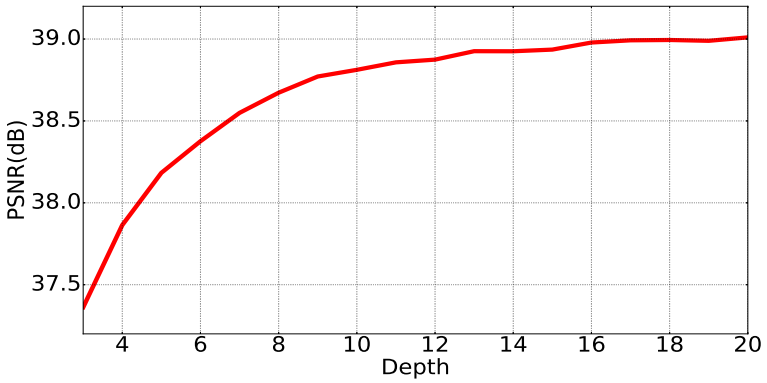


FIGURE 2.9: Depth vs Performance (residual-learning networks with the same filter numbers $n = 64$ and filter size $f = 3$ over 20 training epochs using Adam optimization and tested with isotropic scale factor $\times 2$ using Kirby 21 for training and testing, 32000 patches with size 25^3 for training).

The parameterization of the convolutional filters is also of key interest. Inspired by the VGG network designed for classification [Simonyan and Zisserman, 2014], previous CNN methods for SR mostly focused on small convolutional filters of size ($3\times3\times3$) in [Kamnitsas et al., 2017, Kim et al., 2016a, Oktay et al., 2016]. Small filter size can build deeper networks but reduces the memory for computation cost [Simonyan and Zisserman, 2014]. [Oktay et al., 2016] even argued that such architecture can lead to better non-linear estimations. Regarding the number of filters for each layer, [Dong et al., 2016a] reported greater reconstruction performance when increasing the number of filters. But these experiences were not reported in other CNN-based SR studies [Kim et al., 2016a, Oktay et al., 2016]. Here, we both evaluate the effect of the filter size and of the number of filters.

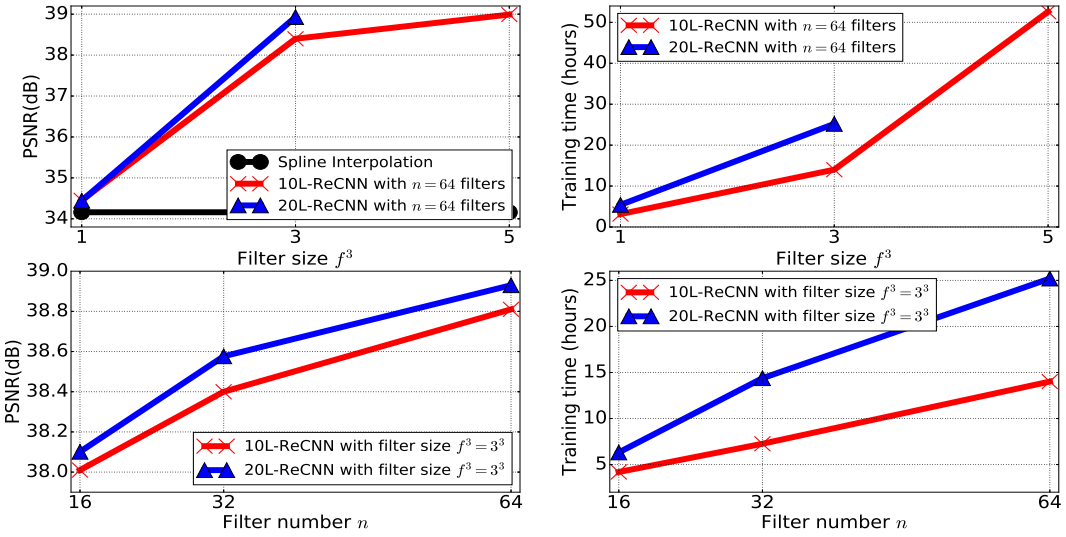


FIGURE 2.10: Impact of convolution filter parameters (sizes $f \times f \times f = f^3$ with n filters) on PSNR and computation time. These 10-layers residual-learning networks are trained from scratch using Kirby 21 with Adam optimization over 20 epochs and tested with the testing images of the same dataset for isotropic scale factor $\times 2$.

Figure 2.10 shows that a 10-layer network with a filter size of 5^3 shows results as well as a 20-layer network with 3^3 filters. Besides reconstruction performance, the use of a larger

filter size decreases the training speed and significantly increases the complexity and memory cost for training. For example, it took us 50 hours to train a 10-layer network with a filter size of 5^3. By contrast, a deeper network with smaller filters (i.e. 20-layer network with 3^3 filters) involves a smaller number of parameters, such that it took us only 24 hours to train. These experiments suggest that deeper architectures with small filters can replace shallower networks with larger filters both in terms of computational complexity and of reconstruction performance. In addition, the increase in the number of filters within networks can increase the performance. However, we were not able to use 128 filters with the baseline architecture due to the limited amount of memory. This stresses out the need to design memory efficient architectures for 3D image processing using deeper CNNs with more filters.

2.2.2.8 Training patch size and subject number

In the context of brain MRI SR, the acquisition and collection of large datasets with homogeneous acquisition settings is a critical issue. We here evaluate the extent to which the number of training subjects affects SR reconstruction performance. As the training samples are extracted as patches of brain MRI images, we also evaluate the impact of the training patch size onto learning and reconstruction performance.

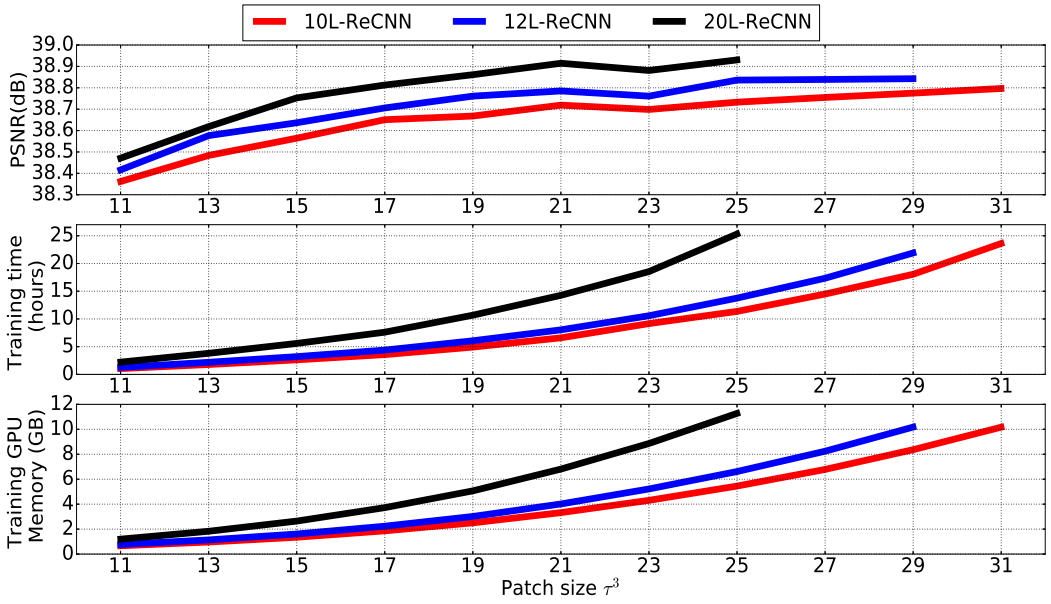


FIGURE 2.11: First row: Training patch size vs Performance. Second row: Patch size vs Training Time. Third row: Patch size vs Training GPU Memory Requirement. These networks with the same $n = 64$ and $f^3 = 3^3$ are trained from scratch using Kirby 21 with batch of 64 and tested with the testing images of the same dataset for isotropic scale factor $\times 2$.

The size of training patches should be larger or equal to the size of the receptive field (the region of the input space affects a particular layer) of the considered network [Kim et al., 2016a, Simonyan and Zisserman, 2014], which is given by $((f-1)D+1)^3$ for a D-layer network with filter size f^3. Figure 2.11 confirms that better performance can be achieved using larger

training patches (from 11^3 to 31^3 with the 10-layer network and from 11^3 to 29^3 with the 12-layer network). However, if the patch size is larger than the receptive field (e.g. 21^3 within the 10-layers network and 25^3 within the 12-layers network), the improvement is very little while we consume considerably more GPU memory and training time.

We stressed previously that the selection of the network depth involves a trade-off between reconstruction performance and GPU memory requirement and training time increase. A similar result can be drawn with respect to the patch size. Figure 2.11 illustrates that larger training patch sizes also require more memory for training. It may be noted that the performance of the 10-layer networks may reach a performance similar to 12-layer and 20-layer networks when using larger training patches but it takes more time and more GPU memory for training.

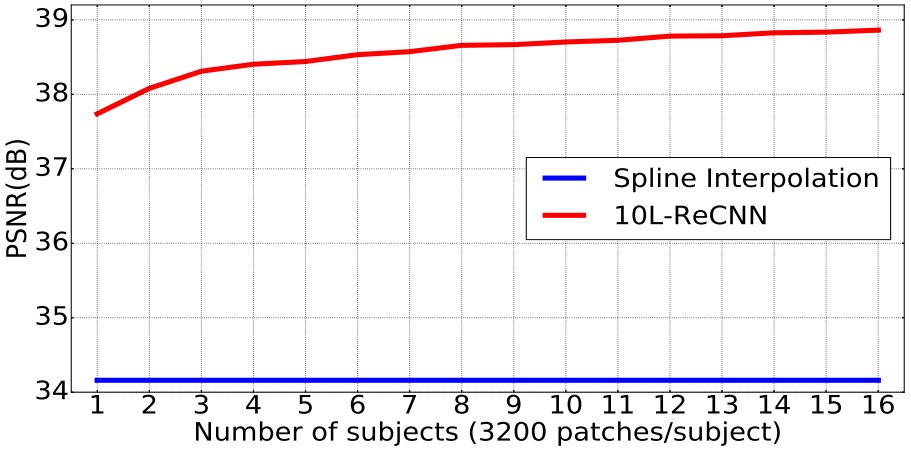


FIGURE 2.12: Number of Subjects vs Performance (10-layer residual-learning networks with the same filter numbers $n = 64$ and filter size $f = 3$ over 20 training epochs using Adam optimization and tested with isotropic scale factor $\times 2$ using Kirby 21 for training and testing, 3200 patches per subject with size 25^3 for training).

Regarding the number of training subjects, Figure 2.12 points out that a single subject is enough to reach better performance than spline interpolation. This has also been discovered in the work of [Shocher et al., 2018, Zhao et al., 2018] in which a super-resolution pipeline using the right testing image (self SR) is proposed. Interestingly, reconstruction performance increases slightly when more subjects are considered, which appears appropriate for real-world applications. However, in fact, more training dataset takes more time within the same experience settings. In the next sections, for saving training time, we propose to use 10 subjects for learning.

2.2.2.9 Handling arbitrary scales

In some CNN-based SR approaches, the networks are learned for a fixed and specified scaling factor. Thus, a network built for one scaling factor cannot deal with any other scale. In medical imaging, [Oktay et al., 2016] have applied CNNs for upscaling cardiac image slices with the scale of 5 (e.g. upscaling the voxel size from $1.25 \times 1.25 \times 10.00mm$ to $1.25 \times 1.25 \times$

$2.00mm$). Typically, their network is not capable of handling other scales due to the use of fixed deconvolutional layers. In brain MRI imaging, the variety of the possible acquisition settings motivates us to explore multi-scale settings.

Test / Train	Full-training			
	Same training samples			Double samples
	×(2,2,2)	×(3,3,3)	×(2,2,2),(3,3,3)	×(2,2,2),(3,3,3)
	PSNR	PSNR	PSNR	PSNR
×(2,2,2)	**39.01**	35.25	**37.35**	**38.80**
×(2,2,3)	36.80	35.11	36.47	37.24
×(2,2.5,2)	37.71	35.41	36.91	37.93
×(2,3,3)	35.23	35.13	35.75	36.20
×(2.5,2.5,2.5)	35.47	35.52	36.09	36.63
×(3,3,3)	33.43	**35.01**	**34.89**	**35.20**

TABLE 2.2: Experiments with multiple isotropic scaling factors with the 20-layers network using the training and testing images of Kirby 21. **Bold numbers** indicate that the tested scaling factor is present in the training dataset. We test two conditions of same training data and double training data

Following [Kim et al., 2016a], we investigate how we may embed multiple scales in a single network. It consists in creating a training dataset within which we consider LR and HR image pairs corresponding to different scaling factors. We test two cases: the first condition where the learning dataset for combined scale factors $(\times 2, \times 3)$ has the same number as a single scale factor and the second one where we double the learning dataset for multiple scale factors. To avoid a convergence towards a local minimum of one of the scaling factors, we learn network parameters on randomly shuffled dataset.

Table 2.2 summarizes experimental results. First, when the training is achieved for the scaling from $(2 \times 2 \times 2)$ on a dataset of $(2 \times 2 \times 2)$ scale, it can be noticed that reconstruction performances decrease significantly when applied to other scaling factors (there is a drop from $39.01dB$ to $33.43dB$ when testing with $(3 \times 3 \times 3)$). Second, it can be noticed that when the training is performed on multi-scale data within the same training samples, there is no significant performance change compared to training from a single-scale dataset. Third, the more training dataset leads to a better performance. Training from multiple scaling factors leads to the estimation of a more versatile network. Overall, these results tend to show that one single network can handle multiple arbitrary scaling factors.

2.2.2.10 Multimodality-guided SR

In some clinical cases, it is common to acquire one isotropic HR image and LR images with different modalities (different contrasts) in order to limit the acquisition time. Hence, a coplanar isotropic HR image might be considered as a complementary information source to reconstruct HR MRI images from LR ones [Rousseau et al., 2010a]. To address this multimodality-guided SR problem, we add a concatenation layer as the first layer of the network as illustrated in Figure 2.13. This layer concatenates the ILR image and a registered

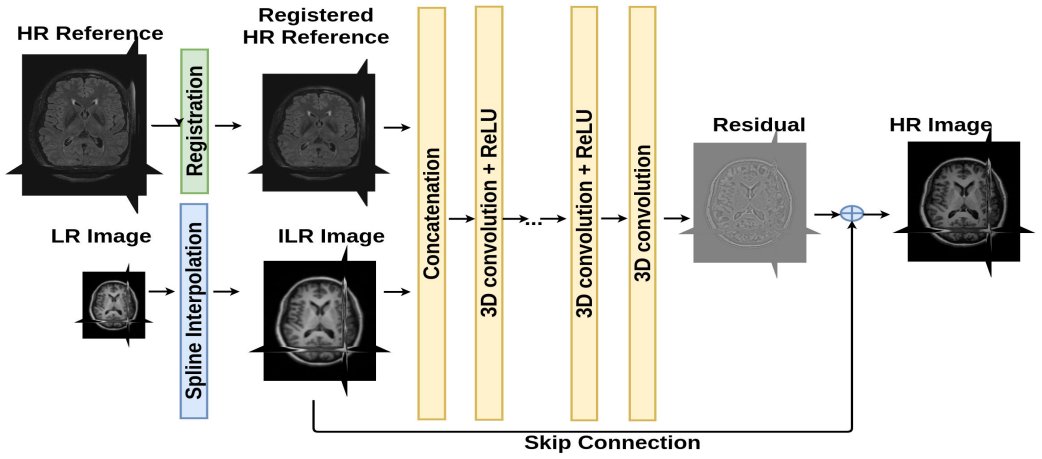
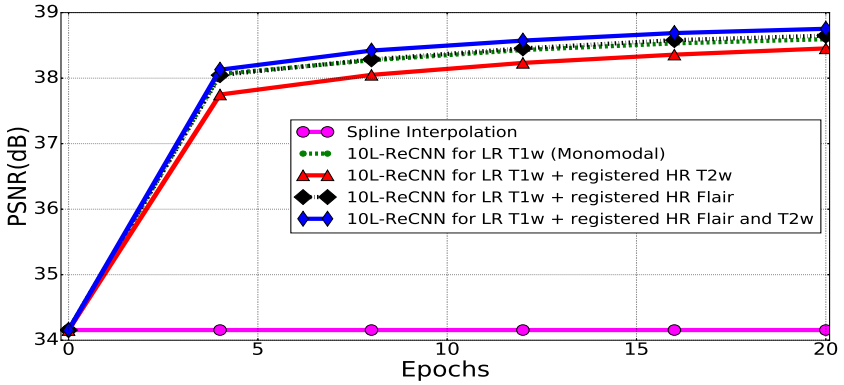
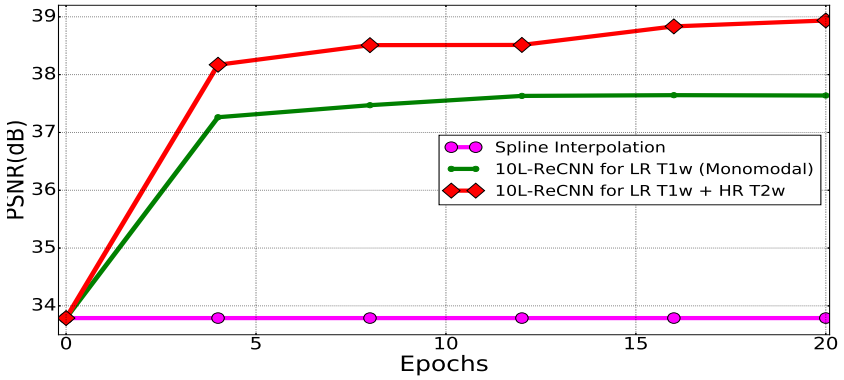


FIGURE 2.13: 3D deep neural network for multimodal brain MRI super-resolution using intermodality priors. Skip connection computes the residual between ILR image and HR image.

HR reference along the channel axis. The registration step of HR reference ensures that the two input images share the same geometrical space.



(a) Multimodal experiments using Kirby dataset for training and testing.



(b) Multimodal experiments using NAMIC dataset for training and testing.

FIGURE 2.14: Multimodality-guided SR experiments. The LR T1-weighted images are upscaled with isotropic scale factor $\times 2$ using respectively monomodal network (10L-ReCNN for LR T1w), HR T2w multimodal network, HR Flair multimodal network and both HR Flair and T2w multimodal images.

We experimentally evaluate the relevance of the proposed multimodality-guided SR model according to the following setting. We investigate whether the complementary use of a Flair or

a T2-weighted MRI image might be beneficial to improve the resolution of a LR T1-weighted MRI image. Concerning the Kirby dataset, we apply an affine transform estimated using FSL [Jenkinson et al., 2012] to register images from the same subject into a common coordinate space. We assume here that the affine registration can compensate motion between two scans acquired during the same acquisition session since here an organ does not undergo significant deformation between two acquisitions. The registration step has been checked visually for all the images. Data of the NAMIC dataset are already in the same coordinate space so no registration step is required.

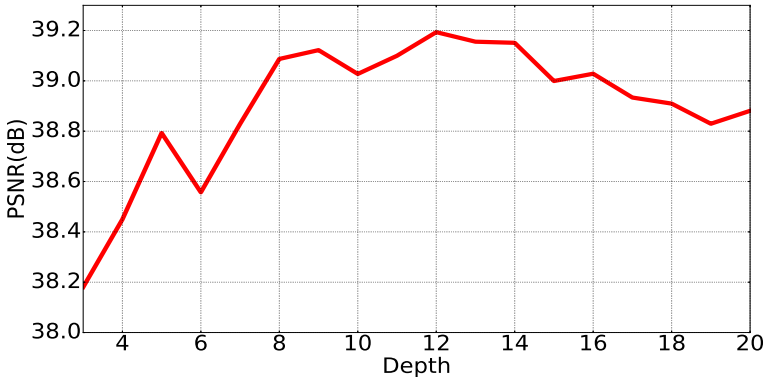
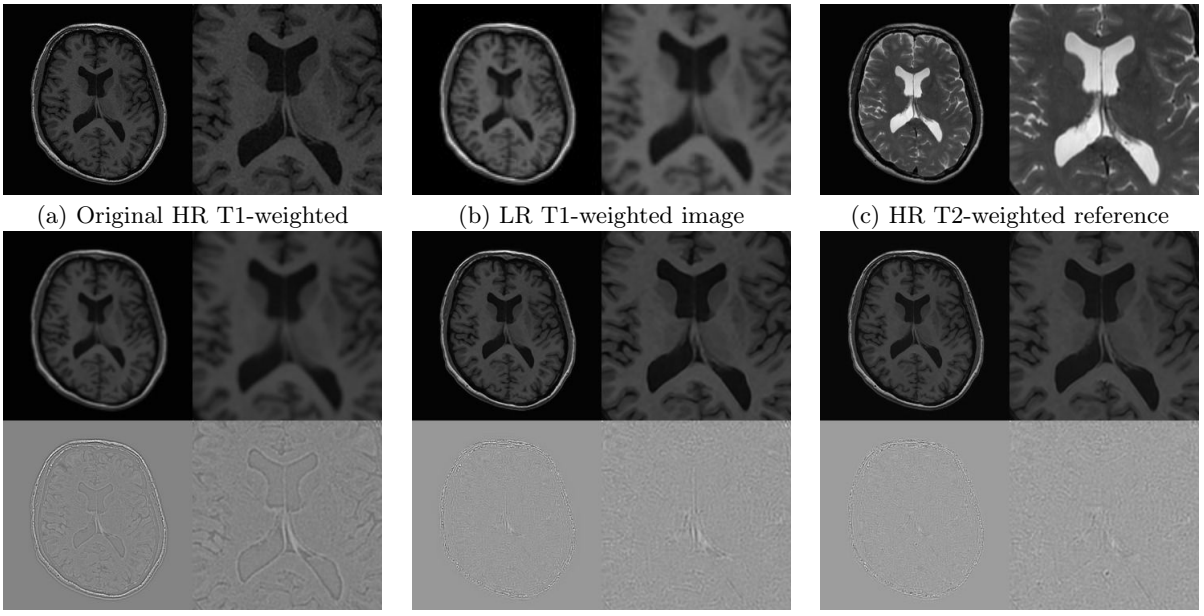


FIGURE 2.15: Depth vs Performance (multimodal SR using residual-learning networks with the same filter numbers $n = 64$ and filter size $f = 3$ over 20 training epochs using Adam optimization and tested with isotropic scale factor $\times 2$ using NAMIC for training and testing).

Figure 2.14 shows the results of the multimodality-guided SR compared to the monomodal SR for both Kirby dataset (a) and NAMIC datasets (b). It can be seen that multimodality driven approach can lead to improved reconstruction results. In these experiments, the overall upsampling result depends on the quality of the HR image used to drive the reconstruction process. Thus, adding high resolution information containing artifacts limits reconstruction performance. This is especially the case for the Kirby dataset. For instance, when considering T2w images, no improvement is observed for Kirby dataset and an improvement greater than $1dB$ is reported for NAMIC dataset. As the T2w image resolution is lower than T1w modality in Kirby dataset, these results may emphasize the requirement for HR information source to expect significant gain with respect to the monomodal model. Figure 2.16 shows visually that edges in the residual image between the ground truth and the reconstruction by the multimodal approach are reduced significantly compared to interpolation and monomodal methods (e.g. the regions of lateral ventricles). This means that the multimodal approach brings the reconstructions which are the most similar to the ground truth. These qualitative results highlight the fact that the proposed multimodal method provides a more favorable performance than other compared methods.

In addition, we explore the impact of the network depth augmentation with regard to the performance of multimodal SR approach. The experiments shown in Figure 2.15 indicate that the deeper structures do not lead to better results within the multimodal method.



(a) Original HR T1-weighted (b) LR T1-weighted image (c) HR T2-weighted reference

(d) Spline Interpolation (e) Monomodal 10L-ReCNN (g) Multimodal 10L-ReCNN

FIGURE 2.16: Illustration of the axial slices of monomodal and multimodal SR results (01018, pathological case) with isotropic voxel upsampling using NAMIC for training and testing. LR T1-weighted image (b) with voxel size $2 \times 2 \times 2mm^3$ is upsampled to size $1 \times 1 \times 1mm^3$. Multimodal network 10L-ReCNN uses the HR T2-weighted reference (c) to upscale LR image. The different between ground truth image and reconstruction results are at the bottom. Their zoom version are at the right.

2.2.2.11 How transferable are learned features?

Training a CNN from scratch requires an amount of training data and may take a long time. Moreover, to avoid overfitting, the training dataset has to reflect the appearance variability of the images to reconstruct. In the context of brain MRI, part of image variability comes from acquisition systems. Hence, we investigate the impact of such image variability onto SR performance by evaluating transfer learning skills among different datasets corresponding to the same imaging modality.

In order to characterize such generalization skills, we evaluate the extent to which the selection of a given training dataset affects the reconstruction performance of the network. We proceed as follows: We train from scratch two 20L-ReCNN networks separately for a 10-image NAMIC T1-weighted dataset and a 10-image Kirby T1-weighted dataset, and we test the trained models for the remaining 10-image NAMIC and Kirby T1-weighted datasets. The considered case-study involves a scaling factor of $(2 \times 2 \times 2)$. For quantitative comparison, the PSNR and the structural similarity (SSIM) (the definition of SSIM can be found in [Wang et al., 2004]) are used to evaluate the performance of each model in Table 2.3. For benchmarking purposes, we also include a comparison with the following methods: cubic spline interpolation, low-rank total variation (LRTV) [Shi et al., 2015], SRCNN3D [Pham et al., 2017a]. The use of 20-layer CNN-based approaches for each training dataset can lead to improvements over spline interpolation, LRTV method and SRCNN3D (with respect to both PSNR and SSIM). Although, we lose a little gain (e.g. PSNR: 0.55dB for testing Kirby and 0.74dB for NAMIC,

SSIM: 0.003 for Kirby and 0.0019 for NAMIC) when using different training and testing dataset (i.e. different resolution), our proposed networks have better results than compared methods.

For qualitative comparison, Figures 2.17 and 2.18 show the results of reconstructed 3D images obtained from all the compared techniques. The zoom version of the reconstructions 20L-ReCNN shows sharpen edges and a grayscale intensity which are closest to the ground truth. In addition, the HR reconstruction of the 20L-ReCNN model shows that its differences from the ground truth are less than other methods (i.e. the contours of the residual image of the 20L-ReCNN method are less occurrences than those of others). Hence, we can infer that our proposed method best preserves contours, geometrical structures and better recovers the image contrast compared with the other methods.

Testing dataset	Spline Interpolation		LRTV		SRCNN3D Kirby (10 images)		20L-ReCNN			
							Kirby (10 images)		NAMIC(10 images)	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Kirby (5 images)	34.16	0.9402	35.08	0.9585	37.51	0.9735	38.93	0.9797	38.06	0.9767
Standard deviation	1.90	0.0111	2.09	0.0083	1.97	0.0053	1.87	0.0044	1.83	0.0045
Gain	-	-	0.92	0.0183	3.36	0.0333	4.77	0.0395	3.9	0.0365
NAMIC (10 images)	33.78	0.9388	34.34	0.9549	36.72	0.9694	37.73	0.9762	38.28	0.9781
Standard deviation	1.82	0.0071	1.79	0.0044	1.76	0.0035	1.81	0.0031	1.78	0.0029
Gain	-	-	0.56	0.0161	2.94	0.0306	3.95	0.0374	4.5	0.0393

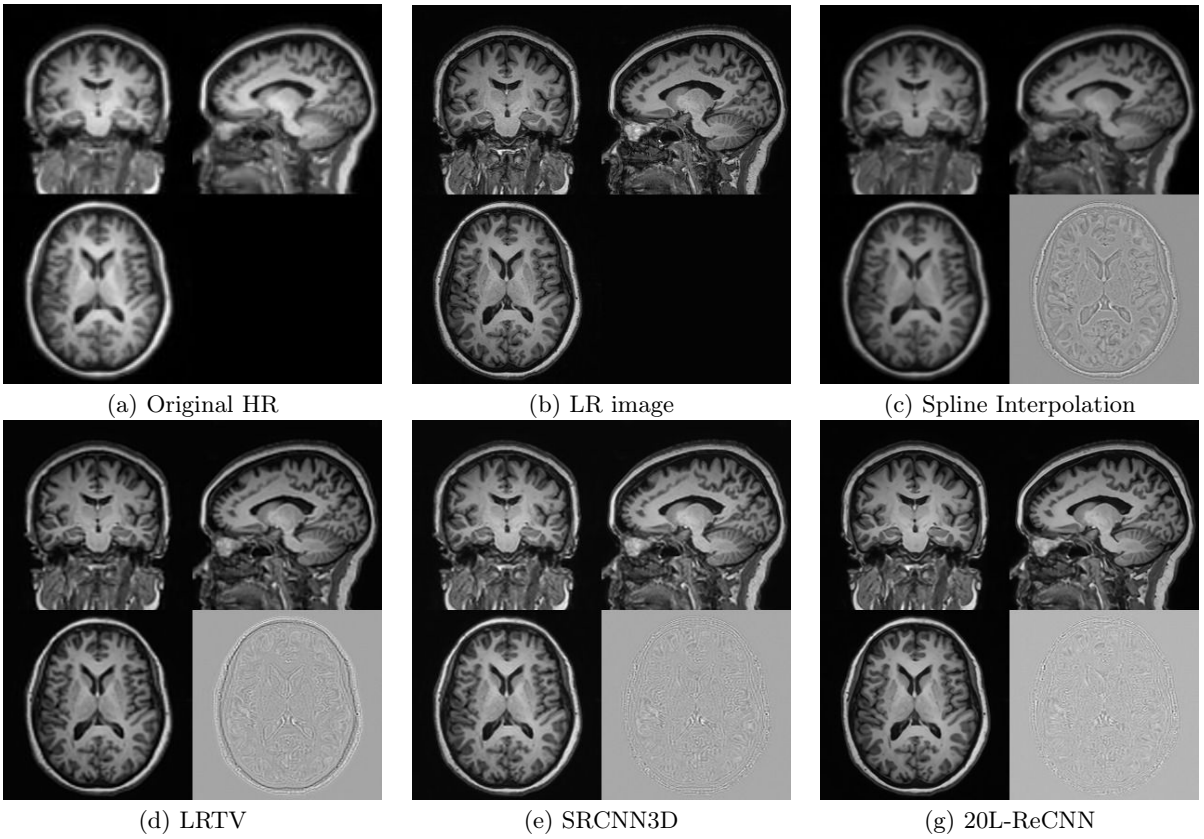
TABLE 2.3: The results of PSNR/SSIM for isotropic scale factor $\times 2$ with the gain between compared methods and the method of spline interpolation. One network 20L-ReCNN trained with 10 images of Kirby and one trained with NAMIC

2.2.3 Practical applications of super-resolution

There are many practical situations, including infant brain MRI scans [Makropoulos et al., 2018], rapid emergency scans [Walter et al., 2003], where the LR images with an anisotropic voxel size are typically acquired due to patient comfort (e.g. infants can not lie on bed for a long time, emergency). These images usually have a high in-plane resolution and a low through-plane resolution. Interpolation is commonly used to upsampled these LR image to isotropic digital resolution. However, interpolated LR images may lead partial volume artifacts that affect segmentation [Ballester et al., 2002]. In such cases, motion correction and multi-image super-resolution can be used to achieve HR isotropic images [Makropoulos et al., 2018]. If these methods are not always available, investigators and clinicians have no choice to process these LR images. For example, the MAIA dataset has the T2w images which acquired with the voxel size of $0.4464 \times 0.4464 \times 3\ mm$. In this section, we attempt to use our single image SR method to enhance the resolution of these clinical data and improve the segmentation methods applied to these images.

