Cross-Modality Feature Learning for Three-Dimensional Brain Image Synthesis

Yawen Huang

Department of Electronic and Electrical Engineering
University of Sheffield

This dissertation is submitted for the degree of
Doctor of Philosophy

August 2018

I would like to dedicate this thesis to my loving parents.

Declaration

I hereby declare that the work contained in this thesis are original and have not been submitted in whole or in part for any other degree or qualification. This thesis is my own work and includes nothing which is the outcome of work done in collaboration except where specifically clarified in the texts and acknowledgements. Some parts of the work presented in this thesis have been, or are due to be published in the following articles.

- **Y. Huang**, and L. Shao. Task-Driven Bidirectional Fault-Aware Adversarial Networks for Three-dimensional Brain Image Analysis. *To be submitted*, 2018.
- **Y. Huang**, and L. Shao. Simultaneous Super-Resolution and Cross-Modality Synthesis in Magnetic Resonance Imaging. *To be submitted*, 2018.
- **Y. Huang**, L. Shao, and A. F. Frangi. Cross-Modality Image Synthesis via Weakly-Coupled and Geometry Co-Regularized Joint Dictionary Learning. *IEEE Transactions on Medical Imaging (TMI)*, 2017.
- **Y. Huang**, L. Shao, and A. F. Frangi. 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 (CVPR)*, July 2017.
- **Y. Huang**, L. Shao, and A. F. Frangi. DOTE: Dual Convolutional Filter Learning for Super-Resolution and Cross-Modality Synthesis in MRI. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 89–98, 2017.
- **Y. Huang**, F. Zhu, L. Shao, and A. F. Frangi. Color Object Recognition via Cross-Domain Learning on RGB-D Images. In *IEEE International Conference on Robotics and Automation (ICRA)* (oral), 2016, pp. 1672–1677.
- **Y. Huang**, L. Beltrachini, L. Shao, and A. F. Frangi. Geometry Regularized Joint Dictionary Learning for Cross-Modality Image Synthesis in Magnetic Resonance Imaging. In

MICCAI Workshop on Simulation and Synthesis in Medical Imaging (SASHIMI) (oral), Springer, 2016, pp. 118–126.

- **Y. Huang**, L. Beltrachini, L. Shao, and A. F. Frangi. Magnetic Resonance Imaging Cross-Modality Synthesis. In *International Society for Magnetic Resonance in Medicine (ISMRM)*, no. 4220, 2016.

- D. Wang, **Y. Huang**, and A. F. Frangi. Region-Enhanced Joint Dictionary Learning for Cross-Modality Synthesis in Diffusion Tensor Imaging. In *MICCAI Workshop on Simulation and Synthesis in Medical Imaging (SASHIMI)*, Springer, 2017, pp. 41–48.

Yawen Huang
August 2018

Acknowledgements

I would like to acknowledge my supervisors Prof. Ling Shao and Prof. Geraint W Jewell for their expert supervisions and continuous encouragements over the past four years. It has been a great pleasure and honor working under the guidance of Prof. Ling, whose profession, foresight and patience for works pulling at me. I am very grateful to him for teaching me not just about some technical skills or the art of communicating ideas, but for teaching me how to think.

A special thanks to Prof. Alejandro F. Frangi for shaping how I think about research and being able to provide considerable intellectual input into the research that I have done to date.

I would like to thank my friends and colleges: Dr. Simon Jones, Dr. Ruomei Yan, Dr. Di Wu, Dr. Fan Zhu, Dr. Mengyang Yu, Dr. Yang Long, Dr. Ziyun Cai, Dr. Feng Zheng, Dr. Zeike Taylor, Dr. Ali Gooya, Dr. Leandro Beltrachini, Dr. Jose Pozo, Dr. Nishant Ravikumar, Dr. Marco Pereanez, Dr. Serkan Cimen, Dr. Bo Dong, Redzuan Bin Abdul Manap, Bingzhang Hu, Yi Zhou, Yunbai Wang, Le Zhang, Yuanjun Lu, Junyu Jiang, Tiantian Dou, and Danyang Wang (who is my best friend). I would also like to thank many other reviewers for their helpful comments and discussions.

I am also very grateful to my thesis examiners Prof. Jianguo Zhang and Dr. Wei Liu, who offered insightful discussions and constructive suggestions for my research. Many special thanks are due to them for serving on reading my thesis, chairing my oral examination and providing valuable feedbacks to refine my thesis.

I would like to thank my assessors Dr. Charith Abhayaratne and Dr. Xiaoli Chu for their invaluable suggestions and comments for assessing my confirmation review report and viva during my first year PhD study.

I specially thank Dr. Thomas Walther in our department for his kind support and suggestions. I also owe a great deal to the postgraduate administrators, Ms. Hilary J. Levesley and Ms. Frances G. Bright.

Thank you to the head of the department Prof. Geraint W Jewell, who has supported me and helped me when I was in trouble. I am very grateful to Prof. Geraint for his continued support and explicit decision during the most difficult times.

I also would like to thank the departmental Learned Society Fund and the British Machine Vision Association for their financial support of my conference trips through my PhD.

Finally, I would like to thank my parents for their selfless love, unconditional support and immense encouragement in the past 26 years.

Abstract

Multi-modality medical imaging is increasingly used for comprehensive assessment of complex diseases in either diagnostic examinations or as part of medical research trials. Different imaging modalities provide complementary information about living tissues. However, multimodal examinations are not always possible due to adversary factors such as patient discomfort, increased cost, prolonged scanning time and scanner unavailability. In addition, in large imaging studies, incomplete records are not uncommon owing to image artifacts, data corruption or data loss, which compromise the potential of multi-modal acquisitions. Moreover, independently of how well an imaging system is, the performance of the imaging equipment usually comes to a certain limit through different physical devices. Additional interferences arise (particularly for medical imaging systems), for example, limited acquisition times, sophisticated and costly equipment and patients with severe medical conditions, which also cause image degradation. The acquisitions can be considered as the degraded version of the original high-quality images.

In this dissertation, we explore the problems of image super-resolution and cross-modality synthesis for one Magnetic Resonance Imaging (MRI) modality from an image of another MRI modality of the same subject using an image synthesis framework for reconstructing the missing/complex modality data. We develop models and techniques that allow us to connect the domain of source modality data and the domain of target modality data, enabling transformation between elements of the two domains. In particular, we first introduce the models that project both source modality data and target modality data into a common multi-modality feature space in a supervised setting. This common space then allows us to connect cross-modality features that depict a relationship between each other, and we can impose the learned association function that synthesizes any target modality image. Moreover, we develop a weakly-supervised method that takes a few registered multi-modality image pairs as training data and generates the desired modality data without being constrained a large number of multi-modality images collection of well-processed (*e.g.*, skull-stripped and strictly registered) brain data. Finally, we propose an approach that provides a generic way of learning a dual mapping between source and target domains while considering both visually high-fidelity synthesis and task-practicability. We demonstrate that this model can be used to take any arbitrary modality and efficiently synthesize the desirable modality data in an unsupervised manner.

We show that these proposed models advance the state-of-the-art on image super-resolution and cross-modality synthesis tasks that need jointly processing of multi-modality images and that we can design the algorithms in ways to generate the practically beneficial data to medical image analysis.

Table of contents

List of figures

List of tables

Nomenclature

Acronyms / Abbreviations

ADMM	Alternating Direction Method of Multipliers
BCCB	Block Circulant with Circulant Block
BP	Basis Pursuit
CNN	Convolutional Neural Network
CSC	Convolutional Sparse Coding
CSF	Cerebral Spinal Fluid
CT	Computerized Tomography
DFT	Discrete Fourier Transform
DFT	Discrete Fourier Transform
DL	Dual Learning
DTI	Diffusion Tensor Imaging
FCSC	Fast Convolutional Sparse Coding
FFT	Fast Fourier Transform
GAN	Generative Adversarial Network
GLA	Generalized Lloyd Algorithm
GM	Grey Matter
HF	High-Frequency

K-SVD	K-Singular Value Decomposition
LF	Low-Frequency
MK-MMD	Multi-Kernel Maximum Mean Discrepancy
MMD	Maximum Mean Discrepancy
MOD	Method of Optimal Directions
MRA	Magnetic Resonance Angiography
MRI	Magnetic Resonance Imaging
NL	Non-Local
NN	Nearest Neighbor
NP-hard	Non-deterministic Polynomial-time Hardness
PCA	Principal Component Analysis
PSNR	Peak Signal to Noise Ratio
QCQP	Quadratically Constrained Quadratic Programing
RKHS	Reproducing Kernel Hilbert Space
sMRI	Structural Magnetic Resonance Imaging
SR	Super-Resolution
SSIM	Structural Similarity Index
T1-w	T1-weighted
PD-w	Proton Density-weighted
T2-w	T2-weighted
WM	White Matter

Chapter 1

Introduction and Literature Review

1.1 Background

The importance of medical imaging for clinical diagnosis, treatment of disease, and medical research has steadily risen over the last decades. Images of difference modalities are usually generated in medical imaging (Fig. 1.1), for example, magnetic resonance imaging (MRI), computed tomography (CT), and positron emission tomography (PET). The resolution and modality diversity of medical acquisitions is constantly improving, but this comes at the cost of expensive equipment, patient comfort, and scanner time availability. Especially for high-quality and multi-modality images, precise and high-resolution scanners are required to extract the most useful data. In real life, these uncertainties may lead to incomplete records owing to image artifacts or corrupted or lost data.