2.2.3.1 Super-resolution of clinical neonatal data

The idea is to use convolutional neural networks to transfer the rich information available from high-resolution experimental dataset to lower-quality image data. The procedure first



(a) Original HR (b) LR image (c) Spline Interpolation

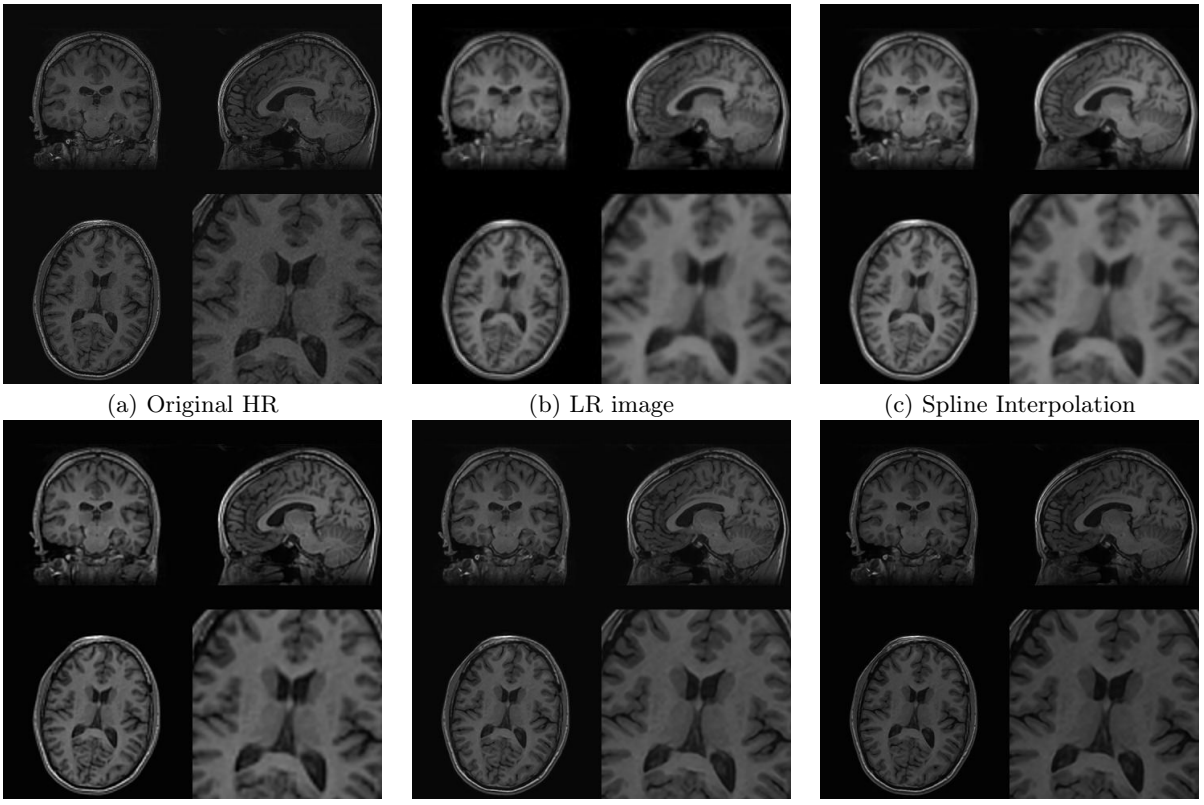
(d) LRTV (e) SRCNN3D (g) 20L-ReCNN

FIGURE 2.17: Illustration of SR results (KKI2009-02-MPRAGE, non-pathological case, of dataset Kirby) with isotropic voxel upsampling. LR data (b) with voxel size $2\times2\times2.4mm^3$ is upsampled to size $1 \times 1 \times 1.2mm^3$. The difference between the ground truth image and the reconstruction results are in the right bottom corners. Both network SRCNN3D and network 20L-ReCNN are trained with the 10 last images of Kirby.

uses CNNs to learn mappings between real HR images and their corresponding simulated LR images with the same resolution of real data. The LR data is generated by the observation model decomposed into a space-invariant blurring model and a downsampling operator. The two most popular choices for MRI PSF approximation for SR evaluation are a rectangular pulse Box-PSF with the box width of slice width [Manjón et al., 2010b], a Gaussian kernel [Greenspan, 2008, Rueda et al., 2013, Shi et al., 2015]. However, the most accurate representation is the use of a Gaussian kernel with the full-width-at-half-maximum (FWHM) equal to slice thickness [Greenspan, 2008]. Once models learned, these mappings enhance the LR resolution of unseen low quality images.

In order to verify the applicability of our CNN-based methods, we have used two neonatal brain MRI dataset: the dHCP dataset [Hughes et al., 2017] and the MAIA dataset. The HR images are T2-weighted MRIs of the Developing Human Connectome Project (dHCP) [Makropoulos et al., 2018], and provided by the Evelina Neonatal Imaging Centre, London, UK. 40 neonatal data were acquired on a 3T Achieva scanner with the repetition (TR) of 12 000 ms and the echo times (TE) of 156 ms respectively. The size of voxels is $0.5 \times 0.5 \times 0.5$ mm3. In-vivo neonatal LR images has a voxel size of $0.4464 \times 0.4464 \times 3$ mm3.

The pipeline of this application is described as follows:



(a) Original HR (b) LR image (c) Spline Interpolation

(d) Low-Rank Total Variation (LRTV) (e) 20L-ReCNN (trained with Kirby) (g) 20L-ReCNN (trained with NAMIC)

FIGURE 2.18: Illustration of SR results (01011-t1w, pathological case, of dataset NAMIC) with isotropic voxel upsampling. LR data (b) with voxel size $2 \times 2 \times 2mm^3$ is upsampled to size $1 \times 1 \times 1mm^3$. The zoom versions of the axial slices are in the right bottom corners.

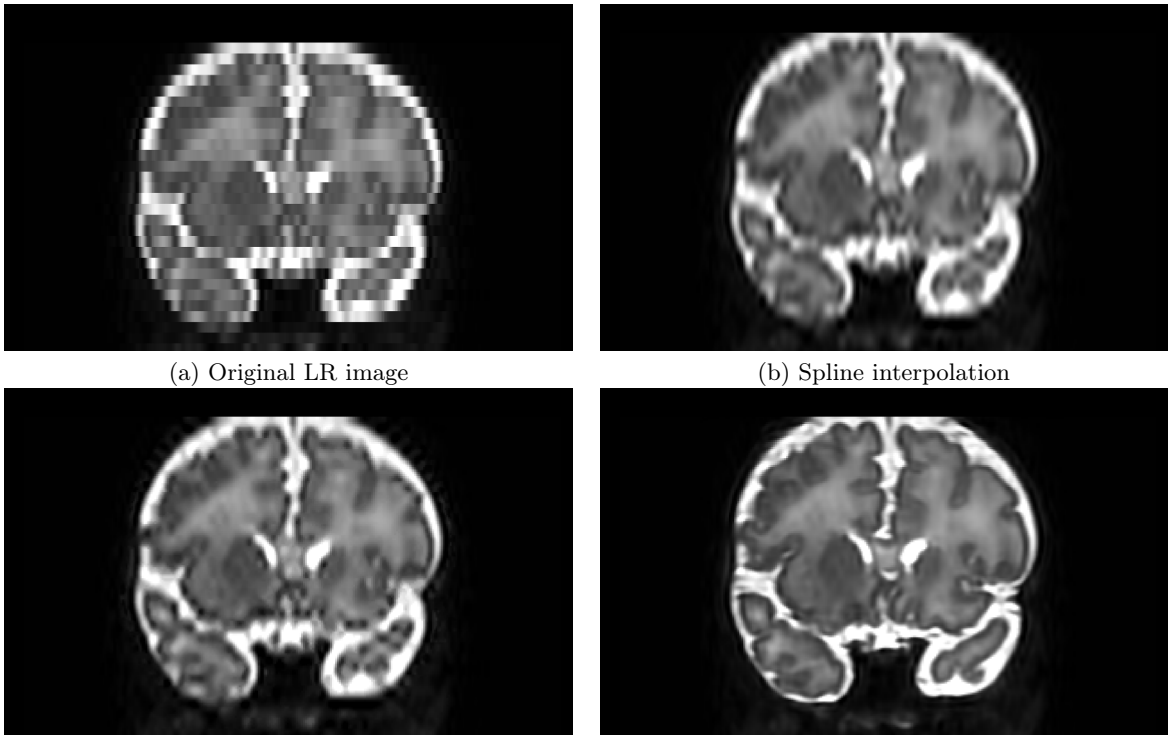
- The HR T2w images of the dHCP dataset are first filtered by a 3D Gaussian kernel with the standard deviation $(\sigma_x, \sigma_y, \sigma_z)$ calculated as :

$$\begin{cases} FWHW_x = 2\sqrt{2\ln 2}\sigma_x = ST_x \\ FWHW_y = 2\sqrt{2\ln 2}\sigma_y = ST_y \\ FWHW_z = 2\sqrt{2\ln 2}\sigma_z = ST_z \end{cases} \tag{2.68}$$

  where (x, y, z) is image coordinates, ST denotes the slice thickness of new images. Concretely, in this case, the slice thickness is calculated as $SW_x = 0.4464$mm, $SW_y = 0.4464$mm, $SW_z = 3$mm. Then, the blurred HR images are downscaled by nearest-neighbour interpolation to generate simulated LR images.

- The simulated LR images are then upscaled by the spline interpolation. HR and corresponding interpolated LR patches with the size of 25 in cube are cropped randomly from 40 pairs of the HR and the interpolated LR images with 3200 patches per image.

- A convolutional neural networks with 20 layers, in which the parameters are described in previous sections, learns the mapping between interpolated LR and HR patches. Once the network learned, the model is stored for the next step.



(a) Original LR image (b) Spline interpolation

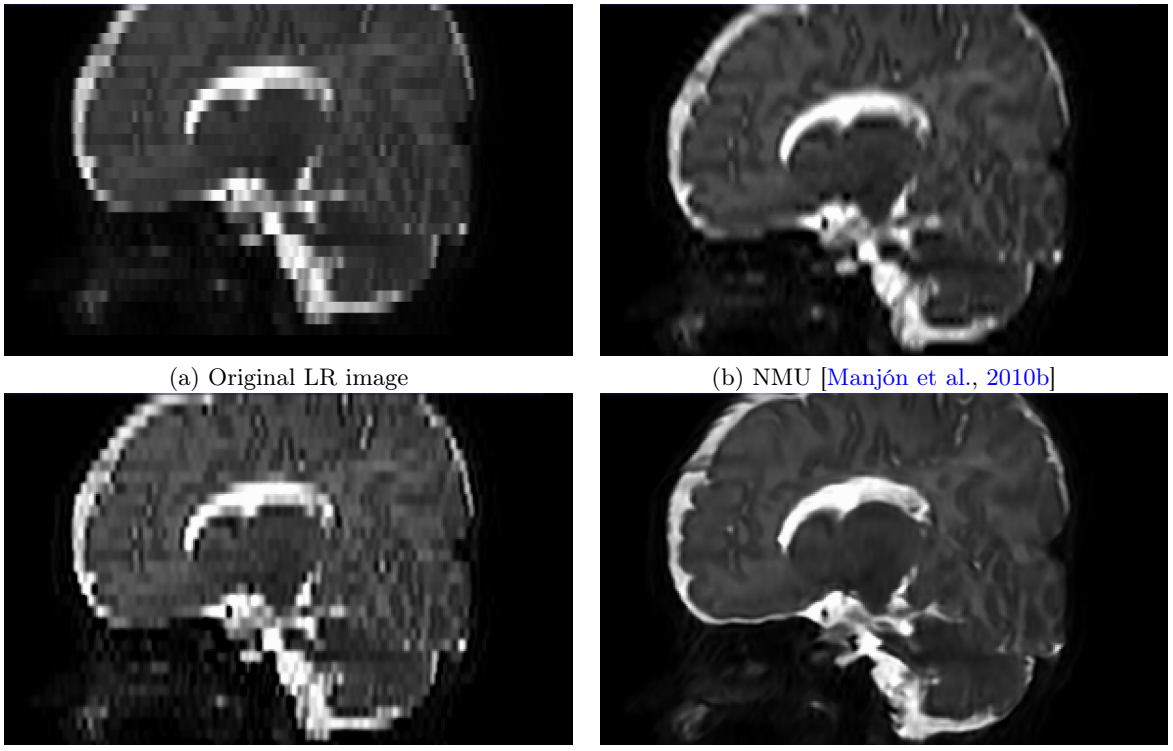
(c) NMU (d) 20L-ReCNN

FIGURE 2.19: Illustration of coronal SR results with isotropic voxel upsampling. Original data with voxel size of 0.4464 × 0.4464 × 3 is resampled to size 0.5 × 0.5 × 0.5 mm^3. 20L-ReCNN is trained with the dHCP dataset

- In order to apply our model, the real LR images are interpolated to have the voxel size equal to the one of HR dataset. Finally, the set of learned convolutional layers applies to the real interpolated LR images to obtain SR images.

Figures 2.19, 2.20 and 2.21 compares the qualitative results of HR reconstructions (spline interpolation, NMU [Manjón et al., 2010b] and our method) of a LR image from MAIA dataset. We also test LRTV [Shi et al., 2015] but do not achieve good reconstructions (shown in Figure 2.20 (c)). Note that we do not have the ground truth of real LR data for calculating quantitative metrics. The comparison reveals that the 20-layers CNNs-based proposed method recovers shaper images and better defined boundaries. For example, the cerebrospinal fluid (CSF) of the cerebellum of proposed method in Figure 2.19 is more visible than compared methods. The cortex of 20L-ReCNN method is less blurry than others in Figure 2.21. The ventricle

These results confirm qualitatively the efficacy of the approach. In addition, these results could support cortex segmentation due to the visibility of cortex boundaries.



(a) Original LR image (b) NMU [Manjón et al., 2010b]

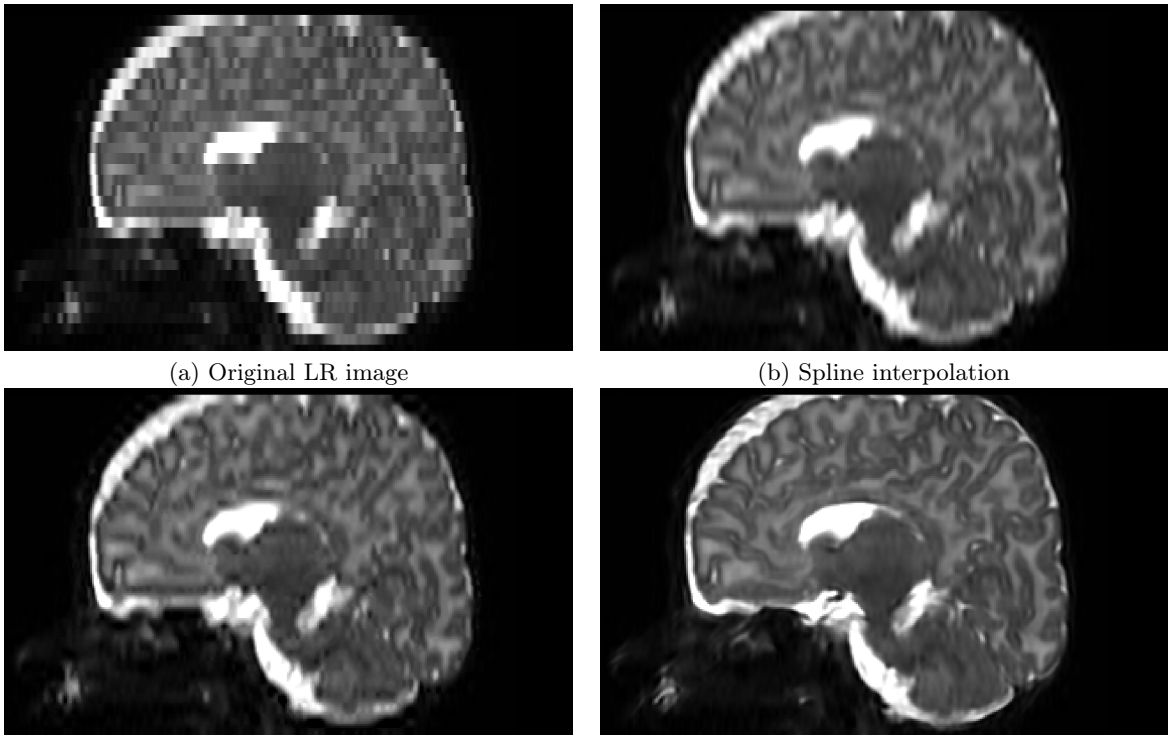
(c) LRTV [Shi et al., 2015] (d) 20L-ReCNN

FIGURE 2.20: Illustration of sagittal SR results with isotropic voxel upsampling. Original data with voxel size of $0.4464 \times 0.4464 \times 3$ is resampled to size $0.5 \times 0.5 \times 0.5\ mm^3$. 20L-ReCNN is trained with the dHCP dataset.

2.2.3.2 Super-resolution for segmentation

In this section, we would like to verify the contribution of SR to medical image segmentation. "SR cannot be viewed as an isolated domain." [Greenspan, 2008]. SR has a strong relationship with image segmentation. Indeed, super-resolution techniques are used to achieve more accurate segmentation maps of brain MRI data [Jog et al., 2016, Rueda et al., 2013]. In order to evaluate state-of-the-art segmentation algorithms in actual clinical settings with respect to our SR results, we use morphologically adaptive neonatal tissue segmentation (MANTIS) toolbox [Beare et al., 2016] to segment the cortex of MAIA SR T2w images. MANTIS proposes a pipeline which combines unified tissue segmentation and morphological adaptation to segment the neonatal brain. BET method of FSL toolbox [Jenkinson et al., 2012] is used to strip skull before applying MANTIS. Figure 2.22 shows the result of segmentation method MANTIS for spline interpolation and two SR technique: NMU [Manjón et al., 2010b] and our proposed method (20L-SRReCNN). The cortex segmentation within our 20L-SRReCNN is more fully connected than others. The outer boundary of cortex segmentation map of our method is smoother than compared methods.

Although, we do not have the ground truth segmentation maps (with the resolution of $0.5 \times 0.5 \times 0.5\ mm^3$) of the clinical T2w images, there are the manual segmentations of these subjects with respect to higher-resolution T1w images with voxel size of $0.268 \times 0.268 \times 1.2\ mm^3$ from a radiologist. We would like to evaluate the segmentation results with respect



(a) Original LR image (b) Spline interpolation

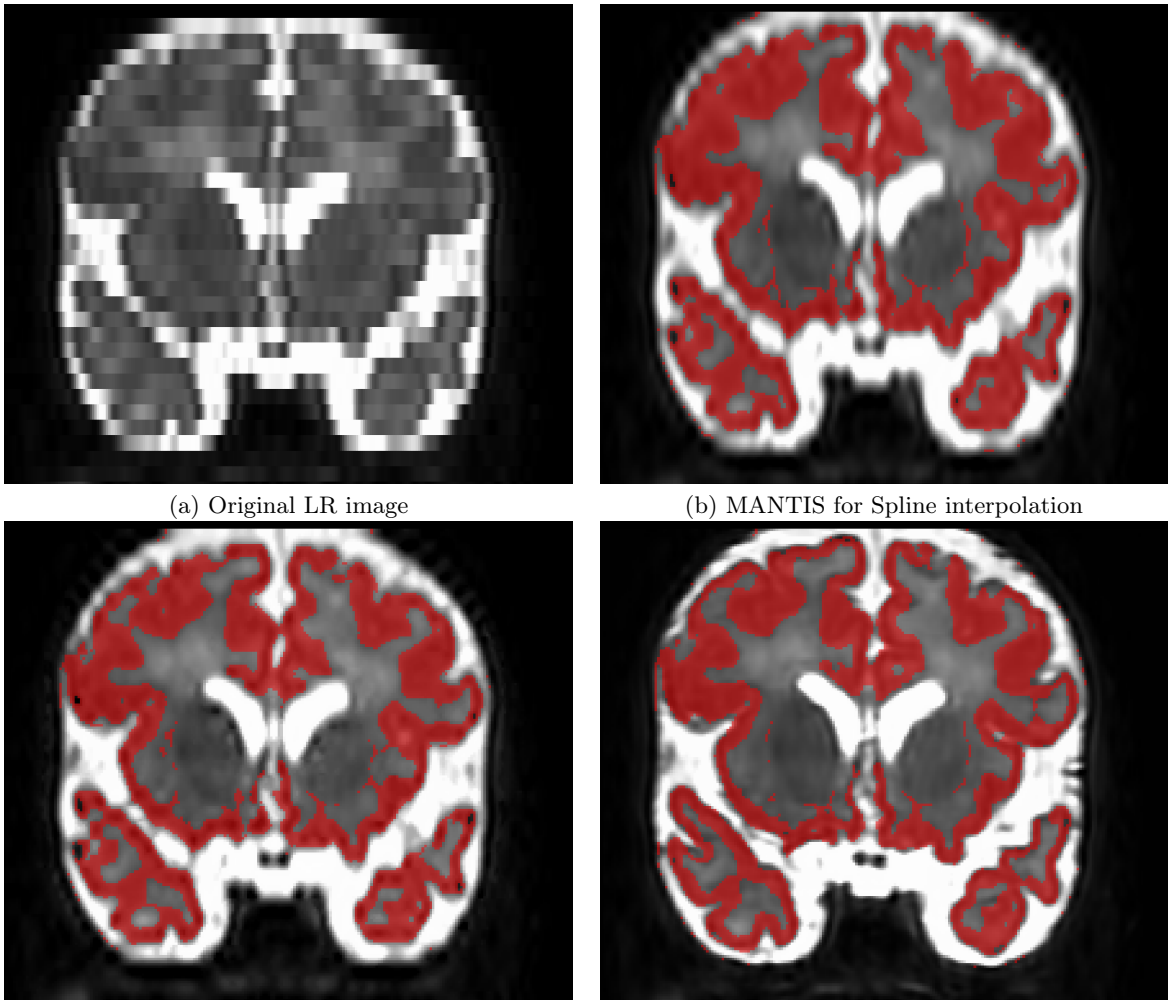
(c) NMU [Manjón et al., 2010b] (d) 20L-ReCNN

FIGURE 2.21: Illustration of sagittal SR results with isotropic voxel upsampling. Original data with voxel size of 0.4464 × 0.4464 × 3 is resampled to size 0.5 × 0.5 × 0.5 mm^3. 20L-ReCNN is trained with the dHCP dataset.

to upsampling methods by these higher-resolution manual segmentation maps. Because these T1w images and T2w images are not paired, the estimated segmentation maps are then mapped onto the original T1w images by a rigid registration between HR T2w and T1w data. A threshold of 0.5 is applied to generate binary segmentation maps. Table 2.4 shows the dice scores of the segmentation method MANTIS on the 2 images of the MAIA testing dataset with respect to different approaches: original T1w images with voxel size of $0.268 \times 0.268 \times 1.2\ mm^3$, interpolated T1w images with voxel size of $0.5 \times 0.5 \times 0.5\ mm^3$, original T2w images with voxel size of $0.4464 \times 0.4464 \times 3\ mm^3$, upsampling T2 images with voxel size of $0.5 \times 0.5 \times 0.5\ mm^3$ using interpolation, NMU and 20L-SRReCNN. The Dice index is described as:

$$Dice = \frac{2TP}{2TP + FP + FN} \tag{2.69}$$

where TP,FP and FN denote true positive, false positive and false negative between the estimated and the original segmentation. First, the segmentation results from isotropic-resolution T2w images are better than higher-resolution T1w images and isotropic T1w images. Secondly, super-resolution methods, which generate better reconstructions, support more accurate segmentation results. Finally, the segmentation method MANTIS for our estimated HR images shows the best results compared to other approaches. These results come from the fact that our SR method estimates more accurate HR reconstructions.



(a) Original LR image

(b) MANTIS for Spline interpolation

(c) MANTIS for NMU [Manjón et al., 2010b]

(d) MANTIS for 20L-ReCNN

FIGURE 2.22: Illustration of coronal cortex segmentation results (red color) using MANTIS toolbox [Beare et al., 2016] with isotropic voxel upsampling. Original data (a) with voxel size of $0.4464 \times 0.4464 \times 3$ is resampled to size $0.5 \times 0.5 \times 0.5\ mm^3$. 20L-ReCNN is trained with the dHCP dataset.

Subject	T1w		T2w			
	Original	Interp.	Original	Interp.	NMU	20L-SRReCNN
MAIA #1	0.6215	0.6205	0.7090	0.7052	0.7190	0.7330
MAIA #2	0.6746	0.6802	0.6694	0.7118	0.7182	0.7333

TABLE 2.4: Dice scores of the segmentation method MANTiS on the 2 images of the MAIA testing dataset with respect to different approaches (columns): original T1w images, interpolated (Interp.) T1w images, original T2w images, upsampling T2 images using interpolation, NMU and 20L-SRReCNN

2.2.4 Conclusion

The section 2.2 investigates CNN-based models for 3D brain MR image SR. Based on a comprehensive experimental evaluation, we would like to draw the following conclusions and recommendations regarding the setup to be considered. We highlight that eight complementary factors may drive the reconstruction performance of CNN-based models. The combination of 1) appropriate optimization with 2) weight initialization and 3) residual learning is a key to

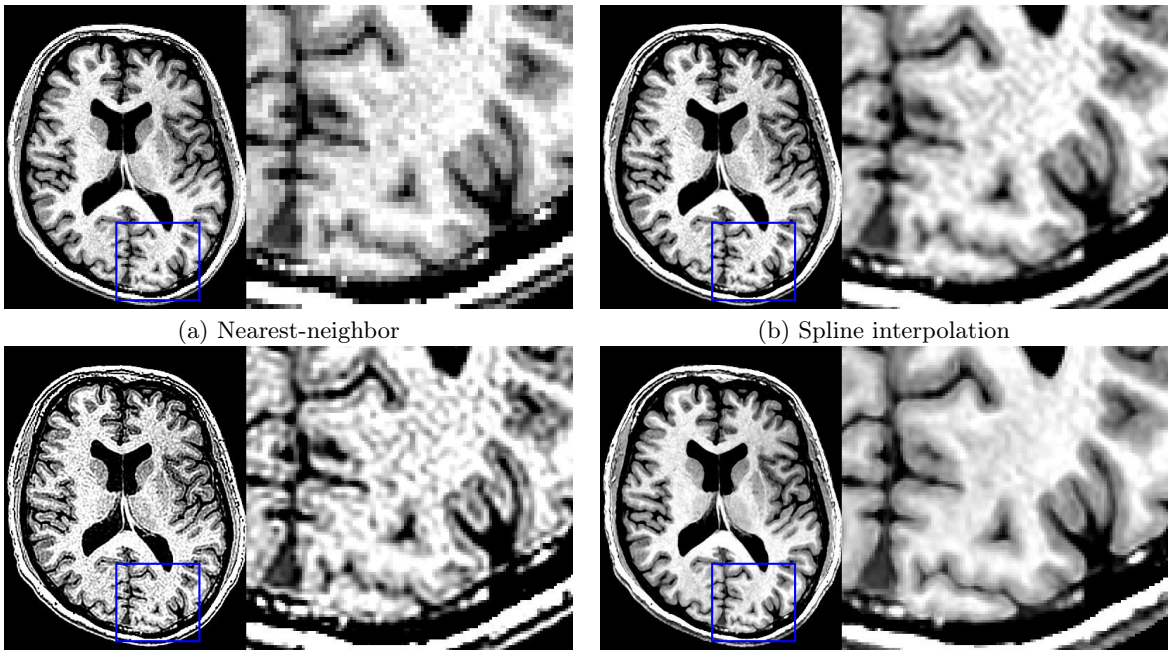
exploit deeper networks with a faster and effective convergence. The choice of an appropriate optimization method can lead to a PSNR improvement of (at least) $1dB$. In this study, it has appeared that Adam method [Kingma and Ba, 2015] provides significantly better reconstruction results than other classic techniques such as SGD, and a faster convergence. Moreover, weights initialization is a very important step. Indeed, some approaches simply do not achieve convergence in the learning phase. This study has also shown that residual modeling for single image SR is a straightforward technique to improve the reconstruction performances ($+0.4dB$) without requiring major changes in the network architecture. Appropriate weight initialization methods described in [Glorot and Bengio, 2010, He et al., 2015] allow us to build deeper residual-learning networks. From our point of view, these three aspects of SR algorithm are the first to require special attention for the implementation of a SR technique based on CNN.

Overall, we show that better performance can be achieved by learning a 4) deeper fully 3D convolution neural network, 5) adding more filters and 6) increasing filter size. In addition, using 7) larger training patch size and 8) augmentation of training subject lead to increase the performance of the networks. The adjustment of these 5 elements provides a similar improvement (about $0.5dB$). Although it seems natural to implement the deepest possible network, this parameter is not always the key to obtaining a better estimate of a high-resolution image. Our study shows that, depending on the type of input data (monomodal or multimodal), network depth is not necessarily the main parameter leading to better image reconstruction. In addition, it is necessary to take into account the time of the learning phase as well as the maximum memory available in the GPU in order to choose the best architecture of the network. For instance, for the monomodal SR case based on the simulations of Kirby dataset, we suggest using 20-layer networks with 64 small filters with size of 3^3 regarding 10 training subjects of size 25^3 to achieve practicable results.

In CNN-based approaches, the upscaling operation can be performed by using transposed convolution (so-called fractionally strided convolutional) layers in [Dong et al., 2016b, Oktay et al., 2016] or sub-pixel layers [Shi et al., 2016]. However, the pre-trained weights of these networks are totally optimized for a specified scale factor. This is a limiting aspect of CNN-based SR for MR data since a fixed upscaling factor is not appropriate in this context. In this study, we have presented a multi-scale CNN-based SR method for single 3D brain MRI that is capable of learning multiple scales by training full all isotropic scale factors due to an independent upsampling technique such as spline interpolation. Handling multiple scales is related to multi-task learning. The lack of flexibility of learned network architecture raises an open issue requiring further studies: how can we build a network that can deal with a set of observation models (*i.e.* multiple scales, arbitrary point spread functions, non uniform sampling, etc.)? For instance, when applying SR techniques in a realistic setting, the choice of the PSF is indeed a key element for SR methods and it depends on the type of MRI sequence. The shape of the PSF also depends on the trajectory in the k-space (cartesian, radial, spiral). Making the network independent from the PSF model (i.e. blind SR) would be a major step for its use in routine protocol. Further research directions could focus on

making more flexible CNN-based SR methods for greater use of these techniques in human brain mapping studies.