Although multi-modality medical imaging plays an important role in the prevention, detection, treatment and even new technology is being developed to improve human health and welfare, the process of estimating a modality transformation regarding anatomical and/or func-

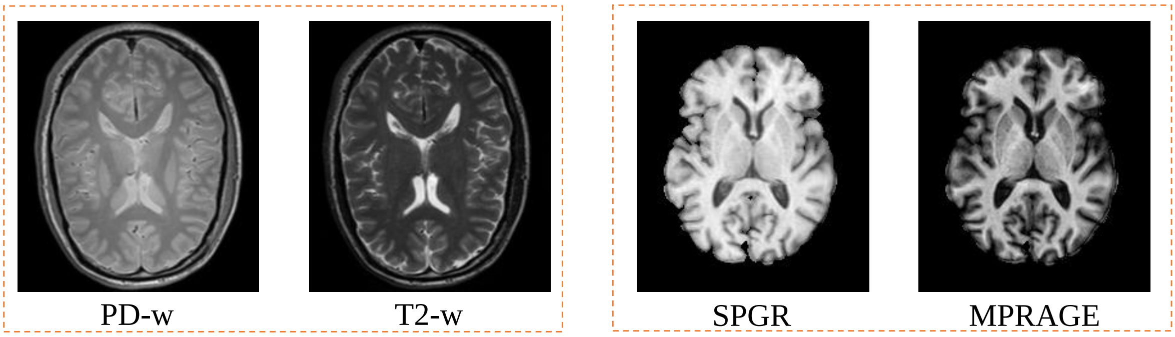


Fig. 1.1 Example of different imaging modalities - PD-w, T2-w, SPGR and MPRAGE images.

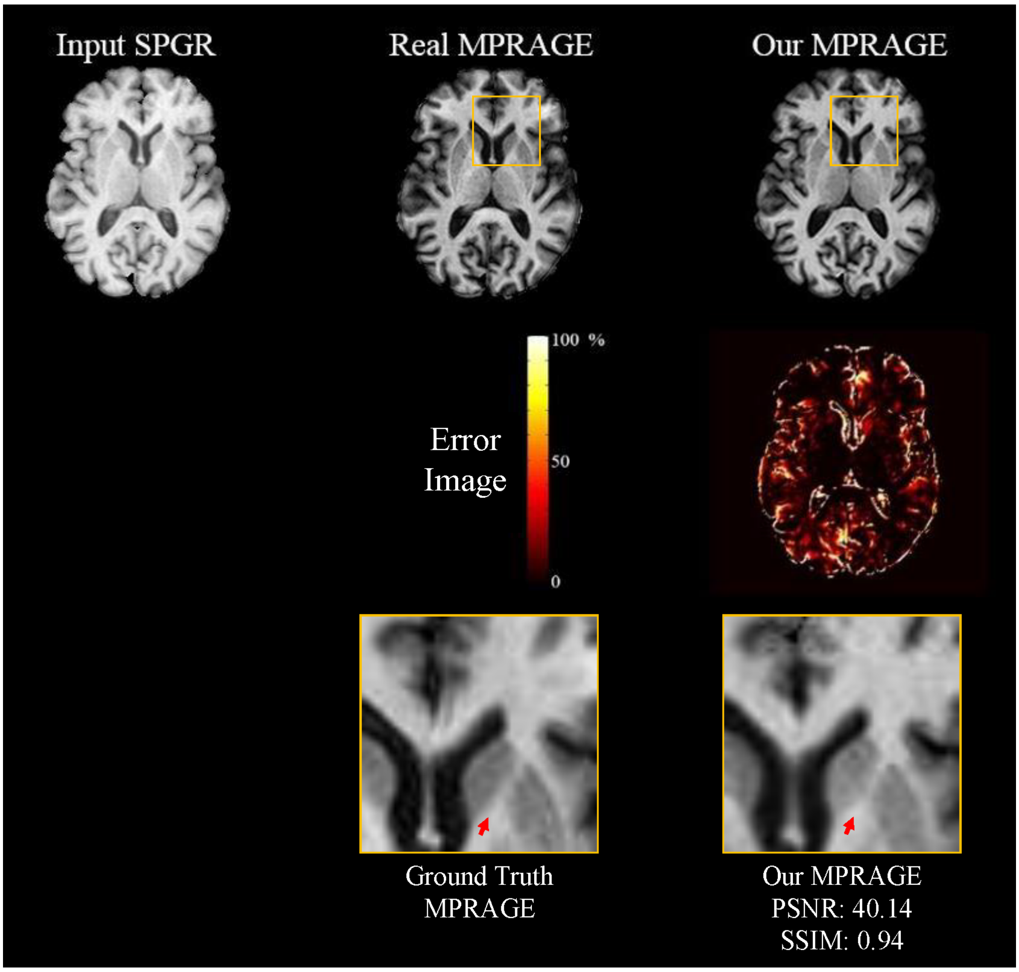


Fig. 1.2 An example of our synthesized result using the patch-based joint dictionary learning method.

tional contrasts between scans is remained, for the most part, unsolved. Despite the fact that the previous effort of researchers focused on the synergy of different modalities, for example, the hybrid positron emission tomography/magnetic resonance imaging [14], no satisfactory algorithms exist that works with simultaneous high-resolution image reconstruction and cross-modality synthesis. The goal of this thesis is to synthesize high-quality image while converting the input modality data to the target modality ones. An example of our synthesized result using the patch-based joint dictionary learning method is shown in Fig. 1.2.

1.2 Imaging Modalities

Medical imaging includes a multitude of imaging modalities such as Magnetic Resonance Imaging, X-ray Computed Tomography, Ultrasound, Positron Emission Tomography, Electrical

Impedance Tomography, etc. Recently, researches have been focused on recovering the missing modalities potentially existing in different modalities of MRI to capture diversified characteristics of the underlying anatomy, especially in the brain research [127, 153].

MRI is a non-ionizing, non-invasive and *in vivo* medical imaging technique used in radiology to create a detailed cross-sectional image of the anatomy and the physiological processes of human body. The acquisitions are generated by forming strong magnetic fields, electric field gradients, and radio waves while avoiding ionizing radiation that can be potentially harmful to the patient. MRI has various modalities for providing useful anatomical and functional diagnostic information, where the contrast of each modality depends on the magnetic properties and number of hydrogen nuclei. By acquiring the comprehensive information, both clinicians and researchers are all likely to benefit from the advances in multi-modality MRI. The clinical applications of such multi-modality MRI exist, containing the assessment of active lesions in multiple sclerosis with MRI.

MRI offers the anatomical information between soft tissues which is usually used to observe/analyze brain activity. The contrast in MR images can be selected for by running different sequences with different weightings. Particularly, there are three main parameters: T1-weighted, T2-weighted, and proton density-weighted. T1-weighted (T1-w) images highlight the differences of longitudinal relaxation time in brain tissue, while T2-weighted (T2-w) images reflect transverse relaxation time, and proton density-weighted (PD-w) images depend on the levels of hydrogen protons within the volume of interest. Tissues with high fat content (e.g. white matter) appear bright and compartments filled with water (e.g. cerebral spinal fluid or CSF) appear dark in T1-w MRI. This is particularly good to depict anatomy. Conversely, in T2-w MRI, compartments filled with water (e.g. CSF) appear bright and those with high fat tissue content (e.g. white matter) appear dark. This is useful for depicting pathology as several lesions (e.g. edema, tumor, infarction, inflammation, and infections) are associated with increased water content. The spoiled gradient recalled (SPGR) pulse sequence spoils the transverse steady state through semi-randomly changing the phase of the RF pulse, leading primarily to PD-w or T1-w contrast. The SPGR sequence is characterized by superior tissue contrast between gray matter and white matter in the brain. As the time of the acquired images can be controlled in a reasonable range, the contrast and spatial resolution of the acquisitions allow the diagnosis to be made with accuracy. The magnetization-prepared rapid gradient-echo (MPRAGE) sequence is designed by using a magnetization-prepared 180 degree inversion pulse for structural brain imaging. The acquisition with T1-w dominance has the improved delineation between gray matter and white matter. The MPRAGE sequence provides high tissue contrast and image quality especially for depicting more focal lesions. Several other contrast mechanisms associate to water diffusion, tissue perfusion, etc.

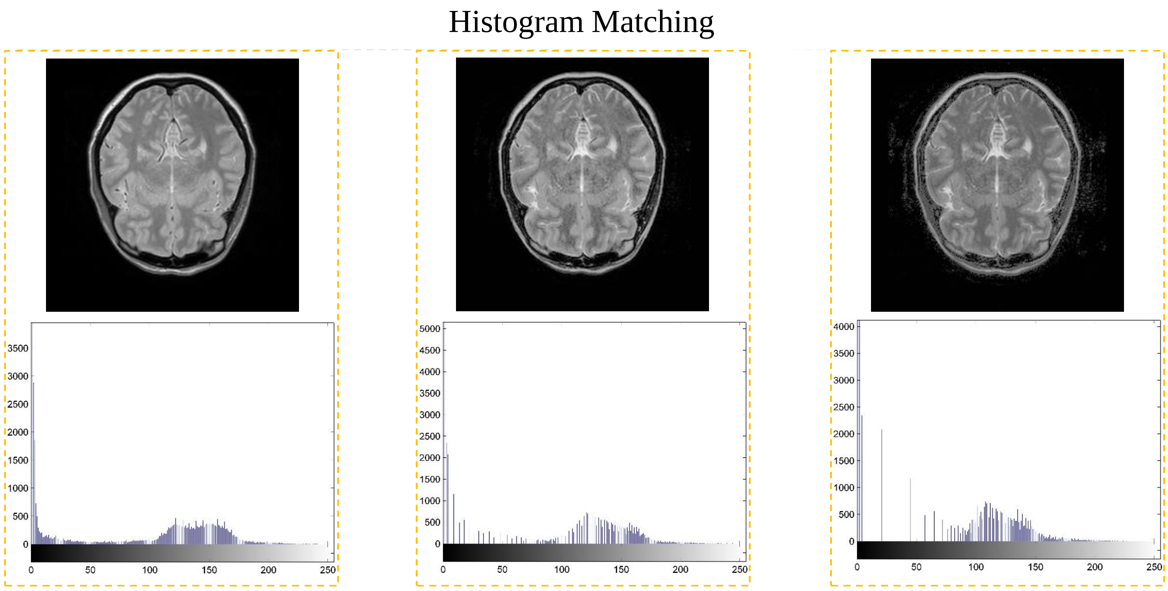


Fig. 1.3 histogram matching model.