Evaluation of SR techniques is done on simulated LR images. However, one potential use of SR techniques would be to improve the resolution of isotropic data acquired in clinical routine. The Figure 2.23 shows upsampling results on isotropic T1-weighted MR images (the resolution was increased from $1 \times 1 \times 1mm^3$ to $0.5 \times 0.5 \times 0.5mm^3$). In this experiment, the applied network has been trained to increase image resolution from $2 \times 2 \times 2mm^3$ to $1 \times 1 \times 1mm^3$. Although quantitative results cannot be computed, visual inspection of reconstructed upsampled images tend to show the potential of this SR method. No external dataset has been used for these experiences. Thus, features learned at a lower scale ($2mm$ in this experiment) may be used to compute high-resolution images that could be used for fine studies of thin brain structures such as the cortex. Further work is required to investigate this aspect or self-super-resolution [Jog et al., 2016, Zhao et al., 2018] and more particularly the link with self-similarity based approaches [Huang et al., 2015a].



(a) Nearest-neighbor (b) Spline interpolation

(c) LRTV [Shi et al., 2015] (d) 20L-ReCNN

FIGURE 2.23: Illustration of SR results (01018-t1w of dataset NAMIC) with isotropic voxel upsampling. Original data with voxel size of $1 \times 1 \times 1\ mm^3$ is upsampled to size $0.5 \times 0.5 \times 0.5\ mm^3$. 20L-ReCNN is trained with the NAMIC dataset.

In this thesis, we have proposed a multimodal method for brain MRI SR using CNNs where a HR reference image of the same subject can drive the reconstruction process of the LR image. By concatenating these HR and LR images, the reconstruction of the LR one can be enhanced by exploiting the multimodality feature of MR data. As shown in previous works [Manjón et al., 2010a, Rousseau, 2008, Rousseau et al., 2010a], the use of HR reference can lead to significant improvements of the reconstruction process. However, unlike the monomodal setup, a deeper network does not lead to better performance within the experiments on

NAMIC dataset. Experiments from our study show that future work is needed to understand the relationship between network depth and the quality of HR image estimation.

Moreover, we have experimentally investigated the performances of CNN for generalizing on a different dataset (*"i.e. how a learned network can be used in another context"*). More specifically, our study illustrates how knowledge learned from one MR dataset is transferred to another one (different acquisition protocol and different scales). We have used Kirby and NAMIC datasets for this purpose. Although a slight decrease in performance can be observed, CNN-based approach can still achieve better performance than existing methods. These results tend to demonstrate the potential applications of CNN-based techniques for MRI SR. Further investigations are required to fully assess the possibilities of transfer learning in medical imaging context, and the contributions of fine-tuning technique [Tajbakhsh et al., 2016].

Finally, future research directions for CNN-based SR techniques could focus on other elements of the network architecture or the learning procedure. For instance, batch normalization (BN) step has been proposed by [Ioffe and Szegedy, 2015]. The purpose of a BN layer is to normalize the data through the entire network, rather than just performing normalization once in the beginning. Although BN has been shown to improve classification accuracy and decrease training time [Ioffe and Szegedy, 2015], we attempt to include BN layers into CNN for image SR but they do not lead to performance increase. Similar observations have been made in a recent SR challenge [Timofte et al., 2017]. From a geometrical point of view, BN does not appear as an important "operation" for regression [Rousseau and Fablet, 2018]. Moreover, while the classical MSE-based loss attempts to recover the smooth component, perceptual losses [Johnson et al., 2016, Ledig et al., 2017, Zhao et al., 2017] are proposed for natural image SR to better reconstruct fine details and edges. Thus, adding this type of layer (residual block) or defining new loss functions may be beneficial for MRI SR and may provide new directions for research.

In this study, we have investigated the impact of adding data (about 3200 patches per added subject of Kirby dataset) on SR performances through PSNR computation. It appeared that using more subjects sightly improves the reconstruction results in this experimental setting. However, further work could focus on SR-specific data augmentation by rotation and flipping, which is usually used in many works [Kim et al., 2016a, Timofte et al., 2016], and intensity variation to handle different contrast and bias field for improving algorithm generalization.

The practical applications of SR are demonstrated in the studies presented: image quality transfer from high-resolution experimental dataset to clinical neonatal low-resolution images and augmenting the performance of segmentation methods. Our CNN-based SR method shows clear improvements over interpolation, which is the standard technique to enhance image quality from visualisation by a radiologist. SR method is therefore an ideal replacement for interpolation.

Chapter 3

Simultaneous super-resolution and segmentation using a generative adversarial network: Application to neonatal brain MRI

Contents

3.1 Introduction

Long-term studies of the outcome of prematurely born infants have clearly documented that the majority of such infants may have significant motor, cognitive, and behavioral deficits. However, there is a limited understanding of the nature of the cerebral abnormality underlying these adverse neurologic outcomes. Magnetic Resonance Imaging (MRI) provides unique opportunities for in vivo investigation of the early developing human brain. However, the analysis of clinical neonatal brain MRI data remains challenging mainly due to low

anisotropic image resolution. Improving morphological data processing such as image resolution enhancement and brain segmentation, is a key-point to provide robust morphometry analysis tools to the community.

One of the first key components of the processing pipeline of clinical MRI data is the upsampling image estimation. Super-resolution (SR) is a post-processing technique that aims at enhancing the resolution of an imaging system [Greenspan, 2008]. SR is a challenging inverse problem; in particular the estimation of texture and details remains difficult. Recently, supervised deep learning-based techniques have shown great improvement over model-based approaches. In this context, applying 3D convolutional neural networks (CNNs) yields promising results for MRI data [Chen et al., 2018b, Pham et al., 2017a]. However, the use of ℓ_2-norm loss leads to smooth, unrealistic high resolution images [Johnson et al., 2016, Ledig et al., 2017]. Generative adversarial networks (GANs) have thus been proposed to estimate textured and sharper images [Chen et al., 2018a, Ledig et al., 2017].

Once the high resolution image reconstruction is performed, the implementation of an automatic segmentation robust approach is crucial for fine brain structure analysis [Makropoulos et al., 2017]. Segmenting thin structures such as the neonatal cortical gray matter remains difficult and the segmentation step is always considered separately from image reconstruction.

In this chapter, we propose an end-to-end GAN-based approach which can generate both the perceptually super-resolved image and a cortical segmentation map from a single low-resolution (LR) image. The proposed approach called SegSRGAN is both assessed on simulated data and real clinical data.

3.2 Method

3.2.1 Formulation of single image super-resolution

The objective of single image SR is to estimate a high-resolution (HR) image $\mathbf{X} \in \mathbb{R}^m$ from one observed LR image $\mathbf{Y} \in \mathbb{R}^n$. SR problem can be formulated using the following linear observation model:

$$\mathbf{Y} = H_{\downarrow} B\mathbf{X} + N = \Theta\mathbf{X} + N \tag{3.1}$$

where N is the additive noise, $B \in \mathbb{R}^{m\times m}$ is a blur matrix (depending on the point spread function), $H_{\downarrow} : \mathbb{R}^m \to \mathbb{R}^n$ is a downsampling decimation and $\Theta = H_{\downarrow} B \in \mathbb{R}^{n\times m} (m > n)$.

A popular approach that solves SR problem defines the matrix Θ^{-1} as the combination of a restoration operator $F \in \mathbb{R}^{m\times m}$ and an upscaling interpolation operator $S^{\uparrow} : \mathbb{R}^n \to \mathbb{R}^m$ computing the interpolated LR image $\mathbf{Z} \in \mathbb{R}^m$ ($\mathbf{Z} = S^{\uparrow}\mathbf{Y}$). In the context of supervised learning, given a set of HR images $\mathbf{X}_i$ and their corresponding LR images $\mathbf{Y}_i$, the restoration

operator F can be estimated by minimizing the following loss function:

$$\widehat{F} = \arg\min_{F} \sum_{i} \|\mathbf{X}_i - F(\mathbf{Z}_i)\|_2^2. \tag{3.2}$$

However, it is known that the use of ℓ_2-norm may lead to oversmoothing high resolution images. In order to provide realistic HR images, perceptual loss function [Johnson et al., 2016] have been used in a GAN [Ledig et al., 2017]. This is a paradigm shift since it is no longer a question of minimizing only the reconstruction error but of estimating a realistic image, i.e. a high resolution image that corresponds to the observation model with a realistic texture aspect.

A perceptual loss can be formulated as the weighted sum of the content loss (based, e.g., on pixel-wise mean squared error loss) and an adversarial loss component. In GAN-based approaches, the purpose is to train a generating network G that estimates for a given LR input image $\mathbf{Y}$ a corresponding HR image $G(\mathbf{Y})$. The goal of the discriminator network D is to classify real images $\mathbf{X}$ and simulated HR images $G(\mathbf{Y})$. The game between the generator G and the discriminator D is expressed as an adversarial loss:

$$\mathcal{L}_{adv} = \min_{G} \max_{D} \mathbb{E}_{\mathbf{X}\sim\mathbb{P}_{\mathbf{X}}}[logD(\mathbf{X})] + \mathbb{E}_{\mathbf{Y}\sim\mathbb{P}_{\mathbf{Y}}}[log(1 - D(G(\mathbf{Y})))] \tag{3.3}$$

where $\mathbb{P}_{\mathbf{X}}$ and $\mathbb{P}_{\mathbf{Y}}$ denote the data distribution of $\mathbf{X}$ and $\mathbf{Y}$ respectively.

3.2.2 Formulation of image segmentation

In this work, image segmentation is viewed as a supervised regression problem:

$$\mathbf{S}_{\mathbf{X}} = R(\mathbf{X}) \tag{3.4}$$

where R denotes a non-linear mapping from the upscaled image $\mathbf{X}$ to the segmentation map $\mathbf{S}_{\mathbf{X}}$. Similarly to the SR problem, assuming that we have a set of images $\mathbf{X}_i$ and corresponding segmentation maps $\mathbf{S}_{\mathbf{X}_i}$, a general approach for solving this segmentation problem is to find the mapping R by minimizing the following loss function:

$$\widehat{R} = \arg\min_{R} \sum_{i} \|\mathbf{S}_{\mathbf{X}_i} - R(\mathbf{X}_i)\|_2^2. \tag{3.5}$$

Unlike the SR problem, the use of ℓ_2-norm is less critical as it is expected to estimate smooth segmentation maps.

3.2.3 Joint mapping by generative adversarial networks

We propose the use of a GAN-based approach to estimate jointly a HR image and its corresponding segmentation map from one LR image. To this end, a convolution-based generator network G takes as input an interpolated LR image $\mathbf{Z}$ and computes a HR image $\widehat{\mathbf{X}}$ and a segmentation map $\widehat{\mathbf{S_X}}$ by minimizing the following reconstruction loss:

$$\mathcal{L}_{rec} = \min_{G} \sum_{i} \rho\left((\mathbf{X}, \mathbf{S_X})_i - G(\mathbf{Z}_i)\right) \tag{3.6}$$

where $(\mathbf{X}, \mathbf{S_X})_i$ are concatenated along the feature channel. In this work, we use a robust loss as Charbonnier loss [Charbonnier et al., 1997, Lai et al., 2017] :

$$\rho(x) = \sqrt{x^2 + \nu^2} \tag{3.7}$$

where ν is set to 10^{-3}.

The discriminator network D attempts to distinguish the real data $(\mathbf{X}, \mathbf{S_X})$ and the generated ones $G(\mathbf{Z})$. The game between the generator G and the discriminator D is usually modeled with a minimax objective as Equation (3.3).

However, using such loss function, GAN may be unstable or can suffer from mode collapse during training. Thus, in this work, we propose to use Wasserstein GAN loss described in [Gulrajani et al., 2017]:

$$\begin{aligned} \mathcal{L}_{adv} = \min_{G} \max_{D} \mathbb{E}_{\mathbf{X}\sim\mathbb{P}_{\mathbf{X}}, \mathbf{S_X}\sim\mathbb{P}_{\mathbf{S_X}}}[D((\mathbf{X}, \mathbf{S_X}))] - \\ \mathbb{E}_{\mathbf{Z}\sim\mathbb{P}_{\mathbf{Z}}}[D(G(\mathbf{Z}))] + \lambda_{gp}\mathbb{E}_{\widehat{\mathbf{XS}}}[(\| (\nabla_{\widehat{\mathbf{XS}}} D(\widehat{\mathbf{XS}}) \|_2 - 1)^2] \end{aligned} \tag{3.8}$$

where $\widehat{\mathbf{XS}}$ is the interpolation of the true data and the generated one as $(1-\epsilon)(\mathbf{X}, \mathbf{S_X}) + \epsilon G(\mathbf{Z})$, $\epsilon \sim \mathcal{U}[0, 1]$. λ_{gp} and ∇ denote the gradient penalty coefficient and gradient operator, respectively. The images $\mathbf{X}$, $\mathbf{S_X}$ and $\mathbf{Z}$ are extracted randomly from the data distributions of HR images $\mathbb{P}_{\mathbf{X}}$, HR segmentation maps $\mathbb{P}_{\mathbf{S_X}}$ and LR images $\mathbb{P}_{\mathbf{Z}}$. The terms $D((\mathbf{X}, \mathbf{S_X}))$, $D(G(\mathbf{Z}))$ and $D(\widehat{\mathbf{XS}})$ are the responses of the discriminator with respect to the real data, the generated data and the interpolated data, respectively. The full objective function is expressed as:

$$\mathcal{L}_{total} = \mathcal{L}_{rec} + \lambda_{adv}\mathcal{L}_{adv} \tag{3.9}$$

where λ_{adv} is a trade-off parameter between reconstruction loss and adversarial loss. Figure 3.1 illustrates our proposed GAN-based method for joint mapping of SR and cortex segmentation.

3.2.4 Architecture of generator and discriminator networks

The generator network (see Figure 3.2 (a)) is a convolution-based network with residual blocks. Let C^i_j-S^k be a block consisting of the following layers: a convolution layer of j filters

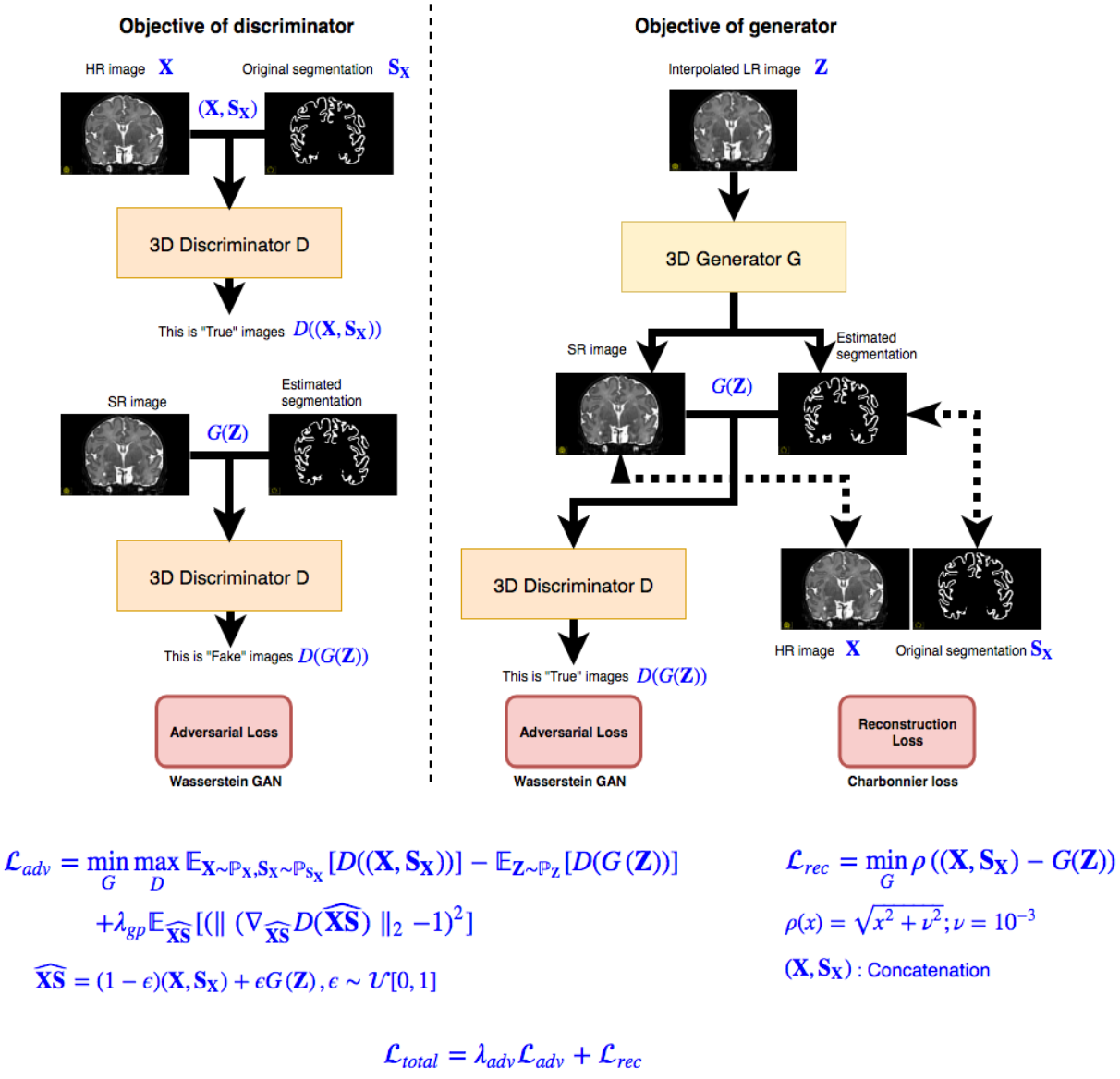
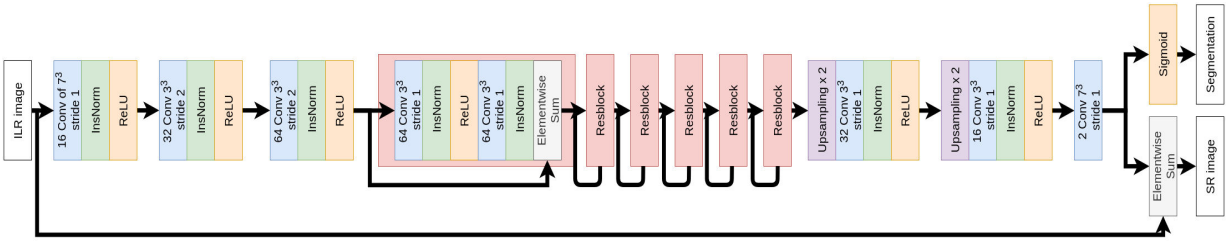


FIGURE 3.1: The illustration of our proposed 3D SegSRGAN for joint mapping of SR and segmentation.

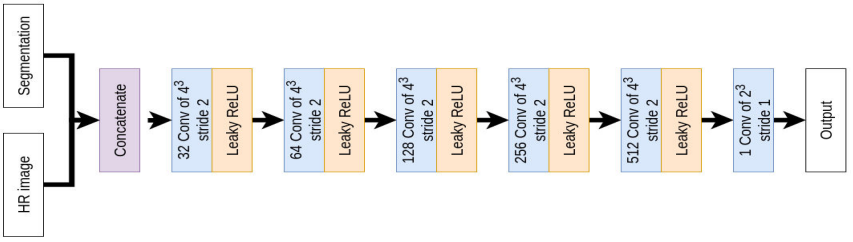
of size i^3 with stride of k, an instance normalization layer (InsNorm) [Ulyanov et al., 2017] and a rectified linear unit (ReLU).

R_k denotes a residual block as Conv-InsNorm-ReLU-Conv-InsNorm that contains 3^3 convolution layers with k filters. U_k denotes layers as Upsampling-Conv-InsNorm-ReLU layer with k filters of 3^3 and stride of 1. After the last layer, we apply a sigmoid activation for the channel of segmentation map and an element-wise sum of the channel of reconstruction and the interpolated LR image (residual-learning as in [Kim et al., 2016a, Pham et al., 2017b]). The generator architecture is: C^7_{16}-S^1, C^3_{32}-S^2, C^3_{64}-S^2, R_{64}, R_{64}, R_{64}, R_{64}, R_{64}, R_{64}, U_{32}, U_{16}, C^7_2-S^1.

The discriminator network (see Figure 3.2 (b)) contains five convolutional layers with an increasing number of filter kernels, increasing by a factor of 2 from 32 to 512 kernels. Let



(a) 3D generator



(b) 3D discriminator

FIGURE 3.2: The architecture of our proposed 3D SegSRGAN for joint mapping of SR and segmentation.

C_k be a block consisting of the following layers: a convolution layer of k filters of size 4^3 with stride of 2 and a Leaky ReLU with a negative slope of 0.01. The last layer C_1^2 is a 2^3 convolution filter with stride of 1. No activation layer is used after the last layer. The discriminator consists of C_{32}, C_{64}, C_{128}, C_{256}, C_{512}, C_1^2.

3.3 Experiments and Results

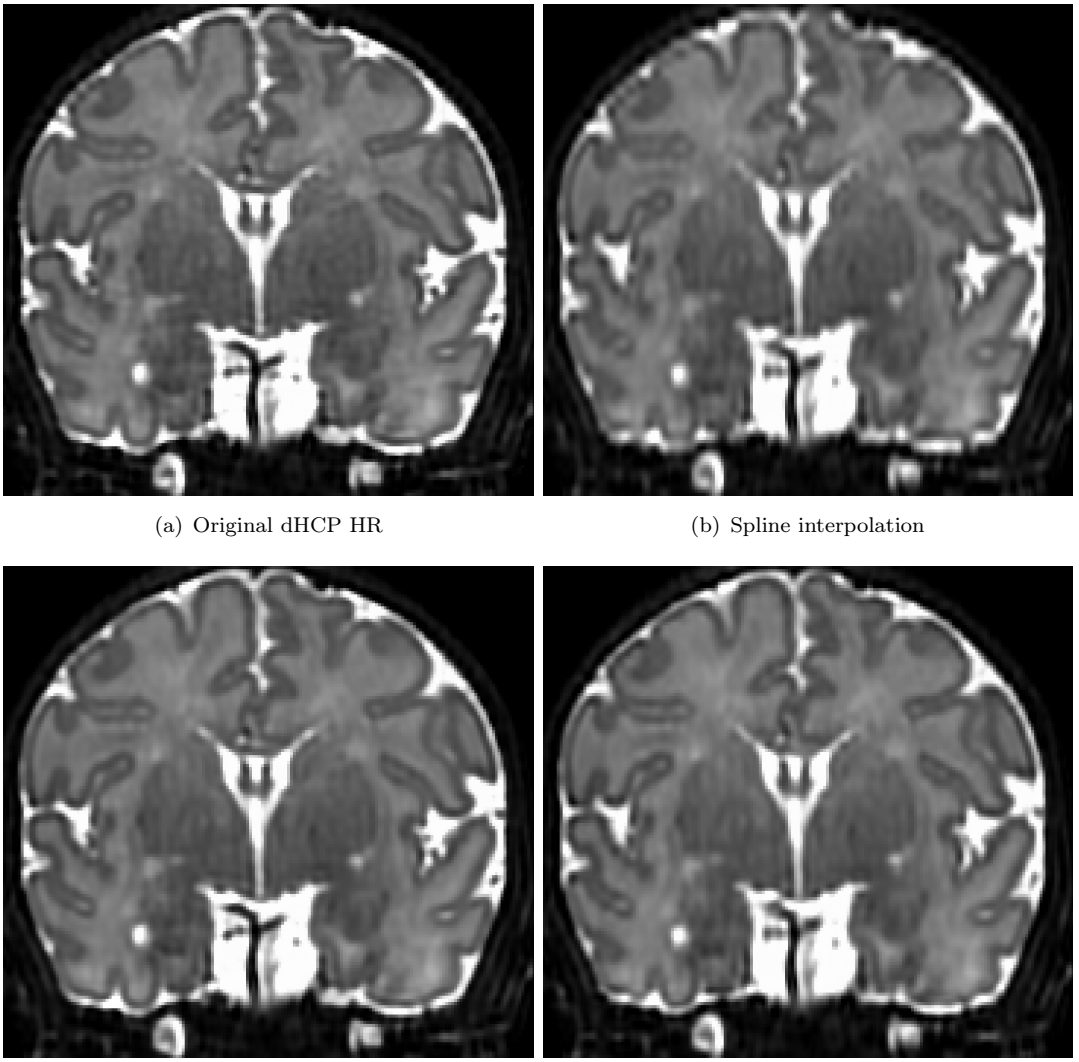
3.3.1 Datasets and network training

To assess the ability to reconstruct HR volume and segment the cerebral cortex , we applied the proposed method on T2-weighted (T2w) MR images of the developing Human Connectome Project[1] (dHCP). 40 T2w images were acquired using a 3T Achieva scanner with a $0.5 \times 0.5 \times 0.5$ mm^3 resolution with TR = 12 000 ms TE = 156 ms, respectively. 30 images were used for training networks, whereas the other 10 were used as testing images. As in [Greenspan, 2008], LR images were generated by using a Gaussian blur with the full-width-at-half-maximum (FWHM) set to slice thickness before a downsampling step to obtain a $0.5 \times 0.5 \times 1.5$ mm^3 resolution.

We have also applied the proposed method onto clinical neonatal MRI data acquired in the neonatology service of Reims Hospital. These LR images have a resolution of $0.446 \times 0.446 \times 3$ mm^3. 40 HR images of the dataset dHCP were filtered and downsampled as in [Greenspan, 2008] in order to generate LR images with a same resolution as clinical data. The network was trained using 40 pairs of simulated data and then applied to real LR images for visual evaluations. All data had bias correction and for network training, they were normalized between 0 and 1. No subjects nor image patches appear twice in the different subsets.

[1] http://www.developingconnectome.org

The 3D network was trained over 200 epochs with batch size of 16, using Adam method with learning rate of 0.0001 and updates the discriminator 5 times before training the generator as in [Gulrajani et al., 2017]. The parameters λ_{adv} and λ_{gp} were set to 0.001 and 10 respectively. The training patch size is 64^3. At test time, the whole HR image and segmentation volume were reconstructed by the weighted predictions of patches. A thresholding at 0.5 has been performed to obtain binary segmentation maps.



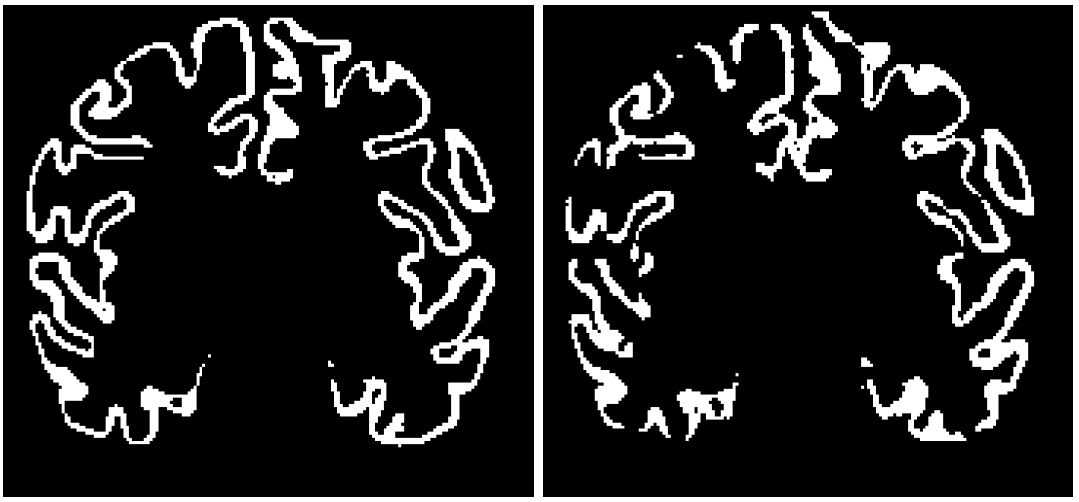
(a) Original dHCP HR (b) Spline interpolation

(c) 20L-SRReCNN [Pham et al., 2017b] (d) Proposed approach

FIGURE 3.3: SR results for one dHCP subject: (a) original HR image; (b–d) SR reconstruction of the LR image generated from (a) ©[2019] IEEE

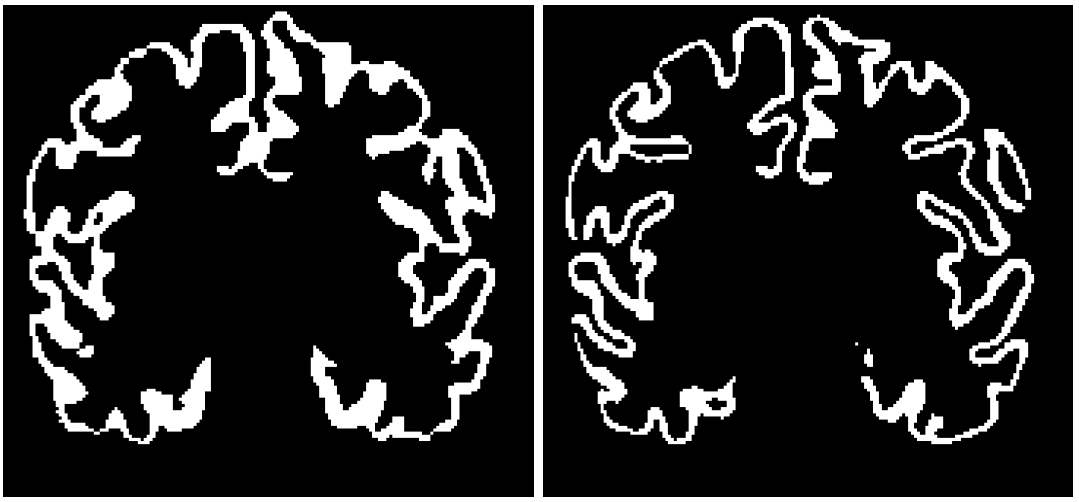
3.3.2 Results

Peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) have been used to evaluate the performance of SR reconstructions. Table 3.1 provides a summary of quantitative evaluations for the following methods: cubic spline interpolation, a 20-layers CNN-based SR approach (20L-SRReCNN) [Pham et al., 2017b] (described in Chapter 2) and our proposed



(a) Original dHCP segmentation

(b) IMAPA [Tor Díez et al., 2018]



(c) DrawEM [Makropoulos et al., 2014]

(d) Proposed approach

FIGURE 3.4: Segmentation results for one dHCP subject: (a) segmentation ground-truth of Figure 3.3 (a); (b,c) segmentation of Figure 3.3 (b); (d) HR segmentation from the LR image using the joint SegSR-GAN method ©[2019] IEEE.

TABLE 3.1: Quantitative evaluation of SR methods on dHCP dataset ©[2019] IEEE.

	Interpolation	20L-SRReCNN	SegSRGAN
PSNR	30.70	35.84	31.75
SSIM	0.9492	0.9739	0.9624

TABLE 3.2: Quantitative evaluation of segmentation methods on dHCP dataset ©[2019] IEEE.