Although the abundance of multiple MRI is clinically advantageous, acquisitions suffer from a number of practical problems. In addition, we investigated most of the multi-modality MRI datasets mainly from the brain imaging. To increase diagnosis capabilities and produce more reliable results, synthesizing the desirable modality MRI from the available brain data is the subject of this thesis.

1.3 Challenges

Synthesizing the unavailable data from the available MRI studies is a common necessity in the medical imaging community and hence attracted the considerable amount of research in the past [121, 127]. Nonetheless, such a cross-modality synthesis problem is not satisfactorily solved in many real-life cases and also poses some challenges. Many applications [150] use subject-specific knowledge to synthesize the desired target modality data from the given source modality images. One common criticism is that in a supervised setting the training process becomes more difficult since collecting multi-modality medical images is both time consuming and expensive. These methods are usually restricted to just considering strictly paired data to the entire dataset. Unfortunately, most datasets are non-normalized (containing unpaired single modality data) and it is imperative to apply them to supply auxiliary support about the training requirement. A more recent, but harder, problem, is the cross-modality synthesis in a weakly-supervised or an unsupervised manner. This is a difficult route and challenge arising when the

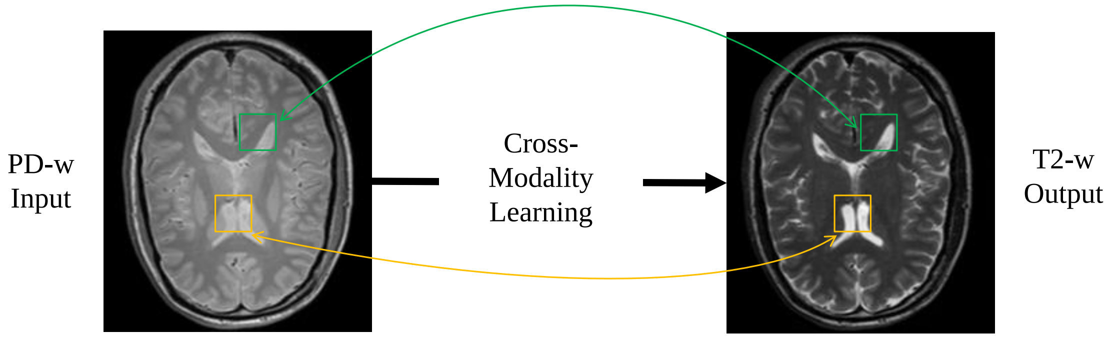


Fig. 1.4 A patch-based cross-modality synthesis schema.

methods attempt to correlate the source modality data and the target modality data. It may feel natural to learn two modalities data in isolation and then compose them to form a relationship. In other words, addressing cross-modality feature learning task jointly needs to formulate an integrated framework that automatically explores the intermediate representations between both modalities during training, without having to explicitly define which two subjects should be aligned while processing the modality-specific image to support the synthesis task. Besides the interplay between different modality subjects and the quantity of data collected, additional problems contain the imaging conditions, the use of histogram matching model (shown in Fig. 1.3), the use of patch-based approaches (shown in Fig. 1.4), effective cross-modality feature representations, and method complexity. Also, how might the researcher even want to utilize the synthesized results in the following applications, e.g., registration, and segmentation. Making the synthesized data not only work but actually be effective of key consideration. A high-level overview of the challenges are given in the following:

1. the lacking of modality-specific information (i.e., missing certain modality data is critical to the feature learning of an algorithm)
2. the lacking of paired data
3. domain discrepancy
4. the relationship between different modality data
5. cross-modality feature learning
6. the fidelity of the synthesized results
7. higher-resolution requirement

8. the validity and effectiveness of the synthesized data

1.4 Dictionary Learning

In computer vision, feature representation is a crucial problem for understanding and learning images. To capture the compact and succinct representation in visual data, a popular way is to adopt dictionary learning to achieve sparse representation using only a few active code elements for representing images. Dictionary leaning for sparse coding has shown promising results in numerous tasks, such as image reconstruction [22, 102, 105, 121, 164], object recognition [66, 71, 180, 185], image super-resolution [43, 167, 171, 178], image denoising [28, 89, 130, 166], and visual classification [81, 169, 170, 181] to name a few.

Dictionary learning-based sparse and redundant representation was first introduced by Olshausen and Field [116] for modeling the spatial receptive fields of simple cells in the mammalian visual cortex. It assumes an ability to represent natural signals (like images) as a linear combination of a few non-zero coefficients of an overcomplete (i.e. the number of basis atoms is greater than the dimension of the data) dictionary. The property of overcomplete is to provide the flexibility in matching data leading to a better approximation of the statistical distribution of the signal. Subsequently, extensive works on the dictionary learning model (according to different criteria) have been investigated in an attempt to understand it better and achieve or improve upon state-of-the-art results. Referring to the most classical ones of dictionary learning, the method of optimal directions (MOD) [37] and the K-Singular Value Decomposition (K-SVD) [1] algorithms have led to the dramatical improvements in infilling missing pixels and image compression tasks.

The MOD method was presented by Engan *et al.* [37] for designing frames to be used for signal compression. This is done by minimizing the representation error to find a dictionary with the corresponding sparse matrix. Similar to the Generalized Lloyd Algorithm (GLA) [45], the MOD is solved by iteratively optimizing over sparse coding and dictionary update steps. Given a set of training data $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ... \mathbf{x}_n]$, the purpose of the MOD algorithm is to learn a reconstructive dictionary $\mathbf{D}$ for obtaining the sparse representation Γ of $\mathbf{X}$, which can be formulated by

$$\begin{aligned} &\arg\min_{\mathbf{D},\Gamma} \|\mathbf{X} - \mathbf{D}\Gamma\|_F^2 \\ &s.t. \|\gamma_i\|_0 \leq T \ \forall i, \end{aligned} \tag{1.1}$$

where $\gamma_i \in \Gamma$ denotes the column vector of Γ, $\|\cdot\|$ is the l_0-norm constraint counting the number of non-zero elements in Γ, $\|\cdot\|_F$ represents the Frobenius norm defined as $\|\mathbf{A}\|_F \triangleq \sqrt{\sum_i \sum_j A_{i,j}^2}$, and T is a sparsity constraint factor that limits the number of non-zero entries of the coefficients.

The resulting optimization problem is highly non-convex, and there is no direct way to find the approximate solution. Instead, an iterative procedure was used in [37] to get a local minimum at best.

While significant steps have been taken to develop the sparsification theory using the MOD, similar problems on the K-SVD algorithm have received competitive attention to efficiently train a generic dictionary for sparse representation. In [1], the K-SVD method was introduced to efficiently learn an overcomplete dictionary from a set of training data. Through learning the dictionary instead of choosing off-the-shelf bases, the K-SVD has been shown to work well in image reconstruction and denoising [34]. The linear decomposition of a signal or an image allows more flexibility to adapt the representation to the data, leading to notable results for various visual inference applications [36, 103, 104, 120, 161, 162]. The K-SVD algorithm is consistent with most of the previous work that relies on iteratively solving sub-problems with the purpose of achieving the optimal solution through an iterative approximation. Despite the wide applications and appealing properties of such an iterative approximation method, the high non-convexity of optimizing the objective function in Eq. (1.1) under the l_0-sparsity penalty measure leads to a general NP-hard (Non-deterministic Polynomial-time hard) problem. Recent results [30, 86] suggest a convexification of the problem posed in Eq. (1.1) by replacing the l_0-norm with an with l_1-norm regularization to enforce sparsity, in which this procedure is also known as Basis Pursuit (BP) [20], or the Lasso [146]. Eq. (1.1) can be rewritten as a joint optimization problem using the l_1-sparsity penalties on the representations:

$$\begin{aligned} &\arg\min_{\mathbf{D},\Gamma} \|\mathbf{X}-\mathbf{D}\Gamma\|_F^2 \\ &s.t. \|\gamma_i\|_1 \leq T \ \forall i, \end{aligned} \tag{1.2}$$

Generally, Eq. (1.2) can be formulated in the form of Lagrange multipliers as

$$\arg\min_{\mathbf{D},\Gamma} \|\mathbf{X}-\mathbf{D}\Gamma\|_F^2 + \lambda \|\Gamma\|_1 , \tag{1.3}$$

where λ is a regularization parameter to balance the sparsity in the objective function.

The optimization problem in Eq. (1.2) is convex in either $\mathbf{D}$ or Γ are fixed, and therefore of iteratively minimizing between sparse coding and dictionary learning. Specifically, when $\mathbf{D}$ is fixed, one optimizes with respect to the sparse codes Γ update (known as a Lasso problem); when Γ is fixed, solving $\mathbf{D}$ becomes to a least square problem with a standard quadratic constraint (known as the quadratically constrained quadratic programming (QCQP) problem).