	IMAPA	DrawEM	SegSRGAN
Dice (mean)	0.788	0.818	0.886
Dice (standard deviation)	0.061	0.014	0.011

SegSRGAN. It can be seen that 20L-SRReCNN provides highest PSNRs as in [Johnson et al., 2016, Ledig et al., 2017] since this approach minimizes a ℓ_2-norm-based loss. However, while the two CNN-based approaches (20L-SRReCNN and SegSRGAN) lead qualitatively to similar

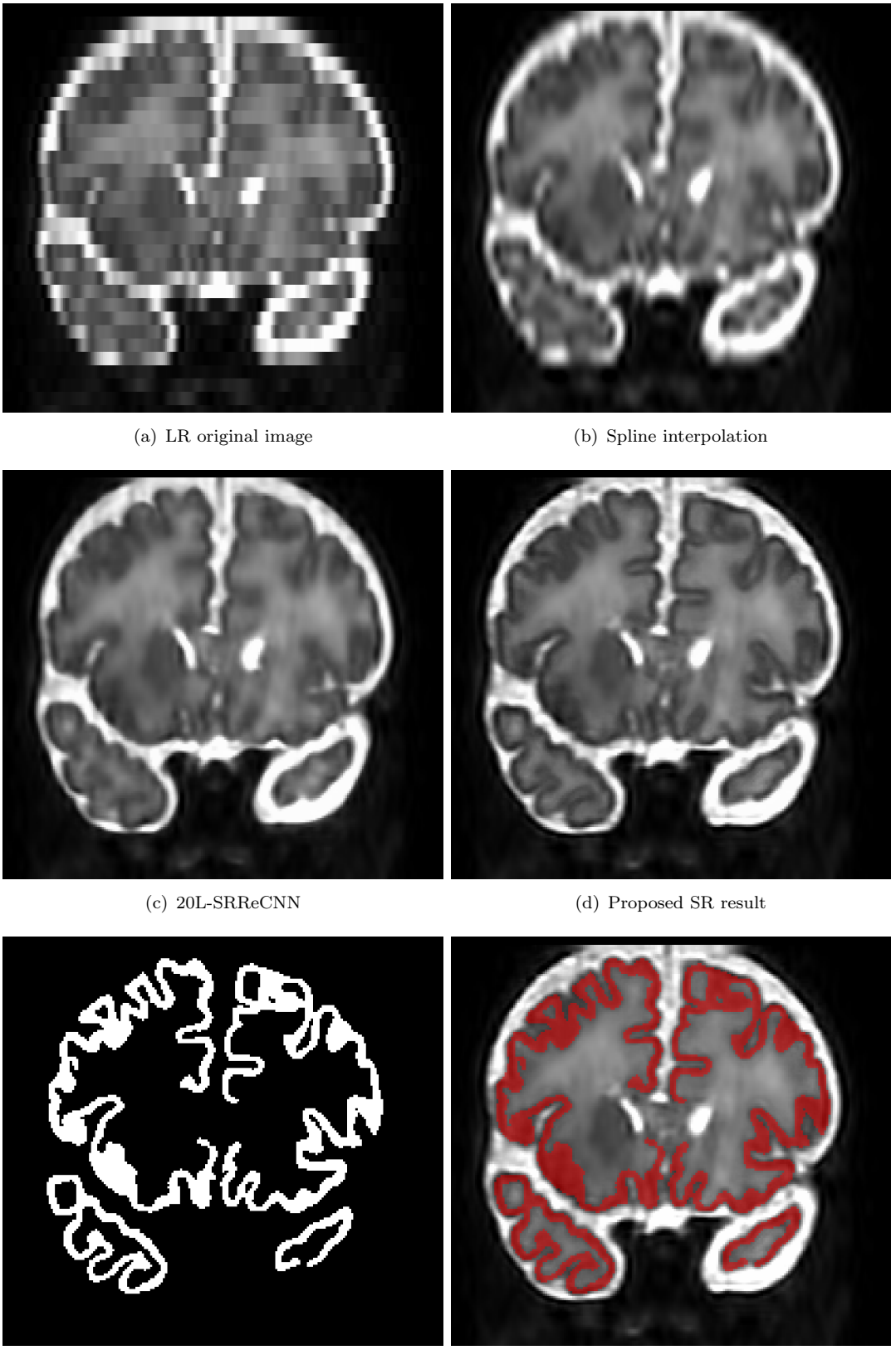
realistic results on dHCP dataset (see Figures 3.3 and 3.4), the proposed approach provides best reconstructed HR images on clinical data with better contrast on cortical gray matter (see Figure 3.5).

The Dice index is used to evaluate the cortical segmentation maps obtained by the following state-of-the-art methods: iterative multi-atlas patch-based approach (IMAPA) [Tor Díez et al., 2018], DrawEM [Makropoulos et al., 2014] and the proposed SegSRGAN. As in a typical clinical setting, the three methods have been applied on interpolated images. Table 3.2 shows that quantitatively the proposed approach lead to the best cortical segmentation results with significant improvement with respect to the two other methods. Moreover, as mentioned in [Tor Díez et al., 2018], the use of IMAPA applied on original HR dHCP images leads to a mean DICE of 0.887 (standard deviation of 0.011) that is very similar to the results obtained with SegSRGAN (applied on interpolated images).

As indicated in Section 2.2.3.2, we would like to evaluate the impact of upsampling methods for clinical LR T2w images with respect to segmentation methods. There are the manual HR segmentations of T1w images (ground truths). The estimated segmentation maps applied to SR results are mapped onto the original T1w images by a rigid registration between estimated HR T2w and original T1w data. Table 3.3 shows the segmentation results of the method MANTIS [Beare et al., 2016] for HR reconstructions of upsampling methods following: interpolation, NMU, 20L-SRReCNN and our SR results of SegSRGAN. The mean dice of the segmentation maps of MANTIS for our estimated HR image is better than the ones of compared upsampling methods. Moreover, we apply the supervised segmentation method (IMAPA) [Tor Díez et al., 2018] for the estimated isotropic T2w images using above upsampling methods (show in Table 3.3). Our proposed method uses the same training dataset of segmentation atlases as IMAPA. Table 3.3 shows that our SR results support other segmentation methods better than compared SR methods. In addition, our segmentation results also give comparable dice scores as the pipeline of 20L-SRReCNN and IMAPA. Results on real LR data (see Figures 3.5 and 3.6, Tables 3.3 and 3.4) confirm the potential of the proposed approach for fine analysis of clinical neonatal brain MRI.

Subject	T1w		T2w				
	Original	Interp.	Original	Interp.	NMU	20L-SRReCNN	Our SR results
MAIA #1	0.6215	0.6205	0.7090	0.7052	0.7190	0.7330	0.7480
MAIA #2	0.6746	0.6802	0.6694	0.7118	0.7182	0.7333	0.7333

TABLE 3.3: Dice scores of the segmentation method MANTiS on the 2 images of the MAIA testing dataset with respect to different approaches (columns): original T1w images, interpolated (Interp.) T1w images, original T2w images, upsampling T2 images using interpolation, NMU, 20L-SRReCNN and our SR results of the proposed SegSRGAN



(a) LR original image

(b) Spline interpolation

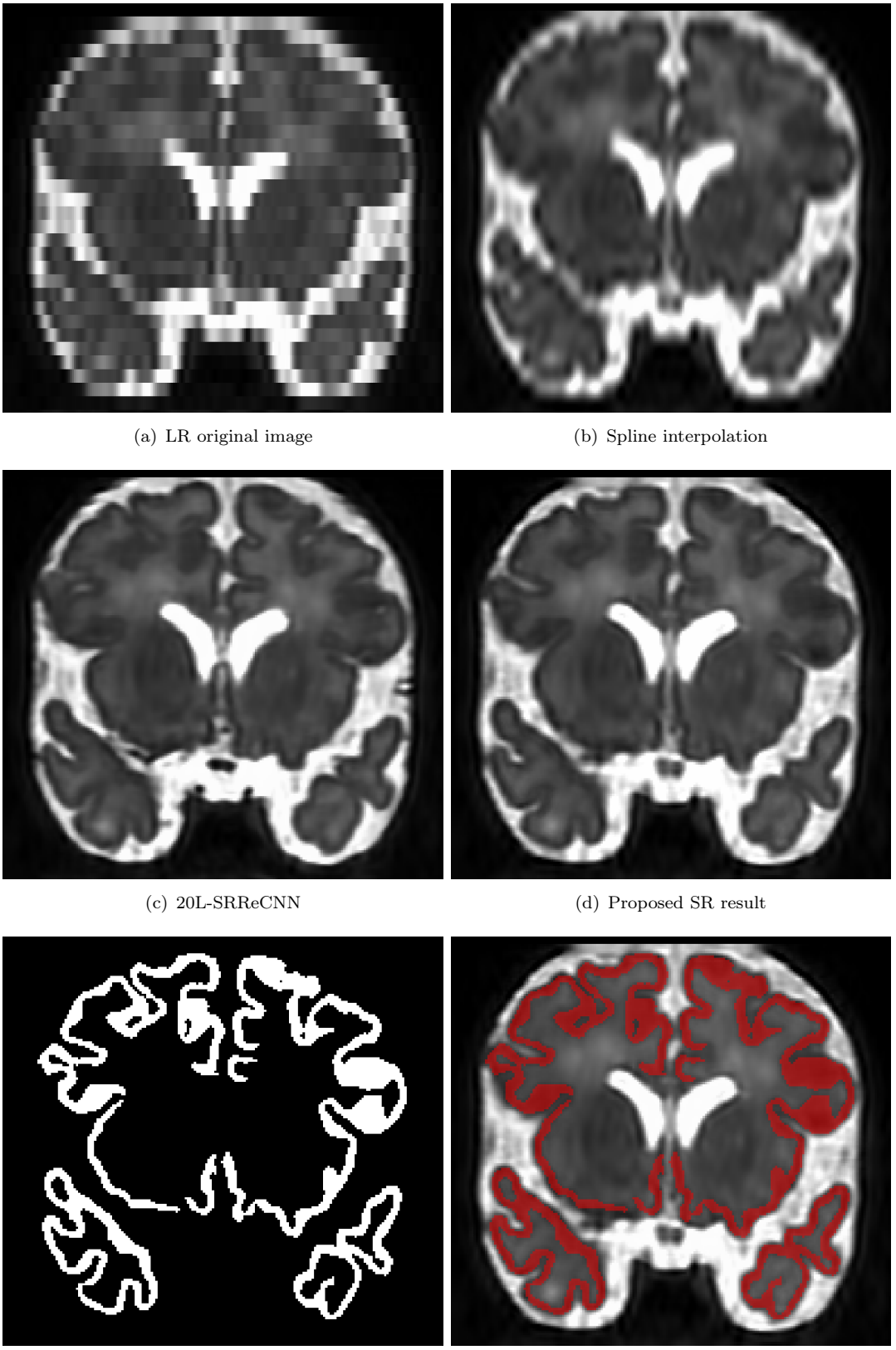
(c) 20L-SRReCNN

(d) Proposed SR result

(e) Proposed segmentation result

(f) Proposed approach

FIGURE 3.5: Reconstruction (b–d) and segmentation results (e) on a real LR neonatal brain image (a) (Subject S00059 of MAIA dataset) with voxel size of $0.446 \times 0.446 \times 3$ mm^3, re-sampled to $0.5 \times 0.5 \times 0.5$ mm^3.



(a) LR original image

(b) Spline interpolation

(c) 20L-SRReCNN

(d) Proposed SR result

(e) Proposed segmentation result

(f) Proposed approach

FIGURE 3.6: Reconstruction (b–d) and segmentation results (e) on a real LR neonatal brain image (a) (Subject S00096 of MAIA dataset) with voxel size of $0.446 \times 0.446 \times 3$ mm^3, re-sampled to $0.5 \times 0.5 \times 0.5$ mm^3.

Subject	IMAPA				Our proposed segmentation
	Interp.	NMU	20L-SRReCNN	Our SR results	
MAIA #1	0.6394	0.6551	0.6698	0.6945	0.6702
MAIA #2	0.6443	0.6497	0.6763	0.6943	0.6658

TABLE 3.4: Dice scores of the supervised segmentation method IMAPA (using the same training dataset with our method, the same segmentation protocol) on the 2 images of the MAIA testing dataset with respect to different approaches (columns): interpolated T2w images, upsampling T2 images using NMU, 20L-SRReCNN and our SR results (SegSR-GAN), and our proposed segmentation map of interpolated T2w images (SegSRGAN).

3.4 Discussion

In this chapter, we have presented a simultaneous super-resolution and segmentation method for 3D brain MR images using a generative adversarial network. Our experiments on both simulated and clinical data have shown that better performance can be achieved by this joint approach compared to state-of-the-art techniques, opening up new perspectives in the processing of clinical LR neonatal brain MRI data.

We have investigated that our proposed GAN-based method is more robust than the CNN-based approach. The CNN-based method achieves the highest PSNR/SSIM because it attempts to minimize the pixel-wise difference between super-resolved images and reference HR images using ℓ_2-norm cost function. This is reasonable as we have presented in Chapter 2. However, CNN-based methods are restricted to the predetermined condition of specific training data and their performance is then decreased when testing real images, where these conditions are not satisfied (also mentioned in [Shocher et al., 2018]). Meanwhile, the SR method using GAN attempts to minimize the difference of the texture between generated images and ground truth HR counterparts using the adversarial loss. This loss makes networks more robust to simulated training data. Future work is required to explore new quality metrics to evaluate better the performance of SR methods.

Our proposed method illustrated that the learned model from high-resolution experimental dataset can be transferred successfully to another low-resolution clinical dataset in order to enhance the image quality. We have used dHCP and MAIA dataset for this purpose. These results demonstrate the potential of GAN-based techniques for practical applications of medical image processing. We believe that our proposed approach can be used to another tasks such as medical image synthesis or other types of segmentation maps such as cerebrospinal fluid or ventricles in brain MRI.

In this study, we assume the paired training dataset, where input images have output counterparts (e.g. LR images and corresponding HR images). In some clinical cases, paired couples are not always available (e.g. T2w images with a specific resolution and the segmentation maps of T1w images with another resolution), that raises the question of self-supervised techniques for mapping of unpaired training dataset.

Chapter 4

Brain MRI cross-modal synthesis of subject-specific scans

Contents

4.1 Introduction

There are many medical imaging modalities in the clinical context such as: radiography, magnetic resonance imaging (MRI), computed tomography (CT) scan, ultrasound. Each modality shows up the physical properties of tissue in organs and special abnormalities for detecting different diseases. The diversity of medical image modalities is useful for diagnosticians but can be a challenge for automated image analysis. In clinical scenarios, the number of tissue contrasts that can be acquired is limited because of time consuming or expensive cost. Collecting all medical images of one subject is impractical. Cross-modal synthesis without real acquisition is considered as an intensity transformation applied to given input images of a source modality to generate new images with a specific tissue contrast. Synthetic images are not intended to be used for diagnostic purposes. Synthesis of a medical image can be used for a preprocessing step before applying more complex image processing algorithms. The objective of cross-modal synthesis is to generate images that are close enough approximations to real images so as to improve automated image processing. Cross-modality synthesis of medical images is proposed for many application such as segmentation [Iglesias et al., 2013, Roy et al., 2010], super-resolution [Pham et al., 2017b, Rousseau, 2008, Rueda et al., 2013], and multimodal registration [Roy et al., 2013, Wein et al., 2008]. The thesis in [Cordier, 2015] shows a review of the annotated data, which can be used to augment the performance of medical image analysis methods for pathological cases.

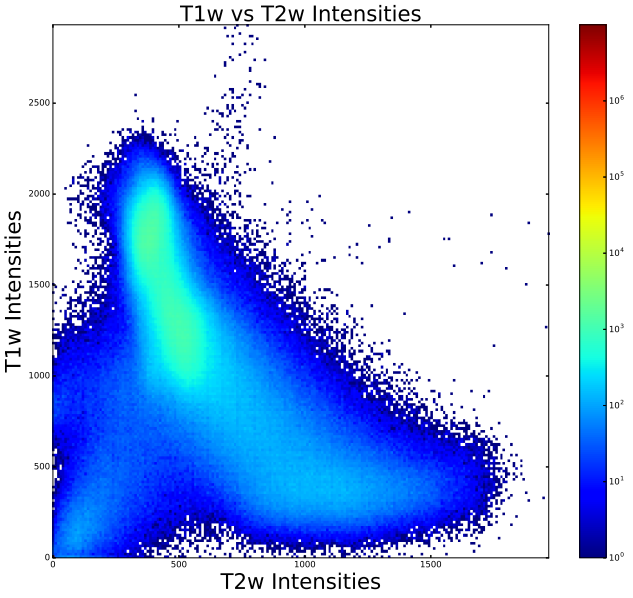
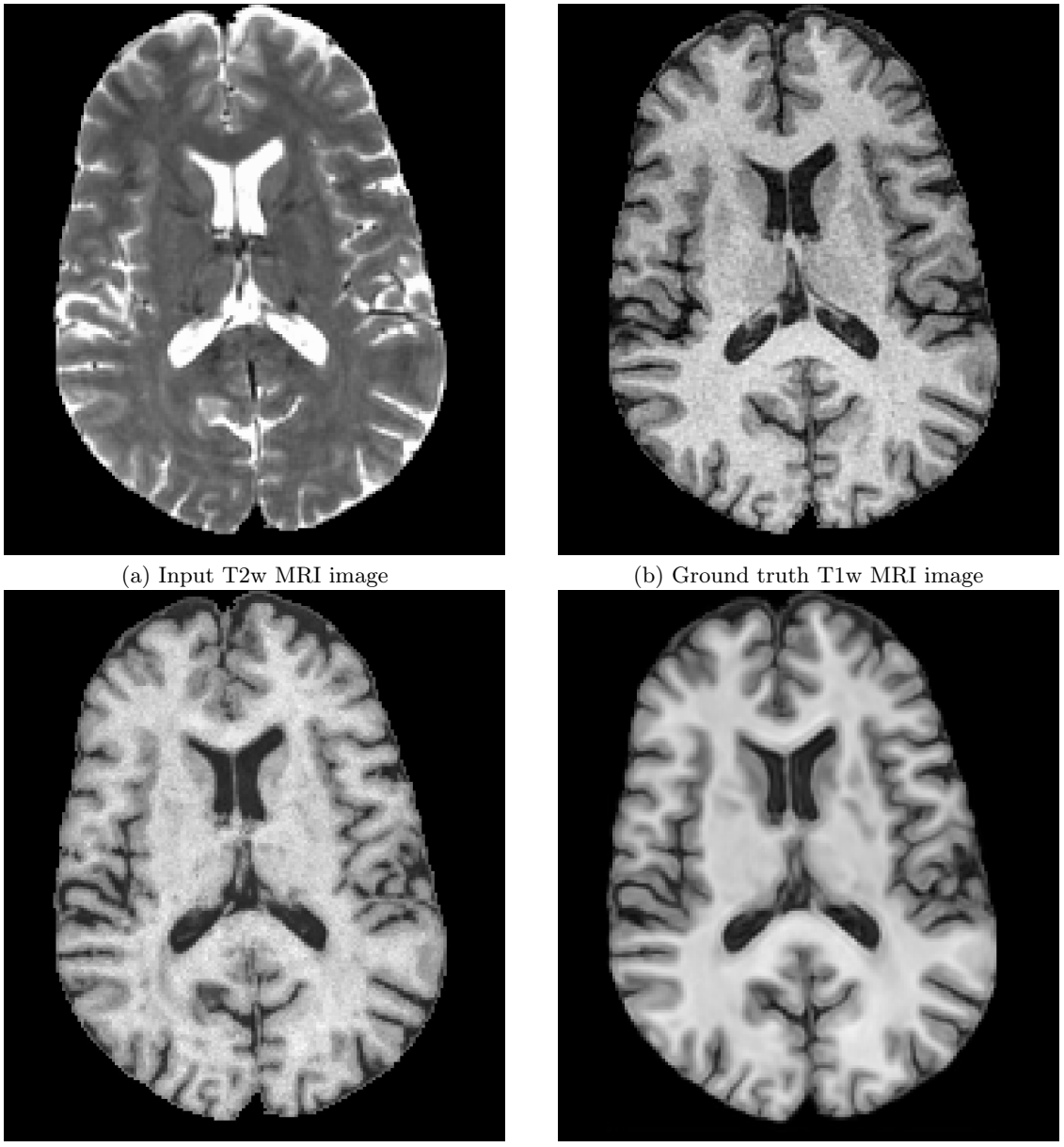


FIGURE 4.1: 2D histogram of intensity correspondences between paired T1w and T2w MRI over an entire image of the same subject form dataset NAMIC. Higher density regions is indicated by brighter color. The figure shows that the relationship between two modalities is not only non-linear but also not unique. It does not exist a function to transform from one T1w image to one T2w image and vice versa.

A statistical model of cross-modal synthesis can be expressed as:

$$\mathbf{Y} = R\mathbf{X} + N \tag{4.1}$$



(a) Input T2w MRI image

(b) Ground truth T1w MRI image

(c) REPLICA [Jog et al., 2017]

(d) 20-layers SRReCNN [Pham et al., 2017b]

FIGURE 4.2: The examples (i.e. the axial slices of a brain MRI) of cross-modal synthesis methods. The input T1w MRI image (a) is synthesized by the random-forest MRI synthesis method REPLICA [Jog et al., 2017] and SRReCNN [Pham et al., 2017b].

where R is a mapping, $\mathbf{Y}$ and $\mathbf{X}$ denote images of source and target domains and N is an additive noise. Figure 4.1 illustrates the intensity of a T1-weighted MR (T1w) image and the corresponding T2-weighted MR (T1w) image (shown Figure 4.2 (a) and (b)) of the same subject. These paired images of the same subject are acquired by the same imaging system with the share the same resolution, orientation, coordinate and the same number of voxels. Despite of paired images, the relation between the T1w and T2w tissue contrasts is totally non-linear as several regions share the opposite gradients but some regions are otherwise. One T2w intensity can be transformed from multiple T1w intensities and vice versa. Figure 4.2 shows synthesized T2 weighted MR (T2w) images (Figure 4.2 (c) and (d)) from a T1 weighted MR (T1w) image (Figure 4.2 (a)). The synthetic images are estimated as closely as

possible to their ground truth.

4.1.1 Paired cross-modal synthesis

The synthesis techniques have been studied in the context of medical imaging analysis using joint histogram [Kroon and Slump, 2009]. Given a dataset with the coupled images of source domain and target domain $\{(\mathbf{Y}_i, \mathbf{X}_i)\}$, the patch-based synthesis method [Iglesias et al., 2013] finds k-nearest neighbor patches $\mathbf{y}_k$ in the training base of the patch $\mathbf{y}$ of observed image $\mathbf{Y}$ as:

$$(\hat{k}, \hat{\mathbf{y}}_k) = \underset{k, \mathbf{y}_k \in \{\mathbf{Y}_i\}}{\operatorname{argmin}} \|\mathbf{y}_k - \mathbf{y}\|_2 \tag{4.2}$$

When the paired set $\{\hat{\mathbf{y}}_k, \hat{\mathbf{x}}_k\}$ is found, the synthesized patch is the average of k optimal patches $\hat{\mathbf{x}}_k$ as: $\hat{\mathbf{x}} = \sum_k \hat{\mathbf{x}}_k$. The patch-based method is improved by the iterative approach [Ye et al., 2013] as:

$$(\hat{k}^{t+1}, \hat{\mathbf{y}}_k^{t+1}) = \underset{k, \mathbf{y}_k^t \in \{\mathbf{Y}_i\}, \mathbf{x}_k^t \in \{\mathbf{X}_i\}}{\operatorname{argmin}} (1-\alpha)\|\mathbf{y}_k^t - \mathbf{y}^t\|_2 + \alpha\|\hat{\mathbf{x}}_k^t - \hat{\mathbf{x}}^t\|_2 \tag{4.3}$$

where, $\hat{\mathbf{x}}^t$ is the synthesized image by the optimal corresponding patches $\hat{\mathbf{x}}_k^t$ at the t^{th} iteration, α denotes the trade-off between two terms. Instead of ℓ_2-norm patch-based approaches, the regression tree method is proposed for synthesis MRI contrasts in [Jog et al., 2013] to find the complex mapping between modalities. Similarly, the improved versions of this technique as random forest decision can be found in [Huynh et al., 2016, Jog et al., 2014, 2017]. In parallel, [Roy et al., 2013, Ye et al., 2013] adapts the sparse-coding-based methods for SR as in [Yang et al., 2010] for synthesis MRI contrasts, assuming jointly dictionaries for T1w and T2w MR images. The multi-layer neural network for cross-domain synthesis is first proposed in [Van Nguyen et al., 2015] (LSDN) for mapping the intensity feature and the spatial coordinates from the input domain to the intensity of target domain as:

$$\min \|\Phi(F\mathbf{Y}_i, P\mathbf{Y}_i) - \mathbf{X}_i\|^2 \tag{4.4}$$

where Φ represents the network, F and P denotes intensity-based feature extractions and spatial informations respectively. Instead of pooling layers for spatial-based voxel connections as CNNs, the method LSDN proposes multiplication operations between layers and the shrinking connection at each layer for reduced the computation cost. The 2D U-net architecture [Ronneberger et al., 2015] (shown in Figure 4.3) is applied to generate CT from discontinuous MRI slices in [Han, 2017]. [Nie et al., 2016] proposes to use fully 3D fully convolutional neural networks inspired by [Dong et al., 2016a] for reconstructing CT scans from MRI volumes. The improved versions of this network in [Nie et al., 2017, 2018] exploit image gradient difference loss and adversarial loss with auto-context model. The process of auto-context model (shown in Figure 4.4) is to refine the synthesized images via an iterative process between input images and the estimated images at each iteration using different training models. Recently, the work in [Xiang et al., 2018] synthesizes the MRI consecutive

axial slices into CT scans using 2D embedding CNNs in which the reconstruction stage and the transform stage are concatenated. The stages are expressed as:

$$\begin{cases} F_{tran,i} = PReLU(W_{tran,i} \star F_{tran,i-1} + B_{tran,i}) \\ F_{rec} = W_{rec} \star F_{tran,j} + B_{rec} \end{cases} \tag{4.5}$$

where F_{rec} and $F_{tran,i}$ denotes the estimated synthesis and the response of the i^{th} layer respectively. $W_{tran,i}$,$B_{tran,i}$, W_{rec} and B_{rec} are network parameters and $\star$ denotes convolution operation. Then, an embedding block is defined as a concatenation of these two stages before a transform stage which maintains the number of response at each layer.

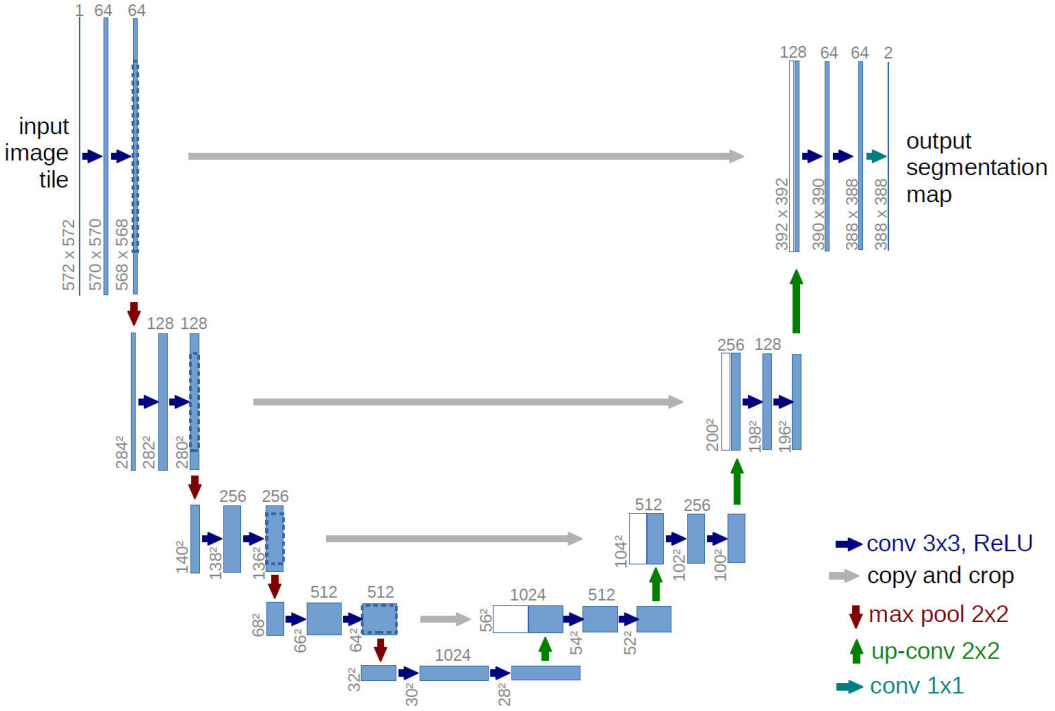


FIGURE 4.3: U-net architecture [Ronneberger et al., 2015]

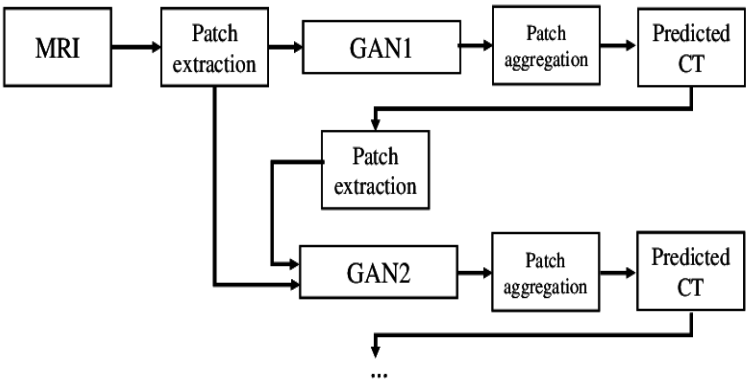
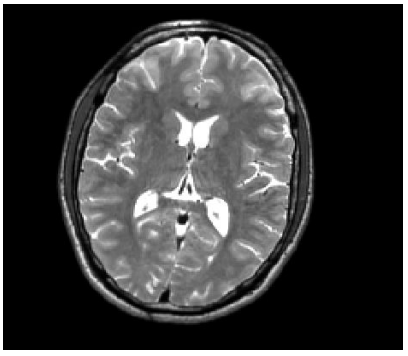


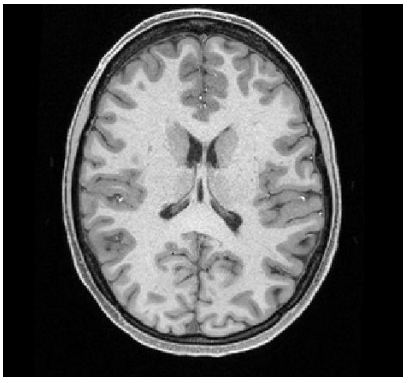
FIGURE 4.4: The architecture for auto-context with generative adversarial networks [Nie et al., 2018]

4.1.2 Unpaired cross-modal synthesis

Coupled training set of one subject is not always available. Figure 4.5 shows an example where the T1w and T2w images are unpaired. These unpaired images of different subjects (of different datasets) are acquired with different resolutions. Thus, they do not share the same general structure. This raises a question about the ability of synthesizing the T2w image of the observed T1w image (e.g. shown in Figure 4.5 (a)) given an unpaired T2w image (e.g. shown in Figure 4.5 (b)).



(a) T1w axial slice (Subject: 01011-t1w of the dataset NAMIC). The voxel size of the images is $1 \times 1 \times 1$mm.



(a) T2w axial slice (Subject: 100307 of the dataset HCP100). The voxel size of the images is $0.7 \times 0.7 \times 0.7$mm.

FIGURE 4.5: Adult brain MRIs of different subjects

Unpaired synthesis methods have recently investigated in [Huang et al., 2017b, 2018, Vemulapalli et al., 2015]. The method in [Vemulapalli et al., 2015] proposes 3 steps to handle unpaired data synthesis. Firstly, a set of patch-based nearest neighbour candidates of the source image is generated using mutual information $MI()$ of the source patches $\mathbf{y}$ and the target patches $\mathbf{x}$ as:

$$MI(\mathbf{y};\mathbf{x}) = H(\mathbf{x}) - H(\mathbf{x}) + H(\mathbf{y} \mid \mathbf{x}) \tag{4.6}$$

where $H(\mathbf{x})$ and $H(\mathbf{y})$ are the marginal entropies and $H(\mathbf{y} \mid \mathbf{x})$ denotes the conditional entropy. The second step attempts to synthesis the source image $\mathbf{Y}$ using best candidates by maximizing the cost function as:

$$\begin{aligned} \max_{w_{vk}} H(\mathbf{X}) - H(\mathbf{X}) + H(\mathbf{Y} \mid \mathbf{X}) + \lambda SC(\mathbf{X}, \mathbf{Y}) \\ s.t. \sum_{k} w_{vk} = 1,\ v \in V \end{aligned} \tag{4.7}$$

where $SC(\mathbf{X}, \mathbf{Y})$ is a regularization term that promote spatial consistency between the neighbour candidates and V denotes two neighboring voxels. Finally, coupled sparse representation of source modality image and the synthesized target modality image is calculated to refine the result of the preceding steps as super-resolution problems [Wang et al., 2015, Yang et al., 2010, Zeyde et al., 2012] as:

$$\hat{\alpha} = \underset{\alpha}{\operatorname{argmin}} \|\alpha\|_1 \, s.t. \, \|D_{\mathbf{y}}\alpha - \mathbf{y}\|_2^2 \leq \epsilon_1, \|D_{\hat{\mathbf{x}}}\alpha - \hat{\mathbf{x}}\|_2^2 \leq \epsilon_2 \tag{4.8}$$

where $D_{\hat{\mathbf{x}}}$ and $D_{\mathbf{y}}$ denote the joint dictionaries for the patches of synthesized and source domains and α is the sparse support. Instead of learning dictionaries of synthesized and source images as [Vemulapalli et al., 2015], improved sparse coding methods based on target/source images for synthesis with convolution representation and regularizations can be found in [Huang et al., 2017b, 2018]. The work in [Huang et al., 2018] proposes to synthesize MRI contrasts using sparse representations and two regularizers as maximum mean discrepancy and geometry preservation based on a few pairs of data. The unpaired couple dictionaries of target and source domain are learned from the sparse representation as:

$$\begin{aligned}\min_{\alpha,D} \mathcal{L}(\alpha, \mathcal{D}, \mathbf{y}, \mathbf{x}) = \min_{\alpha,D} \|D_{\mathbf{y}}\alpha_{\mathbf{y}} - \mathbf{y}\|_2^2 + \|D_{\mathbf{x}}\alpha_{\mathbf{x}} - \mathbf{x}\|_2^2 + \lambda\|\alpha_{\mathbf{y}}\|_1 + \lambda\|\alpha_{\mathbf{x}}\|_1 \\ +\tau\mathcal{F}(\alpha_{\mathbf{y}}, \alpha_{\mathbf{x}}) + \gamma MMD(\alpha_{\mathbf{y}}, \alpha_{\mathbf{x}}) + \mu Geo(\alpha_{\mathbf{y}}, \alpha_{\mathbf{x}})\end{aligned} \tag{4.9}$$

where $D_{\mathbf{x}}$ is now the dictionary of target domain, $\alpha = \{\alpha_{\mathbf{y}}, \alpha_{\mathbf{x}}\}$ denotes the sparse code, and MMD and Geo denotes maximum mean discrepancy regularization and geometry co-regularization. In order to ensure the identity of the sparse codes from the source to the target, we assume the linear projection in the common feature space via a mapping function as $\mathcal{F}(\alpha_{\mathbf{y}}, \alpha_{\mathbf{x}})$. However, the method needs the pair training data to constraint the unpaired image data by the fact that they must share the same high-frequency features. Equation (4.9) is rewritten by adding the objective function on these few pairs:

$$\min_{\alpha,D} \mathcal{L}(\alpha, \mathcal{D}, \mathbf{y}, \mathbf{x}) + \|F_{HF}\mathbf{x} - \hat{T}F_{HF}\mathbf{y}\|_2^2 \tag{4.10}$$

where F_{HF} is the high-frequency feature extractor and $\hat{T}$ denotes the binary matrix which consists of one element of 1 and other set to be 0. The 1 element is set to the maximum value of an affinity matrix which consists of measured distances of paired patches.

4.1.3 Discussion

A brief review of cross-modal synthesis for medical imaging has been described. The learning-based methods such as patch-based techniques, sparse coding, random forest and CNN-based are commonly used for paired cross-modal synthesis. In the context of supervised learning, techniques proposed for image synthesis have the same point of view as example-based learning SR methods. The availability of paired modalities of the same subject is sometimes lacking. Unpaired cross-modal synthesis are proposed to overcome this disadvantage. However, the need of few paired training images is inevitable for the refinement of synthetic results. In the next sections, our CNN-based methods for SR is applied to paired image synthesis. Mostly, we attempt to propose an approach to totally unpaired MRI synthesis using generative adversarial networks.

4.2 Supervised synthesis with convolutional neural networks

4.2.1 Mathematical formulation

In the context of supervised learning, assuming that a training data set which consists of pairs of images in a source modality $\mathbf{Y}_j$ (e.g. T1w images) and the corresponding images in the target modality $\mathbf{X}_j$ (e.g. T2w images), our objective is to find the mapping f that optimize the cost function as:

$$\hat{f} = \arg\min_{f} \sum_{j} \rho(f(\mathbf{Y}_j) - \mathbf{X}_j) \tag{4.11}$$

where ρ can be ℓ_1 or ℓ_2-norm for instance. The convolution neural networks, which are described in the chapter 2, are directly applied to solve our synthesis problem. The mapping f from $\mathbf{Y}_i$ to the residual $(\mathbf{X}_j - \mathbf{Y}_j)$ is decomposed into nonlinear operations as the combination of convolutional layers with the ReLU activation as:

$$\hat{f} = \arg\min_{f} \sum_{j} \|f(\mathbf{Y}_j) - (\mathbf{X}_j - \mathbf{Y}_j)\|^2 \tag{4.12}$$

Residual learning strategies make the convergence of CNNs faster. In principle, residual connections induce the responses of layers to be close to zeros, making the network easier to train. The interest of residual learning is also proposed in [Nie et al., 2018]. The architecture of our networks can be described as follows:

$$\begin{cases} f_1(\mathbf{Y}) = \max(0, W_1 * \mathbf{Y} + B_1) \\ f_i(\mathbf{Y}) = \max(0, W_i * F_{i-1}(\mathbf{Y}) + B_i) \quad for \quad 1 < i < L \\ f_L(\mathbf{Y}) = W_L * F_{L-1}(\mathbf{Y}) + B_L \end{cases} \tag{4.13}$$

where L is the number of layers. Once the training step is done, the synthesized image of a given image is estimated as $\mathbf{X} = \hat{f}(\mathbf{Y}) + \mathbf{Y}$.

4.2.2 Dataset and training parameters

We use T1w and T2w MR images of NAMIC Brain Multimodality [1] to assess the ability of our CNN-based method (20L-SRReCNN). These data have been acquired using a 3T GE. The T1w images are acquired in contiguous spoiled gradient-recalled acquisition (fastSPGR) with the following parameters: TR=7.4ms, TE=3ms, TI=600, 10 degree flip angle, 25.6cm^2 field of view, matrix=256×256. The contiguous T2-weighted images are acquired with the following parameters: TR=2500ms, TE=80ms, 25.6 cm^2 field of view, 1 mm slice thickness. Voxel dimensions of these images are $1 \times 1 \times 1mm^3$.

[1] NAMIC : http://hdl.handle.net/1926/1687

We use a series of 19 convolution layers of $3 \times 3 \times 3$ with 64 filters and the ReLU activations. The last layer is a $3 \times 3 \times 3$ convolution layer with one filter. ADAM method is used to optimize the network with 20 epochs (batch size of 64).

4.2.3 Experimental results

	REPLICA [Jog et al., 2017]		our 20L-SRReCNN	
	PSNR	SSIM	PSNR	SSIM
Synthesized T1w	13.3255	0.9444	15.6848	0.9584
Synthesized T2w	17.8343	0.9542	20.5420	0.9528

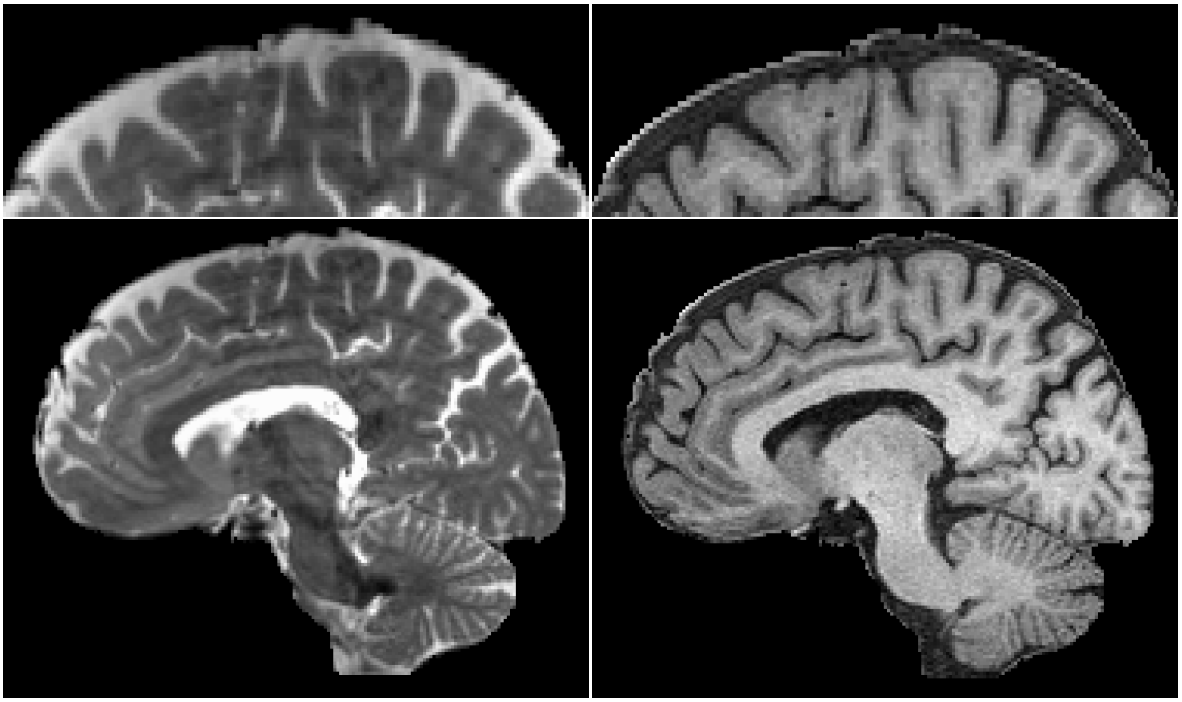
TABLE 4.1: The results of PSNR/SSIM for cross-modal synthesis methods of subject-specific scans. All methods using the training and testing images of NAMIC.

In this section, we study performances of the proposed CNN architecture of SR for supervised synthesis of subject-specific scans. The baseline methods for comparison are random forest regression for synthesis [Jog et al., 2017] (REPLICA). The metrics PSNR and SSIM with respect to normalized results between 0 and 1 are used to evaluate the methods. The quantitative results are shown in Table 4.1. Our method has a gain of about 2.3dB (synthesizing T1w images from T2w images) and 2.7dB (synthesizing T2w images from T1w images) with respect to PSNR compared to REPLICA. Although, our CNN-based approach has a greater SSIM when synthesizing T1w images from T2w images but lower SSIM when synthesizing T2w images from T1w images than those of the random forest-based method.

The qualitative results are shown in Figures 4.6 and 4.7. Visually, our proposed method reconstructs better contours and sharpener edges than the compared method. The cortex (e.g. gray matter) of the result of 20L-SRReCNN shown in 4.6 is more visible and more curvature than those of REPLICA. However, the white matter regions of this CNN-based method is too smooth. In the case of synthesized T2w images of Figure 4.7, the result of our CNN-based technique has salt-and-pepper noise, leading to lower SSIM than REPLICA. The problem of noisy synthesis T2w images comes from the property that we use the residual for training our networks and T1w images are not pre-denoised. The networks attempt to find the mapping for the voxel-wise differences between paired training images without considering their structure as in Equation (4.12).

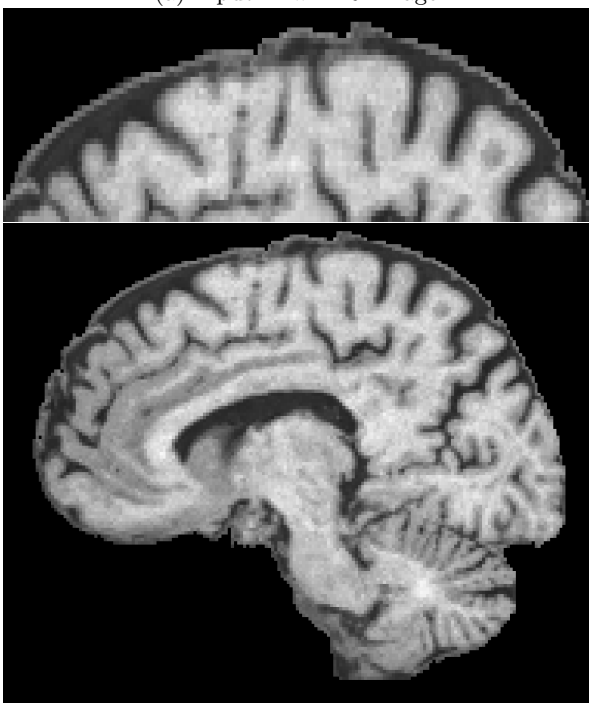
4.2.4 Discussion

We have introduced an approach to synthesize one image of a target modality from an observed image of a source modality. Although the results are sensible to the noise of data, our proposed method could also generate synthesized images which are comparable to the baseline method. Our approach shows the potential of CNN-based technique for cross-modal medical image synthesis problem.

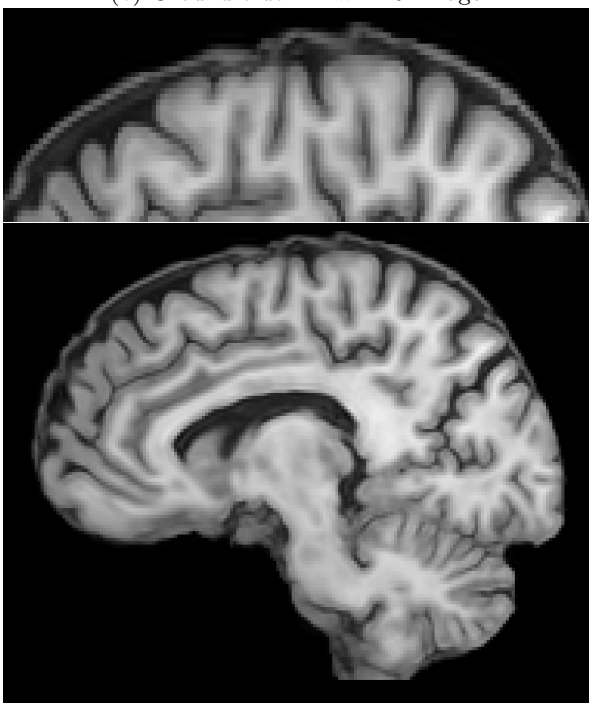


(a) Input T2w MRI image

(b) Ground truth T1w MRI image

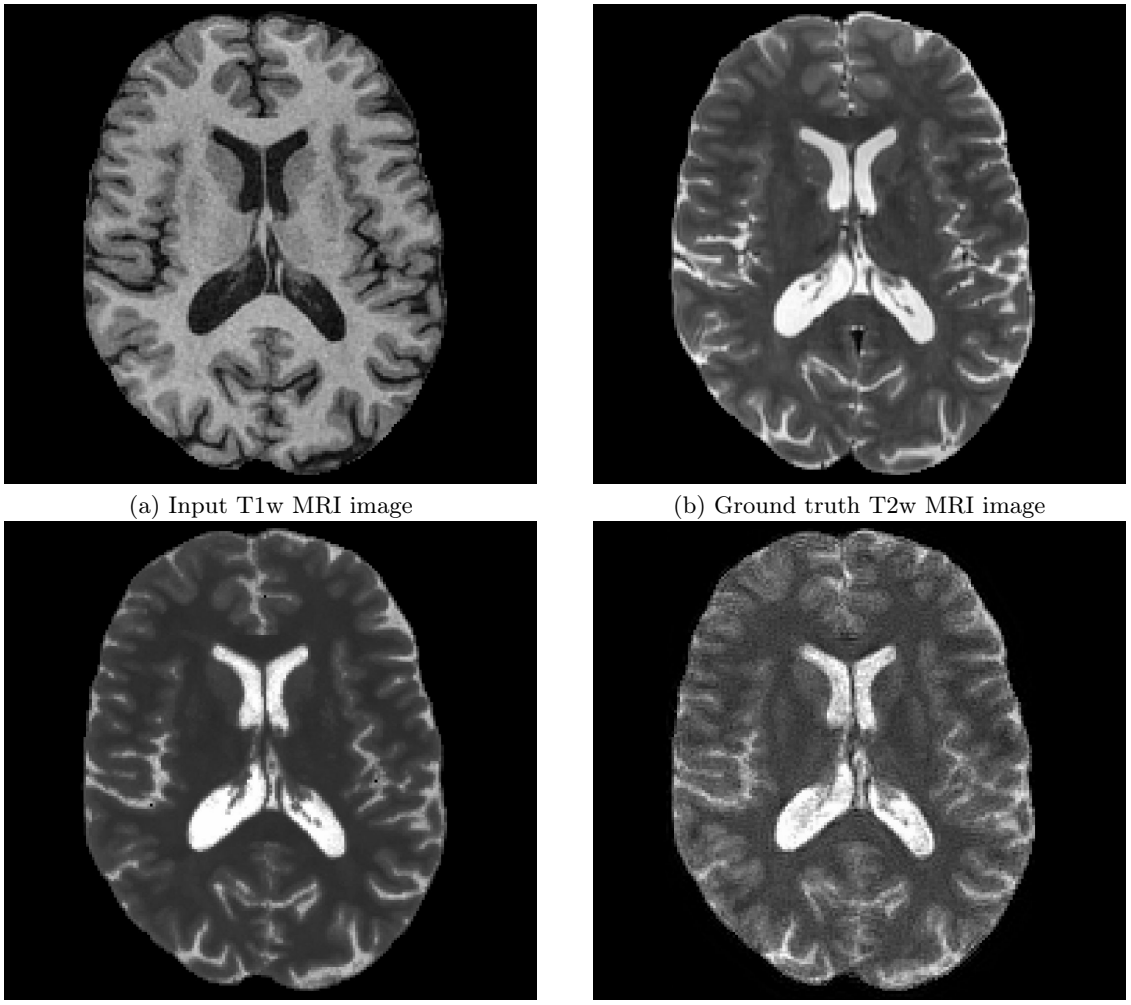


(c) REPLICA [Jog et al., 2017]



(d) 20L-SRReCNN

FIGURE 4.6: The examples (i.e. the coronal slices of a brain MRI) of cross-modal synthesis methods. The input T2w MRI image (a) is synthesized by the method REPLICA [Jog et al., 2017] and our proposed 20L-SRReCNN. The zoom versions are at the upper corners.



(a) Input T1w MRI image

(b) Ground truth T2w MRI image

(c) REPLICA [Jog et al., 2017]

(d) 20L-SRReCNN

FIGURE 4.7: The examples (i.e. the sagittal slices of a brain MRI) of cross-modal synthesis methods. The input T2w MRI image (a) is synthesized by the method REPLICA [Jog et al., 2017] and our proposed 20L-SRReCNN.

Evaluation of cross-modal techniques is done on the original images. However, the results show blurry and noisy. The pre-processing steps such as pre-denoising and bias correction should be applied before training networks. In addition, further works is required to investigate the principal elements of networks with respect to the performance of networks such as: depth of networks, non-residual-learning, the size of filters. Thus, a general pipeline will be drawn based on these works.

Future research direction for supervised learning CNN-based cross-modal techniques could focus on the other networks such as generative adversarial network [Isola et al., 2017, Nie et al., 2018, Yang et al., 2018] or on other elements of the network architecture as the embedded networks [Xiang et al., 2018]. Moreover, the perceptual losses of SR problem as in [Johnson et al., 2016, Ledig et al., 2017] can be applied to reduce the noisy data and generate the naturally synthesized images.

4.3 Unpaired synthesis with generative adversarial networks

Recently, image-to-image translation using GANs have been receiving significant attention from research community [Isola et al., 2017, Kim et al., 2017, Yi et al., 2017, Zhu et al., 2017]. Recent work learns this task in an unpaired learning manner [Kim et al., 2017, Yi et al., 2017, Zhu et al., 2017]. For instance, an architecture with two-block GANs and a connection based ℓ_1-norm cycle-consistent loss has been investigated for translating unpaired images [Zhu et al., 2017]. Another work similar to [Zhu et al., 2017] but with ℓ_2-norm cycle-consistent loss has also proposed in [Kim et al., 2017]. A concurrent work [Yi et al., 2017] with the same approach as cycleGAN has improved the stability of GANs but using Wasserstein GAN [Arjovsky et al., 2017] instead of sigmoid cross-entropy loss used in the original GANs. However, all these frameworks do not take advantage of the shared features between modalities. In this section, we propose an approach to use GANs for unpaired medical image synthesis.

4.3.1 Mathematical formulation

In this work, we propose an unsupervised learning technique for cross-modal synthesis of brain MRI scans using generative adversarial networks. Our GAN consists of one single discriminator and one single generator. Given the training dataset of unpaired images, we assume the "$class$" of each tissue contrast corresponding to a non-negative integer number (e.g. T1w images are the class "1" and T2w images are the class "2"). Our networks use the embedding techniques described in [Gal and Ghahramani, 2016] to turn integer indexes into dense vectors of the fixed size of class numbers : $class \rightarrow c$. The embedding classes make networks easier to train [Gal and Ghahramani, 2016]. In addition, they can separate different classes but learn joint mappings into the same network.

4.3.1.1 Adversarial Loss

A generator G is trained to learn functions between multiple domains. The objective of the generator is to fool the discriminator D by generating the images y_{tar}, which are indistinguishable from real images of the target source x_{tar}. The generator learns a mapping from input images of a source domain x_{src} with a target label $c_{G,tar}$ (embedded by G) to the generated images as $G : \{x_{src}, c_{G,tar}\} \rightarrow y_{tar}$. We use $c_{G,tar}$ to control the desired mapping between different domains. For example, $G(x_{T1w}, c_{G,T2w})$ estimates the synthesized T2w image y_{T2w} of the T1w image x_{T1w} and $G(x_{T1w}, c_{G,Flair})$ estimates the synthesized Flair image y_{Flair} of this T1w input.

Meanwhile the discriminator D is trained to distinguish the generated images and the real image x_{tar} of the target domain. In order to support the discriminator to discriminate different target domains, the discriminator is conditioned by the label $c_{D,tar}$, which is embedded by D, as $D : \{y_{tar}, c_{D,tar}\} \rightarrow D(y_{tar}, c_{D,tar})$ and $D : \{x_{tar}, c_{D,tar}\} \rightarrow D(x_{tar}, c_{D,tar})$. The

training step of the discriminator is independent to those of the generator. Thus, we need one embedding layer for each network in order to it can minimize its weights independently. We express the objective of this adversarial learning as:

$$\begin{aligned}\mathcal{L}_{adv} &= \mathbb{E}_{x_{tar},c_{D,tar}}[log(1 - D(x_{tar}, c_{D,tar}))] + \mathbb{E}_{x_{src},c_{D,tar},c_{G,tar}}[log(D(y_{tar}, c_{D,tar}))] \\ &= \mathbb{E}_{x_{tar},c_{D,tar}}[log(1 - D(x_{tar}, c_{D,tar}))] + \mathbb{E}_{x_{src},c_{D,tar},c_{G,tar}}[log(D(G(x_{src}, c_{G,tar}), c_{D,tar}))]\end{aligned} \tag{4.14}$$

Here, the generator G tries to minimize this cost function while the discriminator D tries to maximize it. When the discriminator gets to local optimal, the use of logarithm term in the adversarial loss is equivalent to minimizing the Jenson-Shannon divergence [Arjovsky et al., 2017] between the real images and the synthetic images. If the real images and the synthetic images share no support, Jenson-Shannon divergence implied by Equation (4.14) becomes saturated (i.e. a constant). Optimizing Equation 4.14 suffers from due to the gradient-vanishing effect. And sometimes, the model tends to the collapsing mode [Arjovsky et al., 2017, Goodfellow et al., 2014, Salimans et al., 2016]. In order to effectively avoiding the mode collapsing problem, [Arjovsky et al., 2017] propose to replace the logarithm cross-entropy loss by Wasserstein distance (WGAN) as:

$$\begin{aligned}\mathcal{L}_{adv} &= -\mathbb{E}_{x_{tar},c_{D,tar}}[D(x_{tar}, c_{D,tar})] + \mathbb{E}_{x_{src},c_{D,tar},c_{G,tar}}[D(G(x_{src}, c_{G,tar}), c_{D,tar}))] \\ &s.t. \parallel D \parallel_L \leq 1\end{aligned} \tag{4.15}$$

where $\parallel D \parallel_L \leq 1$ denotes 1-Lipschitz constraint. Here, the discriminator D becomes a "critic" because it does not distinguish the true and the generated but attempts to minimize the differences between them. In order to implement this constraint, [Arjovsky et al., 2017] use the weight clipping method for the discriminator. However, the value of the clipping threshold, which affects the interactions between the weight constraint and the cost function, is hard to choose. An inappropriate value can induce either vanishing or exploding gradients. An improved version of WGAN proposes to use gradient penalty [Gulrajani et al., 2017] for Equation 4.14 as:

$$\begin{aligned}\mathcal{L}_{adv} &= -\mathbb{E}_{x_{tar},c_{D,tar}}[D(x_{tar}, c_{D,tar})] + \mathbb{E}_{x_{src},c_{D,tar},c_{G,tar}}[D(G(x_{src}, c_{G,tar}), c_{D,tar}))] \\ &+ \lambda_{gp}\mathbb{E}_{\hat{x}}[\parallel (\nabla_{\hat{x}} D(\hat{x}, c_{D,tar}) \parallel_2 -1)^2]\end{aligned} \tag{4.16}$$

where $\hat{x} = \epsilon x_{src} + (1-\epsilon)G(x_{src}, c_{tar})$ denotes the interpolation between the real image and the generated image with a random number $\epsilon \sim U[0, 1]$, and λ_{gp} controls the importance between the objectives. [Gulrajani et al., 2017] investigate that the optimal critic has unit gradient norm almost everywhere under the data distribution of the true images and the generated images. Intuitively, the true image and the optimally generated image should share the same gradient. Thus, this procedure makes the networks easily to train.

4.3.1.2 Cycle Consistency Loss

The adversarial loss does not guarantee that learned mappings can induce a generated image that matches exactly the target image because the networks map the same set of input images to any random images in the target domain [Yi et al., 2017, Zhu et al., 2017]. Equation (4.16) only optimizes the mapping between the domain part between domains. Thus, a cycle consistence loss [Kim et al., 2017, Yi et al., 2017, Zhu et al., 2017] is applied to the generator to preserve the identical content of the input images and the translated one, described as:

$$\mathcal{L}_{cyc} = \mathbb{E}_{x_{src}, c_{G,tar}, c_{G,src}}[\| \, x_{G,src} - G(G(x_{src}, c_{G,tar}), c_{G,src}) \, \|_l] \tag{4.17}$$

where $c_{G,src}$ is the embedded source domain label and l is a norm. The generator translates the input images into the output images and then reconstructs from translated ones to the original images. [Isola et al., 2017] investigate that the use of ℓ_1-norm is better than ℓ_2-norm for training GANs because ℓ_1-norm encourages less blurring. Thus, [Choi et al., 2018, Yi et al., 2017, Zhu et al., 2017] propose the cycle consistence loss with ℓ_1-norm so as to encourage low-frequency correctness. In this work, we propose to use a robust Charbonnier loss function (a differentiable variant of ℓ_1-norm) $\rho(x) = \sqrt{x^2 + \epsilon_\rho^2}$ (ϵ_ρ is set to $1e-3$), which achieves a better high-quality reconstruction than ℓ_2-norm such as in the SR problem [Lai et al., 2017]. The cycle loss can be now expressed as:

$$\mathcal{L}_{cyc} = \mathbb{E}_{x_{src}, c_{G,tar}, c_{G,src}}[\rho \left(x_{src} - G(G(x_{src}, c_{G,tar}), c_{G,src}) \right)] \tag{4.18}$$

4.3.1.3 Total Variation Loss

The output image y_{tar} may be generated with high-frequency artefacts, which are remarked in several GAN methods [Choi et al., 2018, Yi et al., 2017, Zhu et al., 2017]. The TV regularizer has been investigated in the neural style transfer domain [Gatys et al., 2016, Johnson et al., 2016], super-resolution [Gatys et al., 2016, Johnson et al., 2016] and feature inversion [Mahendran and Vedaldi, 2015]. Thus, we apply total variation (TV) regularization to encourage spatial smoothness of the output. The TV loss for a 3D output is described as:

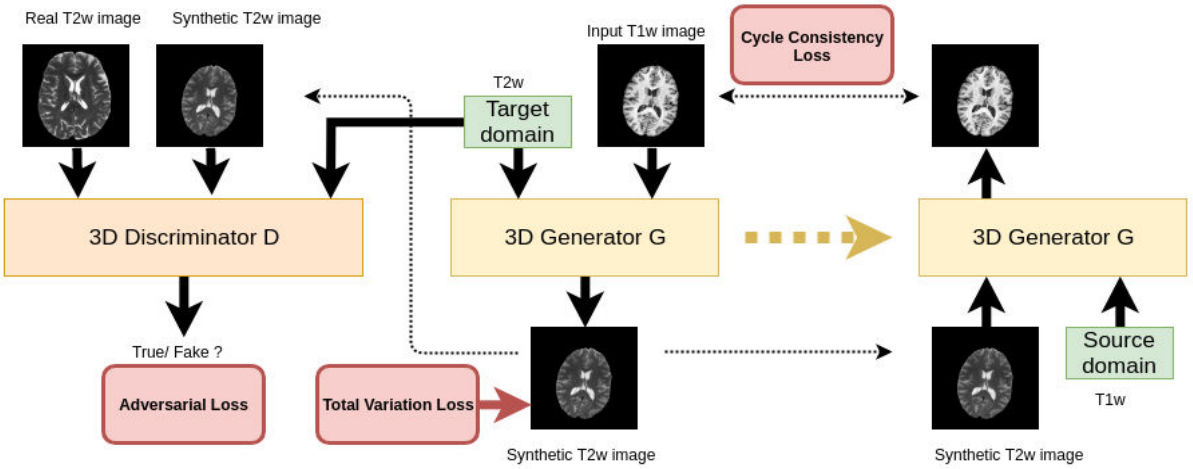
$$\mathcal{L}_{TV}(y) = \sum_{i,j,k} \left((y_{i,j,k+1} - y_{i,j,k})^2 + (y_{i,j+1,k} - y_{i,j,k})^2 + (y_{i+1,j,k} - y_{i,j,k})^2 \right)^{\frac{\beta}{2}} \tag{4.19}$$

4.3.1.4 Full Objective

Our full objective is described as:

$$\begin{aligned}\mathcal{L}_{total} &= \mathcal{L}_{adv} + \lambda_{cycle}\mathcal{L}_{cycle} + \lambda_{TV}\mathcal{L}_{TV} \\ &= -\mathbb{E}_{x_{tar},c_{D,tar}}[D(x_{tar}, c_{D,tar})] + \mathbb{E}_{x_{src},c_{D,tar},c_{G,tar}}[D(G(x_{src}, c_{G,tar}), c_{D,tar}))] \\ &+ \lambda_{gp}\mathbb{E}_{\hat{x}}[\| \ (\nabla_{\hat{x}} D(\hat{x}, c_{D,tar}) \ \|_2 - 1)^2] \\ &+ \lambda_{cycle}\mathbb{E}_{x_{src},c_{G,tar},c_{G,src}}[\rho \left(x_{src} - G(G(x_{src}, c_{G,tar}), c_{G,src})\right)] \\ &+ \lambda_{TV}\mathcal{L}_{TV}(G(x_{src}, c_{G,tar}))\end{aligned} \tag{4.20}$$

where λ_{cycle} and λ_{TV} is parameters which denote the significance of the reconstruction process and the TV regularization. In summary, the generator G is aimed at generating the synthetic images $G(x_{src}, c_{G,tar})$, which are as similar as possible to the images of a target domain x_{tar}, from the images of a source domain x_{src}. The images x_{tar} and x_{src} are unpaired. The discriminator D distinguishes the true image x_{tar} and the generated image $G(x_{src}, c_{G,tar})$, which are conditioned by the embedded target domain $c_{D,tar}$. Meanwhile, the generator G attempts to fool the discriminator by setting the generated image as the true images of the adversarial loss. In addition, the synthetic images are mapped backward to the source images as $G(G(x_{src}, c_{G,tar}), c_{G,src})$. Intuitively, the generated images of the synthetic images back to the source domain must be the source images x_{src}. The cycle consistency loss guarantees this property of images. Besides, the TV regularizer is applied to the synthetic images $G(x_{src}, c_{G,tar})$ so as to ensure their smoothness. Figure 4.8 illustrates our proposed GAN-based method that generates synthetic T2w images from real T1w images of a specific subject using other the T2w images of other subjects.



(a) 3D generator

FIGURE 4.8: Illustration of our proposed 3D GANs for unpaired cross-modal synthesis so as to generate synthetic T2w images from T1w images

4.3.2 Network architectures and training

The works in [Choi et al., 2018, Yi et al., 2017, Zhu et al., 2017] use instance normalization (InsNorm) layers [Ulyanov et al., 2017] for their networks as:

$$InsNorm(x_i) = \gamma \frac{x_i - \mu_B}{\sqrt{\sigma_B^2 + \epsilon}} + \beta \tag{4.21}$$

where x_i denotes values of i^{th} batch of input x over the mini-batch m $(i = 1, ..., m)$, μ_B and σ_B are respectively the average and variance of i^{th} batch and ϵ is a constant. β and γ are respectively learned scale and shift parameters of the layer. However, the use of one β and one γ for all domain-to-domain mappings limits the representation of a rich visual vocabulary for the construction. Instead, we use conditional instance normalization (CondInsNorm) layers [Dumoulin et al., 2017] as:

$$\begin{aligned} CondInsNorm(x_i) &= \gamma_{c_{tar}} \frac{x_i - \mu_B}{\sqrt{\sigma_B^2 + \epsilon}} + \beta_{c_{tar}} \\ &= (\gamma \times c_{tar}) \frac{x_i - \mu_B}{\sqrt{\sigma_B^2 + \epsilon}} + (\beta \times c_{tar}) \end{aligned} \tag{4.22}$$

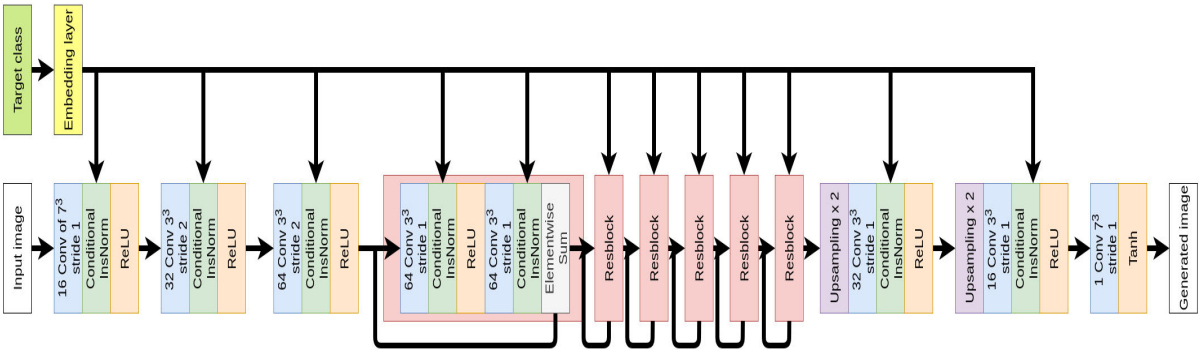
where $\beta_{c_{tar}}$ and $\gamma_{c_{tar}}$ are respectively learned scale and shift parameters of the layer for the mapping from the input image x_{src} to the target domain c_{tar} and $\times$ denotes multiplication. The embedding layer in [Gal and Ghahramani, 2016] is used to create the embedded domains. All our networks are based on 3D layers with training patch size of $128 \times 128 \times 128$.

4.3.2.1 Generator architectures

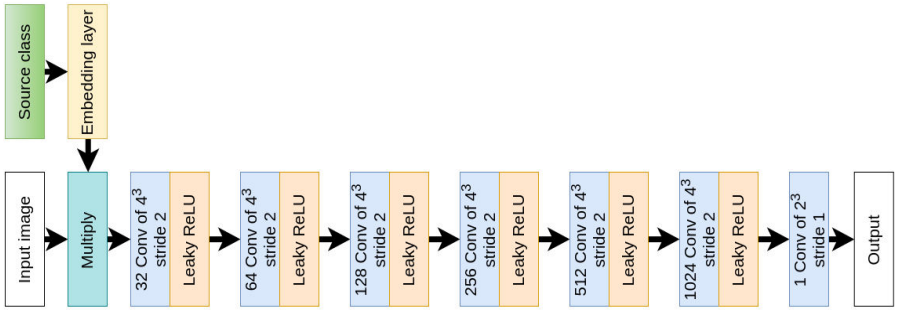
We denote *c7s1-k* as $7 \times 7 \times 7$ Convolution-CondInsNorm-ReLU layer and *c7-k* as $7 \times 7 \times 7$ Convolution layer with k filters and stride 1. Let *c3s2-k* denotes a $3 \times 3 \times 3$ Convolution-CondInsNorm-ReLU layer with k filters, and stride 2. *R-k* denotes a residual block that contains $3 \times 3 \times 3$ Convolution-CondInsNorm-ReLU-Convolution-CondInsNorm layer with the same number of filters of k on both convolution layers. *u-k* denotes $3 \times 3 \times 3$ Upsampling-Convolutional-InstanceNorm-ReLU layer with k filters and the scaling factor of $\times 2$. After the last layer, we apply a tanh activation. The generator architecture is: *c7s1-16*, *c3s2-32*, *c3s2-64*, *R-64*, *R-64*, *R-64*, *R-64*, *R-64*, *R-64*, *u-32*, *u-16*, *c7-1*. The reflecting padding is used to decrease the artefacts of the output. The illustration of the generator is shown in Figure 4.9 (a).

4.3.2.2 Discriminator architectures

The input of our discriminator is a Hadamard product of the input image and a bedded domain/target class label. Wasserstein GAN [Gulrajani et al., 2017] suggests that normalization layer should not be used in the discriminator. We use Leaky ReLU with a negative



(a) 3D generator



(b) 3D discriminator

FIGURE 4.9: The architecture of our proposed 3D GANs for unpaired cross-modal synthesis

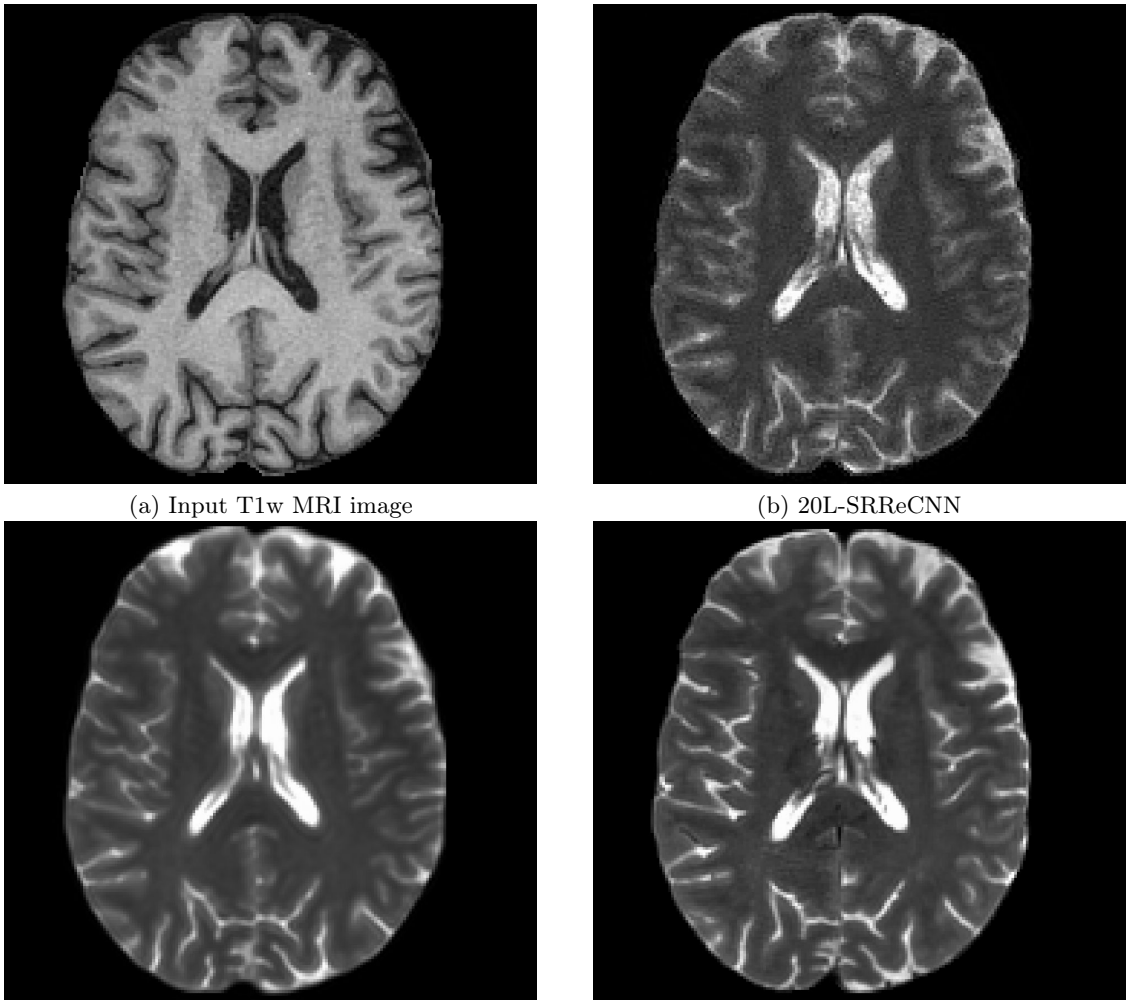
slope of 0.01. Let *c-k* denote a $4 \times 4 \times 4$ Convolution-LeakyReLU layer with k filters, the stride of 2 and same padding. The last layer is a $2 \times 2 \times 2$ convolution layer with stride of 1. No activation layer is used after the last layer. The discriminator consists of: *c-32*, *c-64*, *c-128*, *c-256*, *c-512*, *c-1024*. The illustration of the discriminator is shown in Figure 4.9 (b).

4.3.2.3 Network training

The 3D network is trained over 20 epochs on GPU Titan X using Keras with batch size of 1. Training uses Adam method [Kingma and Ba, 2015] with learning rate of 0.0001 and updates the discriminator 5 times before training the generator as in [Gulrajani et al., 2017]. When updating the generator, we freeze the weights of the discriminator. In our experiments, we set $\lambda_{gp} = 10$, $\lambda_{cyc} = 5000$. At test time, the whole synthesized image is reconstructed by the weighted predictions of patches of the testing image.

4.3.3 Experimental results

We show our qualitative results in Figures 4.10 and 4.11. Although, our approach is an unsupervised learning method, it can generate the synthetic images which have the same contrast to the ground truth images. The CSF, white matter and gray matter regions are reconstructed as close as the ground truth. Synthesized images using our proposed unsupervised approach capture most of the structural information.



(a) Input T1w MRI image (b) 20L-SRReCNN

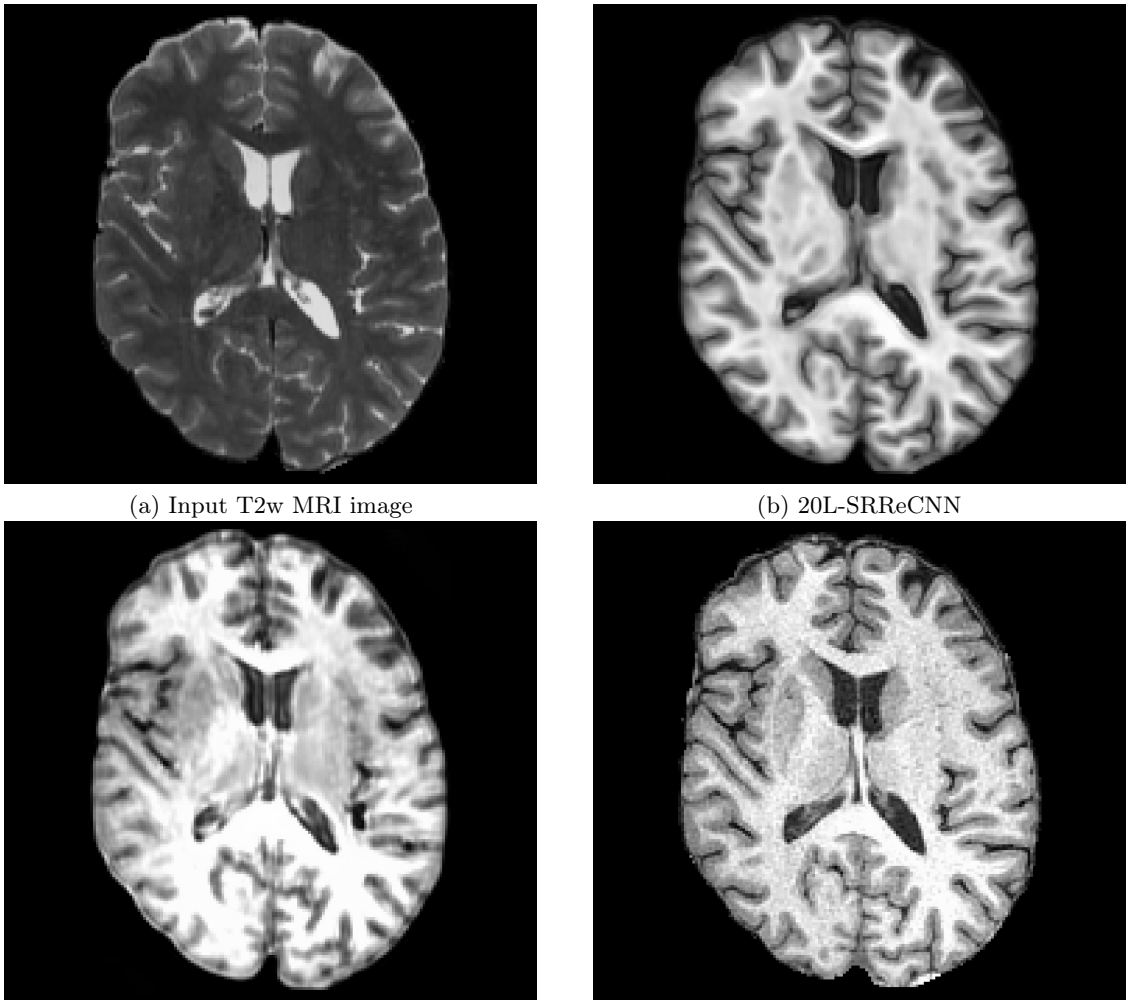
(c) Our proposed GANs with $\lambda_{TV} = 0.001$ (d) Ground truth T2w MRI image

FIGURE 4.10: The examples (i.e. the axial slices of a brain MRI) of cross-modal synthesis methods. The input T2w MRI image (a) is synthesized by the supervised method 20L-SRReCNN and our unsupervised method GAN.

	20L-SRReCNN		our GAN					
			$\lambda_{TV} = 0.005$		$\lambda_{TV} = 0.001$		$\lambda_{TV} = 0$	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Synthesized T1w	15.6848	0.9584	15.8333	0.9562	16.9957	0.9528	14.0984	0.8971
Synthesized T2w	20.5420	0.9528	15.3921	0.9419	16.3176	0.9401	14.6016	0.8925

TABLE 4.2: The results of PSNR/SSIM for our GAN-based cross-modal synthesis methods with respect to the parameter λ_{TV}. All methods using the training and testing images of NAMIC with $\lambda_{gp} = 10$, $\lambda_{cyc} = 5000$.

The quantitative results are illustrated in Table 4.2. We use PSNR/SSIM metrics to evaluate our method with respect to TV regularization ($\lambda_{TV} \in \{0, 0.001, 0.005\}$) and compared to the supervised learning method 20L-SRReCNN (described in Section 4.2). When synthesising T1w images from T2w images, the proposed GAN-based methods have comparable results as 20L-SRReCNN. However, our unsupervised method shows worse PSNR/SSIM than the supervised method. The reason relies on the fact that the GAN-based method attempts to optimize two mappings inside one single network at the same time while two networks



(a) Input T2w MRI image (b) 20L-SRReCNN

(c) Our proposed GANs with $\lambda_{TV} = 0.001$ (d) Ground truth T1w MRI image

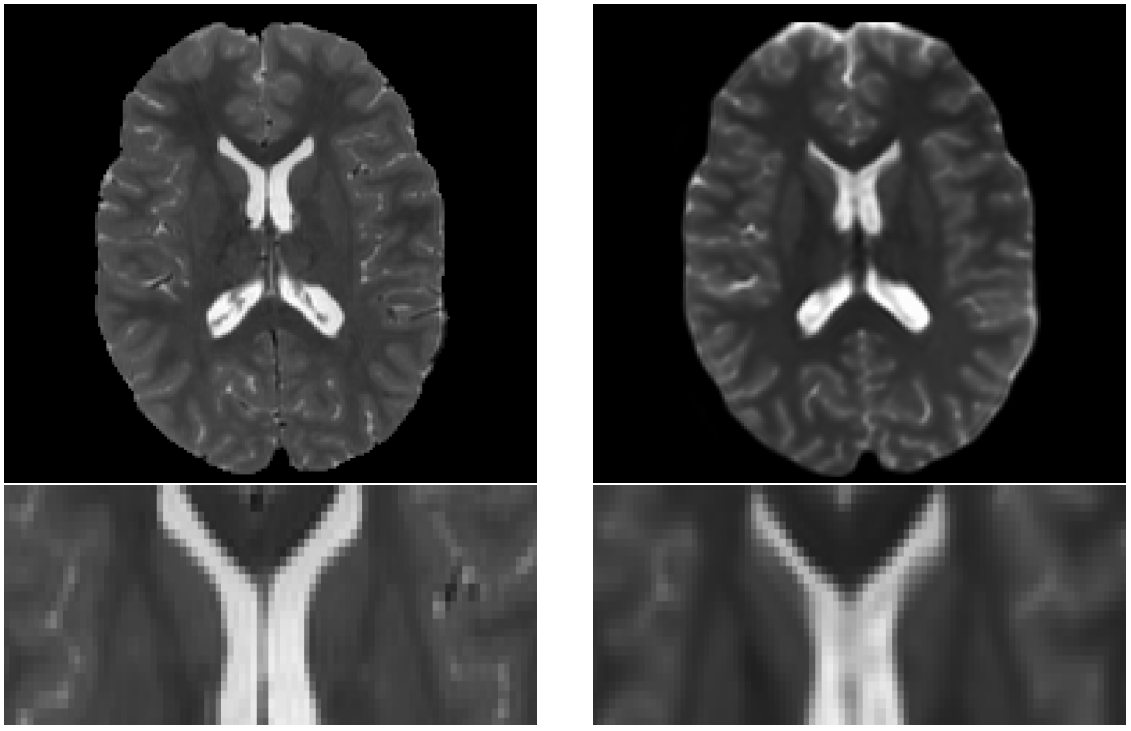
FIGURE 4.11: The examples (i.e. the axial slices of a brain MRI) of cross-modal synthesis methods. The input T1w MRI image (a) is synthesized by the supervised method 20L-SRReCNN and our unsupervised method GAN.

20L-SRReCNN are trained for one mapping.

The role of TV regularization is crucial to our method. TV regularization induces better results than whose of no TV parameter. The lower $\lambda_{TV} = 0.001$ leads to higher PSNR but lower SSIM than $\lambda_{TV} = 0.005$. The illustration of the role of TV regularization is shown in Figure 4.12. No TV regularizer induces high-frequency artefacts. The reconstruction with $\lambda_{TV} = 0.005$ leads to more smoothed results.

4.3.4 Discussion

In this section, we proposed a general unsupervised GAN-based method for cross-modal synthesis of subject-specific scans. Our method works without paired training data from source and target domains. Although the results are a little blurry, the technique shows an approach to solve our cross-modal synthesis problem without the paired dataset. Our method is effective for joint training between different domains thanks to the embedded



(a) Ground truth T2w MRI image

(b) Our proposed GANs with $\lambda_{TV} = 0.005$

(c) Our proposed GANs with $\lambda_{TV} = 0.001$

(d) Our proposed GANs with $\lambda_{TV} = 0$

FIGURE 4.12: The sensibility of TV regularization within our GAN-based method. The zoom versions of ventricle regions are at the lower corners.

labels of the target/source domains. We demonstrate that our conditional GAN can learn mappings between multiple MRI tissue contrasts.

We investigate that better performance can be achieved by adding total variation regularization. Our study shows that a small trade-off parameter for this regularizer is enough to generate better results. A higher parameter can induce too smooth reconstructions. The investigation of other regularizations will be considered in the future works such as ℓ_2-norm or Gaussian kernel.

The future research directions for GAN-based unpaired cross-modal synthesis could focus on applying the synthetic images for CNN-based SR techniques. As shown in [Pham et al., 2017b] that the multimodal method for brain MRI where a HR reference image can leverage the SR results. We believe that the synthetic HR images instead of using the original HR contrast can lead to better results.

Another application of cross-modal synthesis is to support segmentation methods. [Leroy et al., 2011] shows that the cortex segmentation of neonatal brain can be drawn from T2w images. However, the resolution of T2w images is usually lower than the corresponding T1w images of the same subject. Unpaired cross-modal synthesis can generate the T2w synthetic images from the HR T1 images. Then, segmentation methods may apply to these HR synthetic T2w images for cortex segmentation.

"La prochaine révolution de l'IA sera non-supervisée"-Yann LeCun ([2]). More improved GANs technique are growing up. In the future work, we will study more GAN-based techniques (e.g. Fisher GAN [Mroueh and Sercu, 2017]) or other new layers of networks for better synthesis reconstruction. Kernel methods for GAN in [Zhang et al., 2017] would allow to improve the performance of adversarial networks.

[2]RFIAP2018 : https://rfiap2018.ign.fr/

Chapter 5

Conclusions and Perspectives

Contents

Conclusions

In this work, two intended applications of medical image representations have been presented: single image super-resolution (SR) and cross-modal synthesis. SR and cross-modal synthesis have been receiving attention from the research community for recent years. The desire for SR in medical imaging stems from many applications: understanding of the anatomy, helping accurate segmentation and registration, and overcoming the hardware limitations of medical imaging devices. The motivation for cross-modal synthesis raises from many aspects: the mutual support between multi-modality medical imaging, helping accurate segmentation and super-resolution. Several methods for these problems have been introduced: patch-based method, sparse coding, random forest and neural networks.

The first contribution presented, relies on the investigated of 3D convolutional neural networks for brain MRI super-resolution instead of classic 2D networks. Then, several principal elements of networks are analysed to improve the performance such as the optimization methods, the depth of networks, weight initialization schemes, residual learning, multiscale learning. Next, an approach to take advantage of another HR reference image for improve the MRI super-resolution process is proposed. Finally, the application of super-resolution for enhancing the real clinical neonatal brain MRI and supporting segmentation methods is investigated, which demonstrates our proposed networks with respect to practical medical imaging applications.

The second contribution described an approach for joint mappings of high-resolution reconstruction and segmentation using 3D generative adversarial networks. This method is not

only assessed on the simulated low-resolution images of the high-resolution neonatal dataset, but also used to enhance and segment real clinical anisotropic low-resolution images. Our results demonstrate the potential of our GAN-based method with respect to practical medical applications.

The third contribution proposes 3D convolutional neural networks for paired cross-modal synthesis and 3D generative adversarial networks for unpaired cross-modal synthesis. Our CNN-based network for SR applied directly to cross-modal synthesis shows comparable results to the start-of-the-art methods. Moreover, we propose an approach to exploit 3D generative adversarial networks for unpaired cross-modal synthesis. The results of our unsupervised method are encourage. Further improvements of generative adversarial networks are required to improve the performance.

Although, several researchers have proposed many methods to solve these two problems, many challenges still constraint these techniques from wide applications. Firstly, handling a huge number of training examples or complicated models can be induce computational cost. The methods such as CNNs depend on GPU for accelerating the intensive computation. Secondly, the observation model with a given point spread function can not be estimated perfectly, leading the sensitivity of techniques with outliers or the dissatisfaction of ideal conditions. Finally, the metrics such as PNSR or SSIM may not induce more appealing results. Better measures or qualitative results are still needed for performance evaluation. Thus, future researches continue to investigate better methods and more performance evaluation for the developments of these two applications.

Perspectives

In this thesis, we considered MRI contrasts (T1w and T2w). The addition of other contrasts such as FLAIR or dMRI or other modalities such as CT would be used to investigate the robustness of neural networks-based techniques. In [Alexander et al., 2017], image quality transfer (IQT) propagates information from rare or expensive high quality dMRI images to abundant or cheap low quality dMRI images by machine learning technique. The method raises the question of the potential of CNN in dMRI SR. We believe that our proposed cross-modal synthesis can be used to generate MRI brain scans from CT and vice verso, or diffusion-weight MRI or from low dose to high dose CT scan.

When applying CNN-based methods in a realistic setting, the choice of PSF is crucial. Thus, the second future direction would involve blind SR [Michaeli and Irani, 2013, Wang et al., 2005] instead of a simulated PSF so as to approximate better the PSF of observed LR images. On the other hand, the perceptual approaches can also be used to make the network independent from the PSF model.

The objective function of neural networks is based on the differences between pixel-wise or voxel-wise. Thus, this may lead to lack of texture information inside images. The future work

would combine neural networks with nonlocal and statistical priors [Fablet and Rousseau, 2016] to preserve the consistency of high-resolution textured patterns, which are missed in the observed low-resolution images.

Since the segmentation maps in Section 2.2.3 only focus on the cortex of brain, other regions of MRI images such as CSF, WM and GM can also be segmented using our proposed methods. In addition, other supervised learning segmentation algorithms such as atlas-based methods [Rousseau et al., 2010b] or CNN-based methods [Ronneberger et al., 2015] would allow to improve the accuracy of segmentation maps. 0ther organ imaging such as cardiac MRI or other types of medical imaging such as CT could exploit the proposed methods. Moreover, an end-to-tend network would be proposed for joint super-resolution and segmentation or even joint super-resolution, segmentation and synthesis.

Appendix A

Deep learning

Contents

A digital image contains a number of pixels which are arranged in a array. The discontinuities of the value range of pixels (a sudden increase in value) where these points are linked together as a structure termed edges. The closed edges compose of contours of images which may form certain shapes (e.g. a circle or a rectangular) or the pattern of a object. A set of some regular pattern defines an image texture. The content of an image is represented by patterns and textures. In image processing, a key role is to extract and detect these elements for image representation. Before deep learning, classic methods used to extract features of images or data which are the contours, edges or smoothness by a set of predefined filter such as Gaussian filters, Sobel filter. The feature extraction is a fundamental tool in image representation.

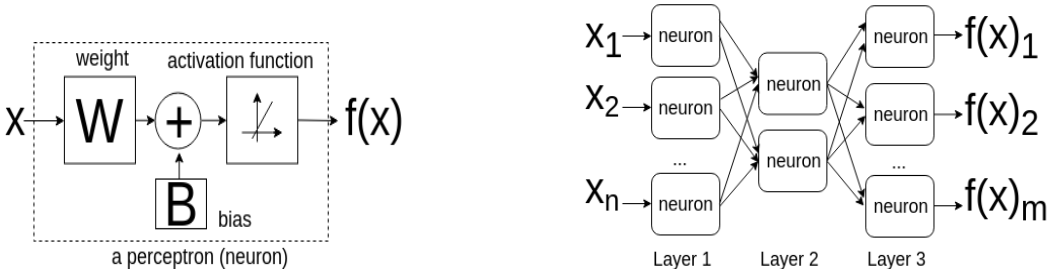
Deep learning, which is a class of machine learning, aims to learn implicitly features via a set of artificial neural networks. Artificial neural networks (ANNs) are systems for computing inspired biological neural networks of human brain. ANNs have many different architectures. One simple class of ANNs is a perceptron (a neuron), described as:

$$f_{neu}(x) = g(W \cdot x + B) \tag{A.1}$$

where $x, f_{neu}(x) \in \mathbb{R}$, g is an activation function, W is a weight and B is a bias. $W \cdot x$ denotes the dot product. For sake of clarity, we denote x and $f(x)$ as the argument (or input) and a model (or a function) respectively. The perceptron attempts to find the weight and bias, which approximate the relation of given input and correspond output. The space of solution of a perceptron is limited in the set of linear separability. In order to extend the solution space, a multilayer perceptron (MLP) connects several perceptrons for higher dimension of x and $f(x)$ with hidden layers. A MLP with one hidden layer can be denoted as:

$$f_{MLP}(x) = g_2(W_2 \cdot g_1(W_1 \cdot x + B_1) + B_2) \tag{A.2}$$

where now $x \in \mathbb{R}^n$ and $f_{MLP}(x) \in \mathbb{R}^m$ $(n, m \geq 1)$. g_i, W_i, B_i denote the activation function, weight and bias of the i^{th} layer. The units of MLPs are fully connected, each node in one layer connects every node in the following layer. A MLP consists of several fully-connected layers, activation layers and a cost function (so-called an objective function). However, MLPs are restricted for one-dimensional training set. In order to better represent higher dimensional patterns (e.g. edges, contours), we can supplement our neural networks with the convolution operation. In the next section, we will go in detail of convolutional neural networks, their characteristics and how to optimize the training of networks.



(a) a neuron

(b) a MLP with hidden layers

FIGURE A.1: The example of a computing neuron with input x and output $f(x)$ as Equation A.1 and a MLP with hidden layers

A.1 Convolutional neural networks

Convolutional neural networks (CNN), which are variants of MLPs, consist of several layers, especially convolutional layers for representing the features of input set. CNNs optimize the weights and biases of their weight layers through the networks in order to find the relation of given training dataset. A example of a CNN with two convolutional layers is decribed as:

$$f_{CNN}(x) = g_2(W_2 \star g_1(W_1 \star x + B_1) + B_2) \tag{A.3}$$

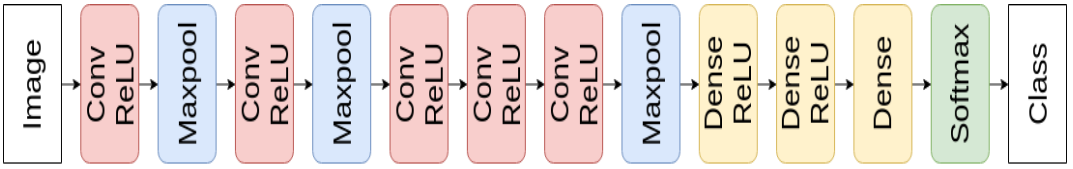
where $\star$ denote a convolution operation, W_i, B_i denote the weight and the bias of the i^{th} convolution layer. Mathematically, a convolution, which is a weighted sum of each element of the input to its local neighbors by filters, between an image I and a filter F can be written

as:

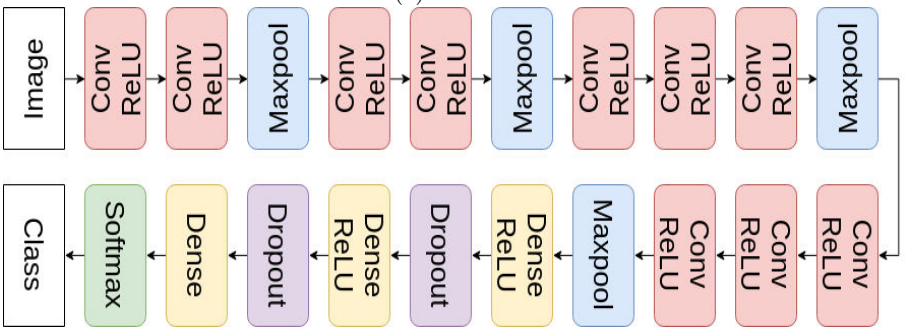
$$I_F(k,l) = \sum_{i=1}^{m}\sum_{j=1}^{n} I(i+k-1, j+l-1)F(i,j) \tag{A.4}$$

where $I \in \mathbb{R}^{M,N}$, F has a size of $m \times n$ and I_F denotes the convoluted image. Depending on our tasks, we could add more types of layers such as pooling layers, transposed convolution layers, embedding layers [Gal and Ghahramani, 2016], batch normalization layers [Ioffe and Szegedy, 2015] or sub-pixel layers [Shi et al., 2016], *etc.*

CNNs have been first studied in [Fukushima and Miyake, 1982, LeCun et al., 1998]. Nevertheless, these networks have been received the most attention from research community since 2014. A CNN called Alexnet has won a challenge of image classification [Krizhevsky et al., 2012]. This network consists of eight layers (convolution and fully-connected layers which need to be trained) and other in-place layers such as Maxpool (i.e. selecting the maximum value of a pooling window), activation functions (e.g. rectified linear unit (ReLU), Softmax). The architecture of AlexNet is drawn in Figure A.2 (a).



(a) Alexnet



(b) An architecture of VGG-nets

FIGURE A.2: The architecture of AlexNet [Krizhevsky et al., 2012] and VGG-net [Simonyan and Zisserman, 2014] for image classification (Recreating from [Krizhevsky et al., 2012, Simonyan and Zisserman, 2014]). Conv and Dense are convolution and fully-connected layers respectively. The block Conv/ReLU and Dense/ReLU denote respectivelt a convolution layer and a fully-connected layer before a ReLU layer.

Later, VGG-nets, very deep CNNs [Simonyan and Zisserman, 2014] drawn in Figure A.2 (b), has improved the performance of the predecessor. In parallel, InceptionNet, which consists of several blocks of the concatenation of filters and pool layers [Szegedy et al., 2015], has more performance than VGG-nets. Several architectures of CNNs have been proposed to many computer vision tasks. In order to generalize CNNs to object detection, the work in [Girshick et al., 2014, Ren et al., 2017] (R-CNN) aims to identify objects via a bounding box in the image.

A.2 Activation layers

The relationship between the input and the output of a problem may be nonlinear. So, they raise the need of the layers which can model the nonlinearity. In order to extend a network to represent nonlinear functions, we can apply nonlinear activation functions such as ReLU, softmax, hyperbolic tangent activation function, etc. The ReLU $f_{ReLU}(x)$ is an activation function which only keeps the positive part of its input x as:

$$f_{ReLU}(x) = x^+ = max(0, x) \tag{A.5}$$

The ReLU layer is close to linear so as the gradient-based methods can easily optimize [Goodfellow et al., 2016]. There are several versions of ReLU layers which extend the negative parts such as: Leaky ReLU [Maas et al., 2013](i.e. retaining the negative part by a small fixed scaling factor), Parametric ReLU [He et al., 2015] (i.e. the scaling factor for negative part is learned). Instead of just scaling the negative part, an exponential function can be apply to this part as ELU [Clevert et al., 2015]:

$$f_{ELU}(x) = max(0, x) + min(0, \alpha(e^x - 1)) \tag{A.6}$$

where α is a scale factor. The purpose of these activation functions is to preserve the properties of linear models for optimization but also model a nonlinear transformation. Another family of nonlinear layers can be used at the end of the networks for predicting a probability, in other word presenting a probability distribution such as sigmoid, softmax or hyperbolic tangent function. The logistic sigmoid function approaches to zero or one when its input is very negative or very positive, thus, it is commonly used for logistic regression as:

$$f_{sigmoid}(x) = \frac{1}{1 + exp(-x)} \tag{A.7}$$

In order to take advantage of sigmoid activation for multiple regression, we can use the softmax function which decomposes the arguments into K distinct linear functions as:

$$f_{softmax}(x)_i = \frac{exp(x_i)}{\sum_{j=1}^{K} exp(x_j)} \tag{A.8}$$

These above functions are often used for classification tasks. For a linear regression, the linear function is the simplest choice as:

$$f_{linear}(x) = x \tag{A.9}$$

However, data is sometimes needed to be normalized into the range of $[-1, 1]$ because of the compatibility of different dataset and the computational cost. The hyperbolic tangent function can be used for this purpose :

$$f_{tanh}(x) = tanh(x) \tag{A.10}$$

Intuitively, the hyperbolic tangent function maintains the linear relationship of the argument in the range of $[-1, 1]$ and it saturates if otherwise. In addition, the property of this function makes gradient-based optimization methods easily calculate the derivatives.

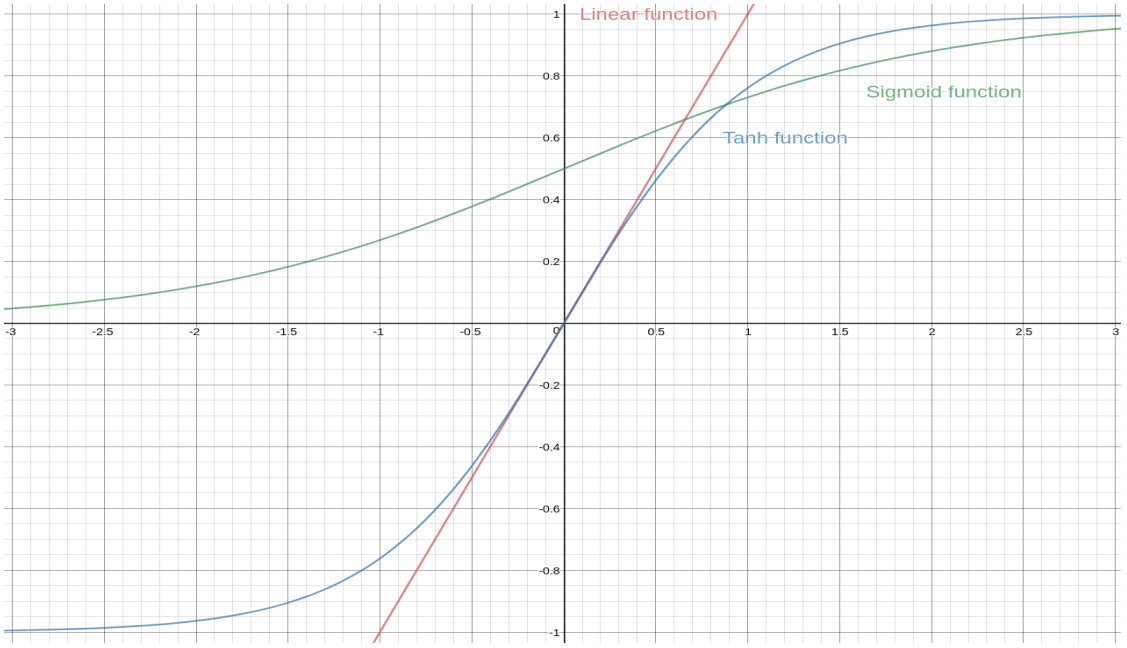


FIGURE A.3: Some activation functions

A.3 Some state-of-the-art CNN architectures

A.3.1 Residual networks

Afterward, an architecture CNN up to 100 layers has been proposed to use the residual blocs (Resnet) [He et al., 2016a]. A residual bloc [He et al., 2016b] draw in Figure A.4, of the i^{th} and $(i+1)^{th}$convolution layer can be :

$$f_{Res}(x) = g_{i+1}(BN_{i+1}\{W_{i+1} \star g_i[BN_i(W_i \star x)]\} + x) \tag{A.11}$$

where BN_i is the batch normalization (BN) function [Ioffe and Szegedy, 2015] of i^{th} layer. When the parameters of training networks are optimized, the distribution of activation functions is changed, leading to internal covariate shift. BN layers propose to normalize the response of convolution layers to produce activations with a stable distribution as:

$$BN(x_i) = \gamma \frac{x_i - \mu_B}{\sqrt{\sigma_B^2 + \epsilon}} + \beta \tag{A.12}$$

where x_i denotes values of i^{th} batch of input x over the mini-batch m $(i = 1, ..., m)$, μ_B and σ_B are respectively the average and variance of the mini-batch and ϵ is a constant. β and γ are respectively learned scale and shift parameters of the layer to ensure network to avoid

forward or backward signals vanish [He et al., 2016a]. Later, Mask R-CNN [He et al., 2017] supplements R-CNN the possibility of object segmentation with Resnet.

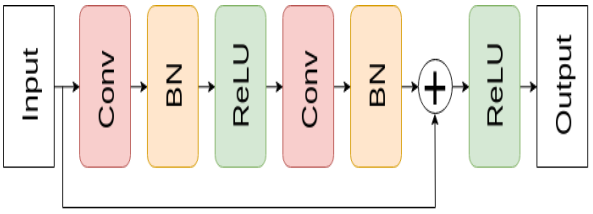


FIGURE A.4: The architecture of a residual block [He et al., 2016b]. BN denotes a batch normalization layer

A.3.2 Densely connected networks

Recently, [Huang et al., 2017a] have proposed to connect all layers in a block or in the networks called Densenet. These latest networks not only decrease number of parameters but also show good performance as the *"ultra deep"* Resnet. An example of densely connected block is illustrated in Figure A.5. Densenet proposes to concatenate all preceding layers $x_1, x_2, ..., x_i$ for the $(i+1)^{th}$ layer as:

$$f_{den,i+1}(x) = g_{i+1}(BN_{i+1}[W_{i+1} \star Concat(x_1, x_2, ..., x_i) + B_{i+1}]) \tag{A.13}$$

where *Concat* denotes the concatenation of feature-maps of all preceding layers. W_{i+1} and B_{i+1} are the parameters of the $(i+1)^{th}$ layer. The intuition of this approach is that each layers share all feature maps as "collective knowledge" [Huang et al., 2017a].

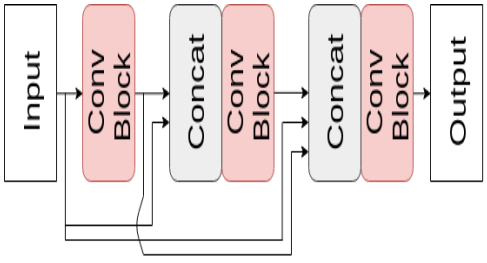


FIGURE A.5: The architecture of a densely connected block [Huang et al., 2017a]. A conv block may consist of convolutional layers, padding layers, BN and ReLU layers

The concatenation technique is also proposed by another famous network called U-Net in biomedical image processing [Ronneberger et al., 2015] (shown in Figure 4.3). However, U-net concatenates the symmetric layers instead of the whole preceding layers, resulting a U-form architecture.

A.4 Application of CNN for neural style transfer

The work in [Gatys et al., 2016] investigated an interesting application of CNNs based on [Simonyan and Zisserman, 2014] that is style transfer. Given a content image p, a style image s and a random image x, we would like to generate x to have the content of p with the style of s as the total loss as:

$$\mathcal{L}(x, p, s) = \alpha \mathcal{L}_{content}(x, p) + \beta \mathcal{L}_{style}(x, s) \tag{A.14}$$

where $\mathcal{L}_{content}(x, p)$ and $\mathcal{L}_{style}(x, s)$ denote the content reconstruction loss and the style loss respectively, and α and β are weights. The authors demonstrated visually that higher layers lost detailed pixel information and capture the high-level content of the image. In order to perverse the content for input image x, the feature reconstruction loss function is calculated by element-wise squared error:

$$\mathcal{L}_{content}(x, p) = \frac{1}{2} \sum_{i} (F_i^l(x) - F_i^l(p))^2 \tag{A.15}$$

where F^l is the feature maps of the l^{th} layer of a pre-trained network (e.g. VGG-net [Simonyan and Zisserman, 2014]). The second loss of Equation A.14 brings stylistic features to the input image, described as:

$$\mathcal{L}_{style}(x, s) = \sum_{l=1}^{L} w_l \sum_{i} (G(F_i^l(x)) - G(F_i^l(p)))^2 \tag{A.16}$$

where L is the number of chosen layers for style transfer, w_l denotes weighting factors and G corresponds the Gram matrix (i.e. inner products of the subsets). But this method slowly finds the solution because of inference processes. [Johnson et al., 2016] propose to add a independent transformation network Φ to transform an input image p to an generated image $\hat{s} = \Phi(p)$ which have the style of image s based on the perceptual loss:

$$\mathcal{L}(s, \hat{s}) = \alpha_\Phi \mathcal{L}_{content}(s, \hat{s}) + \beta_\Phi \mathcal{L}_{style}(s, \hat{s}) \tag{A.17}$$

where α_Φ and β_Φ are the trade-off coefficients. The first version of this method used batch normalization layers [Ioffe and Szegedy, 2015] to encode the mapping $\hat{s} = \Phi(p)$. However, this normalization applies to whole a batch of images leading to the slower optimization of networks. The work in [Ulyanov et al., 2017] proposes another type of normalization called instance normalization. The idea of this normalization layer is to calculate simply the mean and the standard deviation of the input on the sum of a single batch instead of a whole.

A.5 Generative adversarial networks

Since introduced in 2014, generative adversarial networks (GANs) [Goodfellow et al., 2014] have applied for many tasks such as 3D object generation [Wu et al., 2016], super-resolution [Ledig et al., 2017] and image translation [Isola et al., 2017], MRI one-domain synthesis [Bermudez et al., 2018]. GANs consist of two networks in which one network learns how to generate candidates mapped from a latent space while other discriminates them with instances from the true data distribution. The generative possibility and the stability of GANs can be improved by utilizing convolutional neural networks (DCGAN) [Radford et al., 2016], conditioning these two networks with class labels (cGANs) [Mirza and Osindero, 2014], adding auxiliary classifier (AC-GAN) [Odena et al., 2017].

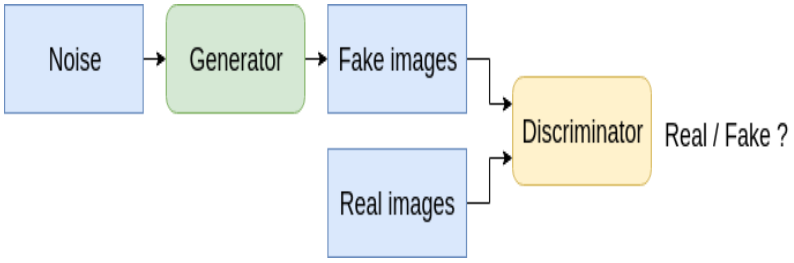


FIGURE A.6: The diagram of generative adversarial networks. Generator and Discriminator consist of convolutional neural networks

The generator learns a mapping from a noise z of the noise distribution $\mathbb{P}_z$ to a image x from the input distribution $\mathbb{P}_x$. Meanwhile the discriminator is trained to distinguish the generated image $G(z)$ and the real image. We express the objective of this adversarial learning [Goodfellow et al., 2014] as:

$$\min_G \max_D \mathcal{L}(D,G) = \mathbb{E}_{x\sim\mathbb{P}_x}[logD(x)] + \mathbb{E}_{z\sim\mathbb{P}_z}[log(1-D(G(z)))] \tag{A.18}$$

where the generator G tries to minimize this object while the discriminator D tries to maximize it. In order to improve the possibility of classification, conditioned GANs (cGAN) [Mirza and Osindero, 2014] attempt to embed the class of images into the generator and the discriminator as:

$$\min_{G,c} \max_{D,c} \mathcal{L}(D,G,c) = \mathbb{E}_{x\sim\mathbb{P}_x,c}[logD(x,c)] + \mathbb{E}_{z\sim\mathbb{P}_z,c}[log(1-D(G(z,c),c))] \tag{A.19}$$

where c is the embedded label of the real image x. Instead of feeding the discriminator with the label information, another approach is to task the discriminator predict the class of image (ACGAN). The former is now defined as :

$$\begin{aligned}\min_{G,c} \max_{D,c} \mathcal{L}(D,G,c) = \mathbb{E}_{x\sim\mathbb{P}_x}[logD(x)] + \mathbb{E}_{z\sim\mathbb{P}_z,c}[log(1-D(G(z,c)))] \\ +\mathbb{E}_{x\sim\mathbb{P}_x,c}[-logD_c(c \mid x)]\end{aligned} \tag{A.20}$$

where $D_c(c \mid x)$ is a probability distribution over labels computed by D. However, the use of logarithm term in the adversarial loss (known as Jensen-Shannon divergence) could

be saturated and in other cases, the model tends to the collapsing mode [Arjovsky et al., 2017, Goodfellow et al., 2014, Salimans et al., 2016]. In order to effectively avoid the mode collapsing problem, [Arjovsky et al., 2017] (WGAN) adopt Wasserstein distance to replace the logarithm cross-entropy loss as:

$$\min_G \max_D \mathcal{L}(D,G) = \mathbb{E}_{x\sim\mathbb{P}_x}[D(x)] - \mathbb{E}_{z\sim\mathbb{P}_z}[D(G(z))] \qquad s.t. \parallel D \parallel_L \leq 1 \tag{A.21}$$

where $\parallel D \parallel_L \leq 1$ denotes the 1-Lipschitz constraint. The authors propose to use the weight-clipping method to perform the constraint. Because of the difficulty of weight clipping on the network optimization, [Gulrajani et al., 2017] alternates this constraint by a gradient penalty. The objective function can be rewritten by an improved version of WGAN as:

$$\min_G \max_D \mathcal{L}(D,G) = \mathbb{E}_{x\sim\mathbb{P}_x}[D(x)] - \mathbb{E}_{z\sim\mathbb{P}_z}[D(G(z))] + \lambda_{gp}\mathbb{E}_{\hat{x}}[\parallel (\nabla_{\hat{x}} D(\hat{x}) \parallel_2 -1)^2] \tag{A.22}$$

where λ_{gp} is a trade-off and $\nabla_{\hat{x}}$ denotes the gradient of the interpolation $\hat{x}$ between the real input and the generated input as: $\epsilon x + (1-\epsilon)G(z)$. ϵ is a random number from a uniform distribution over an interval $[0, 1]$.

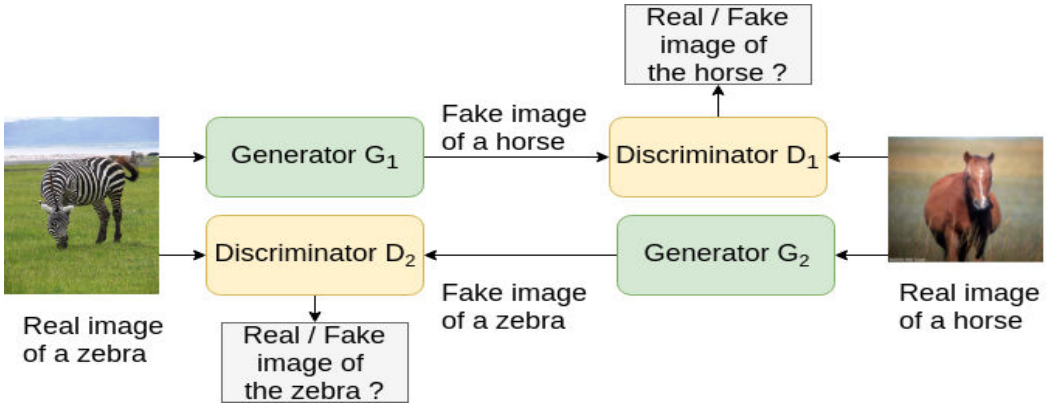


FIGURE A.7: The diagram of cycle-consistent adversarial networks (cycleGAN) [Zhu et al., 2017]. The method consists of two generators and two discriminators with a connection of cycle-consistent loss.

Recently, image-to-image translation using GANs has been receiving significant attention from research community [Isola et al., 2017, Kim et al., 2017, Yi et al., 2017, Zhu et al., 2017]. In the context of supervised learning, [Isola et al., 2017] investigated cGANs for paired image-to-image translation. Recent works learn this task in an unpaired learning manner [Kim et al., 2017, Yi et al., 2017, Zhu et al., 2017]. For instance, an architecture with two-block GANs and a connection based ℓ_1-norm cycle-consistent loss has been investigated for translating unpaired images [Zhu et al., 2017], as demonstrated as Figure A.7. Another work similar to [Zhu et al., 2017] but with ℓ_2-norm cycle-consistent loss has also proposed in [Kim et al., 2017]. A concurrent work [Yi et al., 2017] with the same approach as cycleGAN has improved the stability of GANs but using Wasserstein GAN [Arjovsky et al., 2017] instead of sigmoid cross-entropy loss used in the original GANs.

A.6 Optimization of neural networks

Given a training dataset which consists of N pairs of input x_i and corresponding output y_i: $\mathcal{D} = \{x_i, y_i \mid i = 1, 2, ...N\}$, network parameters are estimated by minimizing the objective function using optimization algorithms. The objective function of the MLP in Equation (A.2) with two hidden layers can be expressed as:

$$\begin{aligned} \mathcal{L}(\theta) &= \underset{\theta}{\operatorname{argmin}} \sum_{i}^{N} \rho(f_{MLP}(x_i, \theta) - y_i) \\ &= \underset{\theta}{\operatorname{argmin}} \sum_{i}^{N} \rho(g_2(W_2 \cdot g_1(W_1 \cdot x_i + B_1) + B_2) - y_i) \end{aligned} \tag{A.23}$$

where $\theta = \{W_1, W_2, B_1, B_2\}$ is the set of learned parameters and ρ denotes a loss function (e.g. ℓ_1-norm or ℓ_2-norm). The role of optimization algorithms is very important in training neural networks. The better optimization techniques result in faster convergence to global minimum, which is the optimal solution of the objective function. One of the classic methods for neural network optimization is a mini-batch stochastic gradient descent with momentum (SGD) [LeCun et al., 1998]. SGD proposes to update the network parameters θ at iteration $t+1$ using the negative gradient of the objective function $\nabla\mathcal{L}(\theta_t)$ at iteration t, described as:

$$\begin{aligned} V_{t+1} &= \mu V_t - \alpha \nabla \mathcal{L}(\theta_t) \\ \theta_{t+1} &= \theta_t + V_{t+1} \end{aligned} \tag{A.24}$$

where V_t denotes the weight update, μ and α are respectively the momentum and learning rate. However, when the optimization process gets closer to a minimum, an fixed momentum causes numerical instabilities. Nesterov's accelerated gradient (NAG) [Nesterov, 1983] proposes to calculate the gradient with added momentum, using the following update:

$$\begin{aligned} V_{t+1} &= \mu V_t - \alpha \nabla \mathcal{L}(\theta_t + \mu V_t) \\ \theta_{t+1} &= \theta_t + V_{t+1} \end{aligned} \tag{A.25}$$

The gradient descent optimization with small learning rates could be lead to slow convergence. On the other hand, high learning rates may lead to vanishing gradients [Bengio et al., 1994, Glorot and Bengio, 2010]. In order to address this issue, the SGD method with an adjustable gradient clipping (SGD-GC) [Pascanu et al., 2013] proposes to scale the gradients over a threshold γ to achieve an optimization with high learning rates (e.g. $\alpha = 0.1$) as follows:

$$\nabla\mathcal{L}(\theta) = \begin{cases} \nabla\mathcal{L}(\theta)/\gamma & \| \nabla\mathcal{L}((\theta) \| > \gamma \\ \nabla\mathcal{L}(\theta) & otherwise \end{cases} \tag{A.26}$$

SGD-GC may not converge quickly because of the predefined clipping range. One family of optimization methods addresses this issue through an automatic adaption of the learning rate for each parameter as RMSProp (root-mean-square propagation) [Tieleman and Hinton,

2012] and Adam (adaptive moment estimation) [Kingma and Ba, 2015]. RMSProp method proposes to rescale the gradients to update trainable weights by the root mean square of its second moments u as:

$$\begin{aligned} u_t &= \delta u_{t-1} + (1-\delta)\nabla\mathcal{L}(\theta_t)^2 \\ \theta_{t+1} &= \theta_t - \alpha\frac{\nabla\mathcal{L}(\theta_t)}{\sqrt{u_t}} \end{aligned} \tag{A.27}$$

where δ is called RMSProp decay. However, RMSProp, which does not take account of the first moment of gradients and bias corrections, may induce divergence or very large step sizes [Kingma and Ba, 2015]. Adam method uses a first-order stochastic gradient-based optimization, which relies on adaptive estimates of both the first and second moments of the gradients (m, u). The Adam method applies the following update:

$$\begin{aligned} m_t &= \beta_1 m_{t-1} + (1-\beta_1)\nabla\mathcal{L}(\theta_t) \\ u_t &= \beta_2 u_{t-1} + (1-\beta_2)\nabla\mathcal{L}(\theta_t)^2 \\ \hat{m}_t &= m_t/(1-(\beta_1)^t) \\ \hat{u}_t &= u_t/(1-(\beta_2)^t) \\ \theta_{t+1} &= \theta_t - \alpha\frac{\hat{m}_t}{\sqrt{\hat{u}_t+\epsilon}} \end{aligned} \tag{A.28}$$

where β_1 and β_2 are the first and second moment decay rates, and α is a predefined parameter. $(\hat{m}_j)_t$ and $(\hat{v}_j)_t$ are called respectively the moment bias corrections of the first and second moment estimates. For further information of other optimization methods, we can refer to [Goodfellow et al., 2016].

A.7 Discussion

Previously, we have introduced many different architectures of CNNs and the structure of each model: from the simple perceptron to the convolutional neural networks. A perceptron may be viewed as a neuron and then a set of this element composes a network. For many image processing tasks, neural networks take advantage of the convolution operation in order to better capture the features of higher-dimensional data. Then, deeper networks (e.g. residual networks) may achieve better performance in many applications such as classification but they need many parameters to train. The densely connected networks show the potential of decreasing the depth of networks but also maintaining the good performance. On the other hand, convolutional neural networks rely on the paired training set. The study of adversarial networks is potential to solve unsupervised learning problems. In addition, the applications of these networks have been provided in order to give readers a general look.

List of Figures

List of Tables

Bibliography

Alexander, D. C., Zikic, D., Ghosh, A., Tanno, R., Wottschel, V., Zhang, J., Kaden, E., Dyrby, T. B., Sotiropoulos, S. N., Zhang, H., et al., 2017. Image quality transfer and applications in diffusion mri. NeuroImage 152, 283–298.

Arjovsky, M., Chintala, S., Bottou, L., 2017. Wasserstein generative adversarial networks. In: International Conference on Machine Learning. pp. 214–223.

Ballester, M. A. G., Zisserman, A. P., Brady, M., 2002. Estimation of the partial volume effect in mri. Medical image analysis 6 (4), 389–405.

Beare, R. J., Chen, J., Kelly, C. E., Alexopoulos, D., Smyser, C. D., Rogers, C. E., Loh, W. Y., Matthews, L. G., Cheong, J. L., Spittle, A. J., et al., 2016. Neonatal brain tissue classification with morphological adaptation and unified segmentation. Frontiers in neuroinformatics 10, 12.

Bengio, Y., Simard, P., Frasconi, P., 1994. Learning long-term dependencies with gradient descent is difficult. IEEE transactions on neural networks 5 (2), 157–166.

Bermudez, C., Plassard, A. J., Davis, L. T., Newton, A. T., Resnick, S. M., Landman, B. A., 2018. Learning implicit brain mri manifolds with deep learning. SPIE Medical Imaging.

Borman, S., Stevenson, R. L., 1998. Super-resolution from image sequences-a review. In: Circuits and Systems, 1998. Proceedings. 1998 Midwest Symposium on. IEEE, pp. 374–378.

Chang, H., Yeung, D.-Y., Xiong, Y., 2004. Super-resolution through neighbor embedding. In: Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on. Vol. 1. IEEE, pp. I–I.

Charbonnier, P., Blanc-Féraud, L., Aubert, G., Barlaud, M., 1997. Deterministic edge-preserving regularization in computed imaging. IEEE Transactions on Image Processing 6 (2), 298–311.

Charron, O., Lallement, A., Jarnet, D., Noblet, V., Clavier, J.-B., Meyer, P., 2018. Automatic detection and segmentation of brain metastases on multimodal mr images with a deep convolutional neural network. Computers in biology and medicine 95, 43–54.

Chen, Y., Shi, F., Christodoulou, A. G., Xie, Y., Zhou, Z., Li, D., 2018a. Efficient and accurate MRI super-resolution using a generative adversarial network and 3D multi-level densely connected network. In: MICCAI. pp. 91–99.

Chen, Y., Xie, Y., Zhou, Z., Shi, F., Christodoulou, A. G., Li, D., 2018b. Brain mri super resolution using 3d deep densely connected neural networks. In: Biomedical Imaging (ISBI 2018), 2018 IEEE 15th International Symposium on. IEEE.

Choi, Y., Choi, M., Kim, M., Ha, J.-W., Kim, S., Choo, J., 2018. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation.

Clevert, D.-A., Unterthiner, T., Hochreiter, S., 2015. Fast and accurate deep network learning by exponential linear units (elus). arXiv preprint arXiv:1511.07289.

Cordier, N., Dec. 2015. Multi-atlas patch-based segmentation and synthesis of brain tumor MR images. Theses, Université Nice Sophia Antipolis.
URL https://tel.archives-ouvertes.fr/tel-01237853

Coupé, P., Yger, P., Prima, S., Hellier, P., Kervrann, C., Barillot, C., 2008. An optimized blockwise nonlocal means denoising filter for 3-d magnetic resonance images. IEEE transactions on medical imaging 27 (4), 425–441.

Dai, D., Timofte, R., Van Gool, L., 2015. Jointly optimized regressors for image super-resolution. In: Computer Graphics Forum. Vol. 34. Wiley Online Library, pp. 95–104.

Dong, C., Loy, C. C., He, K., Tang, X., 2014. Learning a deep convolutional network for image super-resolution. In: Computer Vision–ECCV 2014. Springer, pp. 184–199.

Dong, C., Loy, C. C., He, K., Tang, X., 2016a. Image super-resolution using deep convolutional networks. IEEE transactions on pattern analysis and machine intelligence 38 (2), 295–307.

Dong, C., Loy, C. C., Tang, X., 2016b. Accelerating the super-resolution convolutional neural network. In: European Conference on Computer Vision. Springer, pp. 391–407.

Dumoulin, V., Shlens, J., Kudlur, M., 2017. A learned representation for artistic style. Proc. of ICLR.

Efrat, N., Glasner, D., Apartsin, A., Nadler, B., Levin, A., 2013. Accurate blur models vs. image priors in single image super-resolution. In: Computer Vision (ICCV), 2013 IEEE International Conference on. IEEE, pp. 2832–2839.

Fablet, R., Rousseau, F., 2016. Joint interpolation of multisensor sea surface temperature fields using nonlocal and statistical priors. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 9 (6), 2665–2675.

Fogtmann, M., Seshamani, S., Kroenke, C., Cheng, X., Chapman, T., Wilm, J., Rousseau, F., Studholme, C., 2014. A unified approach to diffusion direction sensitive slice registration

and 3-d dti reconstruction from moving fetal brain anatomy. IEEE transactions on medical imaging 33 (2), 272–289.

Freeman, W. T., Jones, T. R., Pasztor, E. C., 2002. Example-based super-resolution. IEEE Computer graphics and Applications 22 (2), 56–65.

Fukushima, K., Miyake, S., 1982. Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition. In: Competition and cooperation in neural nets. Springer, pp. 267–285.

Gal, Y., Ghahramani, Z., 2016. A theoretically grounded application of dropout in recurrent neural networks. In: Advances in neural information processing systems. pp. 1019–1027.

Gatys, L. A., Ecker, A. S., Bethge, M., 2016. Image style transfer using convolutional neural networks. In: Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on. IEEE, pp. 2414–2423.

Gholipour, A., Estroff, J. A., Warfield, S. K., 2010. Robust super-resolution volume reconstruction from slice acquisitions: application to fetal brain mri. IEEE transactions on medical imaging 29 (10), 1739–1758.

Girshick, R., Donahue, J., Darrell, T., Malik, J., 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 580–587.

Glasner, D., Bagon, S., Irani, M., 2009. Super-resolution from a single image. In: Computer Vision, 2009 IEEE 12th International Conference on. IEEE, pp. 349–356.

Glorot, X., Bengio, Y., 2010. Understanding the difficulty of training deep feedforward neural networks. In: Aistats. Vol. 9. pp. 249–256.

Gonzalez, R. C., Woods, R. E., 2006. Digital Image Processing (3rd Edition). Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

Goodfellow, I., Bengio, Y., Courville, A., Bengio, Y., 2016. Deep learning. Vol. 1. MIT press Cambridge.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial nets. In: Advances in neural information processing systems. pp. 2672–2680.

Greenspan, H., 2008. Super-resolution in medical imaging. The Computer Journal 52 (1), 43–63.

Gregor, K., LeCun, Y., 2010. Learning fast approximations of sparse coding. In: Proceedings of the 27th International Conference on International Conference on Machine Learning. Omnipress, pp. 399–406.

Gu, S., Zuo, W., Xie, Q., Meng, D., Feng, X., Zhang, L., 2015. Convolutional sparse coding for image super-resolution. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1823–1831.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A. C., 2017. Improved training of wasserstein gans. In: Advances in Neural Information Processing Systems. pp. 5769–5779.

Han, X., 2017. Mr-based synthetic ct generation using a deep convolutional neural network method. Medical physics 44 (4), 1408–1419.

He, K., Gkioxari, G., Dollár, P., Girshick, R., 2017. Mask r-cnn. In: Computer Vision (ICCV), 2017 IEEE International Conference on. IEEE, pp. 2980–2988.

He, K., Zhang, X., Ren, S., Sun, J., 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE international conference on computer vision. pp. 1026–1034.

He, K., Zhang, X., Ren, S., Sun, J., 2016a. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778.

He, K., Zhang, X., Ren, S., Sun, J., 2016b. Identity mappings in deep residual networks. In: European Conference on Computer Vision. Springer, pp. 630–645.

He, Y., Yap, K.-H., Chen, L., Chau, L.-P., 2009. A soft map framework for blind super-resolution image reconstruction. Image and Vision Computing 27 (4), 364–373.

Hou, H., Andrews, H., 1978. Cubic splines for image interpolation and digital filtering. IEEE Transactions on acoustics, speech, and signal processing 26 (6), 508–517.

Huang, G., Liu, Z., Weinberger, K. Q., van der Maaten, L., 2017a. Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. Vol. 1. p. 3.

Huang, J.-B., Singh, A., Ahuja, N., 2015a. Single image super-resolution from transformed self-exemplars. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5197–5206.

Huang, J.-J., Siu, W.-C., Liu, T.-R., 2015b. Fast image interpolation via random forests. IEEE Transactions on Image Processing 24 (10), 3232–3245.

Huang, Y., Shao, L., Frangi, A. F., 2017b. Simultaneous super-resolution and cross-modality synthesis of 3d medical images using weakly-supervised joint convolutional sparse coding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6070–6079.

Huang, Y., Shao, L., Frangi, A. F., 2018. Cross-modality image synthesis via weakly coupled and geometry co-regularized joint dictionary learning. IEEE transactions on medical imaging 37 (3), 815–827.

Hughes, E., Grande, L. C., Murgasova, M., Hutter, J., Price, A., Gomes, A. S., Allsop, J., Steinweg, J., Tusor, N., Wurie, J., et al., 2017. The developing human connectome: announcing the first release of open access neonatal brain imaging. Organization for Human Brain Mapp, 25–29.

Huynh, T., Gao, Y., Kang, J., Wang, L., Zhang, P., Lian, J., Shen, D., 2016. Estimating ct image from mri data using structured random forest and auto-context model. IEEE transactions on medical imaging 35 (1), 174–183.

Iglesias, J. E., Konukoglu, E., Zikic, D., Glocker, B., Van Leemput, K., Fischl, B., 2013. Is synthesizing mri contrast useful for inter-modality analysis? In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 631–638.

Ioffe, S., Szegedy, C., 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International Conference on Machine Learning. pp. 448–456.

Irani, M., Peleg, S., 1991. Improving resolution by image registration. CVGIP: Graphical models and image processing 53 (3), 231–239.

Isola, P., Zhu, J.-Y., Zhou, T., Efros, A. A., 2017. Image-to-image translation with conditional adversarial networks.

Jain, S., Sima, D. M., Nezhad, F. S., Hangel, G., Bogner, W., Williams, S., Van Huffel, S., Maes, F., Smeets, D., 2017. Patch-based super-resolution of mr spectroscopic images: Application to multiple sclerosis. Frontiers in neuroscience 11.

Jenkinson, M., Beckmann, C. F., Behrens, T. E., Woolrich, M. W., Smith, S. M., 2012. Fsl. Neuroimage 62 (2), 782–790.

Jia, Y., Gholipour, A., He, Z., Warfield, S. K., 2017. A new sparse representation framework for reconstruction of an isotropic high spatial resolution mr volume from orthogonal anisotropic resolution scans. IEEE Transactions on Medical Imaging 36 (5), 1182–1193.

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T., 2014. Caffe: Convolutional architecture for fast feature embedding. In: Proceedings of the 22nd ACM international conference on Multimedia. ACM, pp. 675–678.

Jog, A., Carass, A., Pham, D. L., Prince, J. L., 2014. Random forest flair reconstruction from t 1, t 2, and p d-weighted mri. In: Biomedical Imaging (ISBI), 2014 IEEE 11th International Symposium on. IEEE, pp. 1079–1082.

Jog, A., Carass, A., Prince, J. L., 2016. Self super-resolution for magnetic resonance images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 553–560.

Jog, A., Carass, A., Roy, S., Pham, D. L., Prince, J. L., 2017. Random forest regression for magnetic resonance image synthesis. Medical image analysis 35, 475–488.

Jog, A., Roy, S., Carass, A., Prince, J. L., 2013. Magnetic resonance image synthesis through patch regression. In: Biomedical Imaging (ISBI), 2013 IEEE 10th International Symposium on. IEEE, pp. 350–353.

Johnson, J., Alahi, A., Fei-Fei, L., 2016. Perceptual losses for real-time style transfer and super-resolution. In: European Conference on Computer Vision. Springer, pp. 694–711.

Kainz, B., Steinberger, M., Wein, W., Kuklisova-Murgasova, M., Malamateniou, C., Keraudren, K., Torsney-Weir, T., Rutherford, M., Aljabar, P., Hajnal, J. V., et al., 2015. Fast volume reconstruction from motion corrupted stacks of 2d slices. IEEE transactions on medical imaging 34 (9), 1901–1913.

Kamnitsas, K., Ledig, C., Newcombe, V. F., Simpson, J. P., Kane, A. D., Menon, D. K., Rueckert, D., Glocker, B., 2017. Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation. Medical Image Analysis 36, 61–78.

Kim, J., Lee, J. K., Lee, K. M., 2016a. Accurate image super-resolution using very deep convolutional networks. In: in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition.

Kim, J., Lee, J. K., Lee, K. M., 2016b. Deeply-recursive convolutional network for image super-resolution. In: in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition.

Kim, K. I., Kwon, Y., 2010. Single-image super-resolution using sparse regression and natural image prior. IEEE transactions on pattern analysis and machine intelligence 32 (6), 1127–1133.

Kim, T., Cha, M., Kim, H., Lee, J., Kim, J., 2017. Learning to discover cross-domain relations with generative adversarial networks. In: International Conference on Machine Learning.

Kingma, D., Ba, J., 2015. Adam: A method for stochastic optimization. In: International Conference on Learning Representations.

Krizhevsky, A., Sutskever, I., Hinton, G. E., 2012. Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. pp. 1097–1105.

Kroon, D.-J., Slump, C. H., 2009. Mri modalitiy transformation in demon registration. In: Biomedical Imaging: From Nano to Macro, 2009. ISBI'09. IEEE International Symposium on. IEEE, pp. 963–966.

Lai, W.-S., Huang, J.-B., Ahuja, N., Yang, M.-H., 2017. Deep laplacian pyramid networks for fast and accurate super-resolution. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. pp. 624–632.

Landman, B. A., Huang, A. J., Gifford, A., Vikram, D. S., Lim, I. A. L., Farrell, J. A., Bogovic, J. A., Hua, J., Chen, M., Jarso, S., et al., 2011. Multi-parametric neuroimaging reproducibility: a 3-t resource study. Neuroimage 54 (4), 2854–2866.

LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., 1998. Gradient-based learning applied to document recognition. Proceedings of the IEEE 86 (11), 2278–2324.

Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al., 2017. Photo-realistic single image super-resolution using a generative adversarial network. in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition.

Leroy, F., Mangin, J.-F., Rousseau, F., Glasel, H., Hertz-Pannier, L., Dubois, J., Dehaene-Lambertz, G., 2011. Atlas-free surface reconstruction of the cortical grey-white interface in infants. PloS one 6 (11), e27128.

Lim, B., Son, S., Kim, H., Nah, S., Lee, K. M., 2017. Enhanced deep residual networks for single image super-resolution. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. Vol. 1. p. 3.

Liu, J., Musialski, P., Wonka, P., Ye, J., 2013. Tensor completion for estimating missing values in visual data. IEEE transactions on pattern analysis and machine intelligence 35 (1), 208–220.

Luo, J., Mou, Z., Qin, B., Li, W., Yang, F., Robini, M., Zhu, Y., 2017. Fast single image super-resolution using estimated low-frequency k-space data in mri. Magnetic resonance imaging 40, 1–11.

Maas, A. L., Hannun, A. Y., Ng, A. Y., 2013. Rectifier nonlinearities improve neural network acoustic models. In: Proc. icml. Vol. 30. p. 3.

Mahendran, A., Vedaldi, A., 2015. Understanding deep image representations by inverting them. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5188–5196.

Makropoulos, A., Counsell, S. J., Rueckert, D., 2017. A review on automatic fetal and neonatal brain MRI segmentation. NeuroImage 170, 231–248.

Makropoulos, A., Gousias, I. S., Ledig, C., Aljabar, P., Serag, A., Hajnal, J. V., Edwards, A. D., Counsell, S. J., Rueckert, D., 2014. Automatic whole brain MRI segmentation of the developing neonatal brain. IEEE Transactions on Medical Imaging 33 (9), 1818–1831.

Makropoulos, A., Robinson, E. C., Schuh, A., Wright, R., Fitzgibbon, S., Bozek, J., Counsell, S. J., Steinweg, J., Vecchiato, K., Passerat-Palmbach, J., et al., 2018. The developing human connectome project: A minimal processing pipeline for neonatal cortical surface reconstruction. Neuroimage 173, 88–112.

Manjón, J. V., Coupé, P., Buades, A., Collins, D. L., Robles, M., 2010a. Mri superresolution using self-similarity and image priors. Journal of Biomedical Imaging 2010, 17.

Manjón, J. V., Coupé, P., Buades, A., Fonov, V., Collins, D. L., Robles, M., 2010b. Non-local mri upsampling. Medical image analysis 14 (6), 784–792.

Meyer, P., Noblet, V., Mazzara, C., Lallement, A., 2018. Survey on deep learning for radiotherapy. Computers in biology and medicine.

Michaeli, T., Irani, M., 2013. Nonparametric blind super-resolution. In: Computer Vision (ICCV), 2013 IEEE International Conference on. IEEE, pp. 945–952.

Milanfar, P., 2010. Super-resolution imaging. CRC press.

Mirza, M., Osindero, S., 2014. Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784.

Mroueh, Y., Sercu, T., 2017. Fisher gan. In: Advances in Neural Information Processing Systems. pp. 2513–2523.

Nesterov, Y., 1983. A method of solving a convex programming problem with convergence rate o (1/k2). In: Soviet Mathematics Doklady. Vol. 27. pp. 372–376.

Nie, D., Cao, X., Gao, Y., Wang, L., Shen, D., 2016. Estimating ct image from mri data using 3d fully convolutional networks. In: Deep Learning and Data Labeling for Medical Applications. Springer, pp. 170–178.

Nie, D., Trullo, R., Lian, J., Petitjean, C., Ruan, S., Wang, Q., Shen, D., 2017. Medical image synthesis with context-aware generative adversarial networks. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 417–425.

Nie, D., Trullo, R., Lian, J., Wang, L., Petitjean, C., Ruan, S., Wang, Q., Shen, D., 2018. Medical image synthesis with deep convolutional adversarial networks. IEEE Transactions on Biomedical Engineering.

Odena, A., Olah, C., Shlens, J., 06–11 Aug 2017. Conditional image synthesis with auxiliary classifier GANs. In: Proceedings of the 34th International Conference on Machine Learning. Vol. 70 of Proceedings of Machine Learning Research. PMLR, pp. 2642–2651.

Oktay, O., Bai, W., Lee, M., Guerrero, R., Kamnitsas, K., Caballero, J., de Marvao, A., Cook, S., O'Regan, D., Rueckert, D., 2016. Multi-input cardiac image super-resolution using convolutional neural networks. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 246–254.

Park, S. C., Park, M. K., Kang, M. G., 2003. Super-resolution image reconstruction: a technical overview. IEEE signal processing magazine 20 (3), 21–36.

Pascanu, R., Mikolov, T., Bengio, Y., 2013. On the difficulty of training recurrent neural networks. ICML (3) 28, 1310–1318.

Pham, C.-H., Ducournau, A., Fablet, R., Rousseau, F., 2017a. Brain mri super-resolution using deep 3d convolutional networks. In: Biomedical Imaging (ISBI 2017), 2017 IEEE 14th International Symposium on. IEEE, pp. 197–200.

Pham, C.-H., Fablet, R., Rousseau, F., 2017b. Multi-scale brain mri super-resolution using deep 3d convolutional networks.

Poot, D. H., Jeurissen, B., Bastiaensen, Y., Veraart, J., Van Hecke, W., Parizel, P. M., Sijbers, J., 2013. Super-resolution for multislice diffusion tensor imaging. Magnetic resonance in medicine 69 (1), 103–113.

Radford, A., Metz, L., Chintala, S., 2016. Unsupervised representation learning with deep convolutional generative adversarial networks. In: International Conference on Learning Representations.

Ramos-Llordén, G., Arnold, J., Van Steenkiste, G., Jeurissen, B., Vanhevel, F., Van Audekerke, J., Verhoye, M., Sijbers, J., 2017. A unified maximum likelihood framework for simultaneous motion and $t\_\{1\}$ estimation in quantitative mr $t\_\{1\}$ mapping. IEEE transactions on medical imaging 36 (2), 433–446.

Ren, S., He, K., Girshick, R., Sun, J., 2017. Faster r-cnn: towards real-time object detection with region proposal networks. IEEE transactions on pattern analysis and machine intelligence 39 (6), 1137–1149.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. Springer, pp. 234–241.

Rousseau, F., 2008. Brain hallucination. In: European Conference on Computer Vision. Springer, pp. 497–508.

Rousseau, F., Fablet, R., 2018. Residual networks as geodesic flows of diffeomorphisms. arXiv preprint arXiv:1805.09585.

Rousseau, F., Initiative, A. D. N., et al., 2010a. A non-local approach for image super-resolution using intermodality priors. Medical image analysis 14 (4), 594–605.

Rousseau, F., Kim, K., Studholme, C., 2010b. A groupwise super-resolution approach: application to brain mri. In: Biomedical Imaging: From Nano to Macro, 2010 IEEE International Symposium on. IEEE, pp. 860–863.

Rousseau, F., Kim, K., Studholme, C., Koob, M., Dietemann, J.-L., 2010c. On super-resolution for fetal brain mri. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 355–362.

Rousseau, F., Studholme, C., 2013. A supervised patch-based image reconstruction technique: Application to brain mri super-resolution. In: Biomedical Imaging (ISBI), 2013 IEEE 10th International Symposium on. IEEE, pp. 346–349.

Roy, S., Carass, A., Prince, J. L., 2013. Magnetic resonance image example-based contrast synthesis. IEEE transactions on medical imaging 32 (12), 2348–2363.

Roy, S., Carass, A., Shiee, N., Pham, D. L., Prince, J. L., 2010. Mr contrast synthesis for lesion segmentation. In: Biomedical Imaging: From Nano to Macro, 2010 IEEE International Symposium on. IEEE, pp. 932–935.

Rudin, L. I., Osher, S., Fatemi, E., 1992. Nonlinear total variation based noise removal algorithms. Physica D: nonlinear phenomena 60 (1-4), 259–268.

Rueda, A., Malpica, N., Romero, E., 2013. Single-image super-resolution of brain mr images using overcomplete dictionaries. Medical image analysis 17 (1), 113–132.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X., 2016. Improved techniques for training gans. In: Advances in Neural Information Processing Systems. pp. 2234–2242.

Salvador, J., Perez-Pellitero, E., 2015. Naive bayes super-resolution forest. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 325–333.

Scherrer, B., Gholipour, A., Warfield, S. K., 2012. Super-resolution reconstruction to increase the spatial resolution of diffusion weighted images from orthogonal anisotropic acquisitions. Medical image analysis 16 (7), 1465–1476.

Schulter, S., Leistner, C., Bischof, H., 2015. Fast and accurate image upscaling with super-resolution forests. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3791–3799.

Shi, F., Cheng, J., Wang, L., Yap, P., Shen, D., 2015. Lrtv: Mr image super-resolution with low-rank and total variation regularizations. IEEE transactions on medical imaging 34 (12), 2459–2466.

Shi, W., Caballero, J., Huszar, F., Totz, J., Aitken, A. P., Bishop, R., Rueckert, D., Wang, Z., 2016. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1874–1883.

Shocher, A., Cohen, N., Irani, M., 2018. Zero-shot" super-resolution using deep internal learning. In: Conference on computer vision and pattern recognition (CVPR).

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations.

Steenkiste, G., Jeurissen, B., Veraart, J., Den Dekker, A. J., Parizel, P. M., Poot, D. H., Sijbers, J., 2016. Super-resolution reconstruction of diffusion parameters from diffusion-weighted images with different slice orientations. Magnetic resonance in medicine 75 (1), 181–195.

Sun, J., Xu, Z., Shum, H.-Y., 2008. Image super-resolution using gradient profile prior. In: Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on. IEEE, pp. 1–8.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., et al., 2015. Going deeper with convolutions. Cvpr.

Tai, Y., Yang, J., Liu, X., 2017. Image super-resolution via deep recursive residual network. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Vol. 1.

Tajbakhsh, N., Shin, J. Y., Gurudu, S. R., Hurst, R. T., Kendall, C. B., Gotway, M. B., Liang, J., 2016. Convolutional neural networks for medical image analysis: full training or fine tuning? IEEE transactions on medical imaging 35 (5), 1299–1312.

Tappen, M. F., Russell, B. C., Freeman, W. T., 2003. Exploiting the sparse derivative prior for super-resolution and image demosaicing. In: In IEEE Workshop on Statistical and Computational Theories of Vision. Citeseer.

Thévenaz, P., Blu, T., Unser, M., 2000. Image interpolation and resampling. Handbook of medical imaging, processing and analysis 1 (1), 393–420.

Tieleman, T., Hinton, G., 2012. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. COURSERA: Neural networks for machine learning 4 (2), 26–31.

Timofte, R., Agustsson, E., Van Gool, L., Yang, M.-H., Zhang, L., et al., July 2017. Ntire 2017 challenge on single image super-resolution: Methods and results. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops.

Timofte, R., De, V., Van Gool, L., 2013. Anchored neighborhood regression for fast example-based super-resolution. In: Computer Vision (ICCV), 2013 IEEE International Conference on. IEEE, pp. 1920–1927.

Timofte, R., De Smet, V., Van Gool, L., 2014. A+: Adjusted anchored neighborhood regression for fast super-resolution. In: Computer Vision–ACCV 2014. Springer, pp. 111–126.

Timofte, R., Rothe, R., Van Gool, L., 2016. Seven ways to improve example-based single image super resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1865–1873.

Tor Díez, C., Passat, N., Bloch, I., Faisan, S., Bednarek, N., Rousseau, F., 2018. An iterative multi-atlas patch-based approach for cortex segmentation from neonatal MRI. Computerized Medical Imaging and Graphics.
URL https://hal.univ-reims.fr/hal-01761063

Tsai, R., Huang, T., 1984. Multiple frame image restoration and registration. Advances in Computer Vision and Image Processing, 317–339.

Ulyanov, D., Vedaldi, A., Lempitsky, V., 2017. Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis. In: Proc. CVPR.

Van Nguyen, H., Zhou, K., Vemulapalli, R., 2015. Cross-domain synthesis of medical images using efficient location-sensitive deep network. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 677–684.

Van Steenkiste, G., Poot, D. H., Jeurissen, B., Den Dekker, A. J., Vanhevel, F., Parizel, P. M., Sijbers, J., 2017. Super-resolution t1 estimation: Quantitative high resolution t1 mapping from a set of low resolution t1-weighted images with different slice orientations. Magnetic resonance in medicine 77 (5), 1818–1830.

Vemulapalli, R., Van Nguyen, H., Kevin Zhou, S., 2015. Unsupervised cross-modal synthesis of subject-specific scans. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 630–638.

Walter, C., Kruessell, M., Gindele, A., Brochhagen, H., Gossmann, A., Landwehr, P., 2003. Imaging of renal lesions: evaluation of fast mri and helical ct. The British journal of radiology 76 (910), 696–703.

Wang, Q., Tang, X., Shum, H., 2005. Patch based blind image super resolution. In: Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on. Vol. 1. IEEE, pp. 709–716.

Wang, Z., Bovik, A. C., 2009. Mean squared error: Love it or leave it? a new look at signal fidelity measures. IEEE signal processing magazine 26 (1), 98–117.

Wang, Z., Bovik, A. C., Sheikh, H. R., Simoncelli, E. P., 2004. Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing 13 (4), 600–612.

Wang, Z., Liu, D., Yang, J., Han, W., Huang, T., 2015. Deep networks for image super-resolution with sparse prior. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 370–378.

Wein, W., Brunke, S., Khamene, A., Callstrom, M. R., Navab, N., 2008. Automatic ct-ultrasound registration for diagnostic imaging and image-guided intervention. Medical image analysis 12 (5), 577–585.

Wu, J., Zhang, C., Xue, T., Freeman, B., Tenenbaum, J., 2016. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. In: Advances in Neural Information Processing Systems. pp. 82–90.

Xiang, L., Wang, Q., Nie, D., Qiao, Y., Shen, D., 2018. Deep embedding convolutional neural network for synthesizing ct image from t1-weighted mr image. Medical image analysis 47, 31–44.

Yang, C.-Y., Yang, M.-H., 2013. Fast direct super-resolution by simple functions. In: Computer Vision (ICCV), 2013 IEEE International Conference on. IEEE, pp. 561–568.

Yang, J., Wright, J., Huang, T. S., Ma, Y., 2008. Image super-resolution as sparse representation of raw image patches. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition. pp. 1–8.

Yang, J., Wright, J., Huang, T. S., Ma, Y., 2010. Image super-resolution via sparse representation. IEEE transactions on image processing 19 (11), 2861–2873.

Yang, Q., Li, N., Zhao, Z., Fan, X., Chang, E. I., Xu, Y., et al., 2018. Mri image-to-image translation for cross-modality image registration and segmentation. arXiv preprint arXiv:1801.06940.

Ye, D. H., Zikic, D., Glocker, B., Criminisi, A., Konukoglu, E., 2013. Modality propagation: coherent synthesis of subject-specific scans with data-driven regularization. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 606–613.

Yi, Z., Zhang, H., Tan, P., Gong, M., 2017. Dualgan: Unsupervised dual learning for image-to-image translation. In: International Conference on Computer Vision.

Zeyde, R., Elad, M., Protter, M., 2012. On single image scale-up using sparse-representations. In: Curves and Surfaces. Springer, pp. 711–730.

Zhang, Y., Gan, Z., Fan, K., Chen, Z., Henao, R., Shen, D., Carin, L., 2017. Adversarial feature matching for text generation.

Zhao, C., Carass, A., Dewey, B. E., Prince, J. L., 2018. Self super-resolution for magnetic resonance images using deep networks. In: Biomedical Imaging (ISBI 2018), 2018 IEEE 15th International Symposium on. IEEE.

Zhao, H., Gallo, O., Frosio, I., Kautz, J., 2017. Loss functions for neural networks for image processing. IEEE Transactions on Computational Imaging 2017 (TCI).

Zhu, J.-Y., Park, T., Isola, P., Efros, A. A., 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: International Conference on Computer Vision.

UNIVERSITE BRETAGNE LOIRE / MATHSTIC

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

Titre : Apprentissage profond pour la super-résolution et la segmentation d'images médicales.

Mots clés : Analyse d'images, Apprentissage profond, Super-résolution, Segmentation, IRM

Résumé : L'objectif de cette thèse est d'étudier le comportement de différentes représentations d'images, notamment apprentissage profond, dans le contexte d'application en imagerie médicale. Le but est de développer une méthode unifiée efficace pour les applications visées que sont la super résolution, la segmentation et la synthèse. La super-résolution est un procès d'estimation d'une image haute-résolution à partir d'une ou plusieurs images basses résolutions. Dans cette thèse, nous nous concentrons sur la super-résolution unique, c'est-à-dire que l'image haute résolution (HR) est estimée par une image basse-résolution (LR) correspondante. Augmenter la résolution de l'image grâce à la super-résolution est la clé d'une compréhension plus précise de l'anatomie. L'application de la super résolution permet d'obtenir des cartes de segmentation plus précises. Étant donné que deux bases de données qui contiennent les images différentes (par exemple, les images d'IRM et les images de CT), la synthèse est un procès d'estimation d'une image qui est approximative aux images dans la base de données de cible à partir d'une image de la base de données de source. Parfois, certains contrastes tissulaires ne peuvent pas être acquis pendant la séance d'imagerie en raison du temps et des coûts élevés ou de l'absence d'appareils. Une solution possible est à utiliser des méthodes de synthèse d'images médicales pour générer les images avec le contraste différent qui est manquée dans le domaine à cible à partir de l'image du domaine donnée. L'objectif des images synthétiques est d'améliorer d'autres étapes du traitement automatique des images médicales telles que la segmentation, la super-résolution ou l'enregistrement. Dans cette thèse, nous proposons les réseaux neurones pour la super-résolution et la synthèse d'image médicale. Les résultats démontrent le potentiel de la méthode que nous proposons en ce qui concerne les applications médicales pratiques.

Title : Deep learning for medical image super resolution and segmentation.

Keywords : Image Analysis, Deep Learning, Super-Resolution, Segmentation, MRI

Abstract : In this thesis, our motivation is dedicated to studying the behaviors of different image representations and developing a method for super-resolution, cross-modal synthesis and segmentation of medical imaging. Super-Resolution aims to enhance the image resolution using single or multiple data acquisitions. In this work, we focus on single image super-resolution (SR) that estimates the high-resolution (HR) image from one corresponding low-resolution (LR) image. Increasing image resolution through SR is a key to more accurate understanding of the anatomy. The applications of super-resolution have been shown that applying super-resolution techniques leads to more accurate segmentation maps. Sometimes, certain tissue contrasts may not be acquired during the imaging session because of time-consuming, expensive cost or lacking of devices. One possible solution is to use medical image cross-modal synthesis methods to generate the missing subject-specific scans in the desired target domain from the given source image domain. The objective of synthetic images is to improve other automatic medical image processing steps such as segmentation, super-resolution or registration. In this thesis, convolutional neural networks are applied to super-resolution and cross-modal synthesis in the context of supervised learning. In addition, an attempt to apply generative adversarial networks for unpaired cross-modal synthesis brain MRI is described. Results demonstrate the potential of deep learning methods with respect to practical medical applications.