Given the training data $\mathbf{X}$, we first need to initialize $\mathbf{D}$ with a Gaussian random matrix. Then, an iterative algorithm can be performed alternatively over Γ and $\mathbf{D}$:

1. With $\mathbf{D}$ fixed, sparse coding coefficients Γ can be calculated as

$$\Gamma = \arg\min_{\Gamma} \|\mathbf{X} - \mathbf{D}\Gamma\|_F^2 + \lambda \|\Gamma\|_1 , \tag{1.4}$$

2. With Γ fixed, dictionary $\mathbf{D}$ can be updated by

$$\begin{aligned} \mathbf{D} = \arg\min_{\mathbf{D}} \|\mathbf{X} - \mathbf{D}\Gamma\|_F^2 \\ s.t. \|\mathbf{D}\|_2^2 \leq 1 \ \forall i, \end{aligned} \tag{1.5}$$

Once the optimization is completed, i.e., iteration between step 1 and step 2 until converge, we can get the learned dictionary on which the sparse codes have a stable linear decomposition for image reconstruction.

1.5 Convolutional Sparse Coding

Convolutional Sparse Coding (CSC) [10, 175] has been demonstrated as a promising direction for learning the convolutional image representations in machine learning and computer vision. The concept of CSC is closely related to classic patch-based feature learning and image reconstruction methods [1, 11, 86]. However, feature representation with a patch-based mechanism is highly redundant and may lead to the loss of shifted copies (i.e., overlapped samples) of the same features in the overlapped area of adjacent patches. A more elegant way to solve the above problem, i.e. removing much of the overhead of the patch-based sparse coding, is to use a convolution image formation model for consistently capturing the sparsely-distributed convolutional features. Through decomposing all training data into the defined number of sparse feature maps by the corresponding filters, CSC can avoid missing any latent structures of the underlying signal and thus naturally keep the consistency prior in the decomposition procedure. Recently, CSC has proven essential for many important applications in a wide range of computer vision problems [2, 34, 92, 108, 190]. For example, robust feature learning [77], as part of hierarchical networks in high-level computer vision tasks [77, 175], and low-level image reconstruction [18, 142]. In addition, CSC-based methods have been proposed to solve many practical applications including super-resolution [51], cross-modality synthesis [65], inpaiting [57], demosaicing [23] and reconstruction [57].

To take the property of shift invariance into account, CSC models local interactions through the convolution operator to sparsely encode the whole image. This is done by directly representing an image as the summation of convolutions of the feature maps and the corresponding filters, thus avoiding the sparse decomposition on every single vector. Given a set of input vectors $\{\mathbf{x}\}_{n=1}^{N}$, CSC can be expressed as the following optimization problem:

$$\begin{aligned} \arg\min_{\mathbf{d},\mathbf{z}} \frac{1}{2} \left\| \mathbf{x} - \sum_{i=1}^{K} \mathbf{d}_i * \mathbf{z}_i \right\|_2^2 + \beta \sum_{i=1}^{K} \|\mathbf{z}_i\|_1 \\ s.t. \, \|\mathbf{d}_i\|_1 \leq 1 \; \forall i \in \{1,...,K\}, \end{aligned} \tag{1.6}$$

where each vectorized input image $\mathbf{x}$ can be denoted as the sum of sparse feature maps $\mathbf{z}_i$ convolved with the corresponding filters $\mathbf{d}_i$ of fixed spatial support, $*$ represents the convolution operator processed on the vectorized inputs. The l_2-norm constraint on $\mathbf{d}_i$ ensures the learned filters do not absorb the system energy, and therefore of removing the scaling ambiguity. Rather than averaging in patch-based model, CSC directly approximates the whole image as in the objective of Eq. (1.6) to avoid inconsistent reconstruction.

Despite the benefit of convolutional implementation of sparse coding to solve the inconsistency problem, CSC also brings some difficulties in optimization. Zeiler *et al.* [175] proposed to solve the objective function by an alternation way with the auxiliary variable $\mathbf{t}$, where one solves a set of convex subproblems until convergence and $\mathbf{t}$ is used to separate the convolution from the l_1 regularization. Similar to the solution of conventional sparse coding problem, CSC alternatively updates between processing the subproblem $\mathbf{d}$ given a fixed $\mathbf{z}$, and the subproblem $\mathbf{z}$ given a fixed $\mathbf{d}$. A shortcoming of this method, however, is the computational overhead associated with the iterative subproblems. Bristow *et al.* [10] introduced a fast CSC algorithm through exploiting the property of block circulant with circulant block (BCCB) matrix solving in the Fourier domain. It has been shown the remarkable improvements in efficiency by utilizing Parseval's theorem for solving Eq. (1.6). Following [10], we can reformulate Eq. (1.6) as a constrained optimization problem by involving two auxiliary variables $\mathbf{t}$ and $\mathbf{s}$ in the Fourier domain:

$$\begin{aligned} \arg\min_{\mathbf{d},\mathbf{z},\mathbf{s},\mathbf{t}} \frac{1}{2} \left\| \hat{\mathbf{x}} - \sum_{i=1}^{K} \hat{\mathbf{d}}_i \odot \hat{\mathbf{z}}_i \right\|_2^2 + \beta \sum_{i=1}^{K} \|\mathbf{t}_i\|_1 \\ s.t. \, \|\mathbf{s}_i\|_2^2 \leq 1 \; \forall i \in \{1,...,K\} \\ \mathbf{s}_i = \mathbf{S}\Phi^T \hat{\mathbf{d}}_i \; \forall i \in \{1,...,K\} \\ \mathbf{t}_i = \mathbf{z}_i \; \forall i \in \{1,...,K\}. \end{aligned} \tag{1.7}$$

where the symbol ˆ applied to any vector denotes the frequency representation of a vectorized signal, for example, $\hat{\mathbf{x}} = [F(\mathbf{x}_1)^T, ..., F(\mathbf{x}_N)^T]$ and $F(\cdot)$ is the Fourier transform operator, $\odot$ denotes the Hadamard (component-wise) product, $\mathbf{S}$ projects $\mathbf{d}$ onto a corresponding small spatial support, Φ represents the Discrete Fourier Transform (DFT) matrix, $\mathbf{s}_i$ and $\mathbf{t}_i$ are two

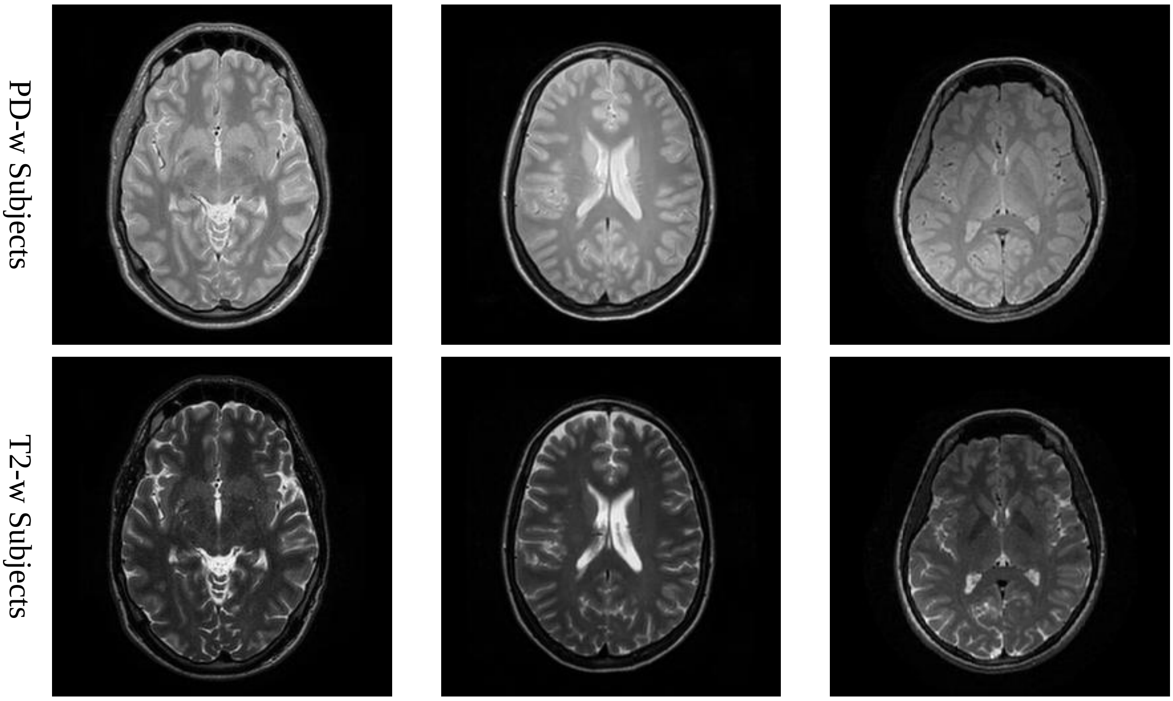


Fig. 1.5 Illustration of the example PD-w and T2-w MR images from the IXI dataset. In each panel, the first row shows the PD-w image while the second row shows the corresponding T2-w data.

slack variables allowing for an explicit and efficient solution to Eq. (1.7) by splitting the objective into several subproblems.

1.6 Datasets

Two distinct datasets were used in this dissertation. The first dataset is taken from the Information eXtraction from Images (IXI)[1] [126] including 578 Magnetic Resonance (MR) images from normal and healthy subjects. The images have been collected at three different hospitals (i.e., Hammersmith Hospital using a Philips 3T system, Guy's Hospital using a Philips 1.5T system, and Institute of Psychiatry using a GE 1.5T system) and stored in NIFTI format. The acquisition protocol for each subject contains: T1-weighted, T2-weighted, Proton Density (PD)-weighted images; Magnetic Resonance Angiography (MRA) images; and Diffusion-weighted images (15 directions). Some examples collected from the IXI dataset are shown in Fig. 1.5.

[1] http://brain-development.org/ixi-dataset/

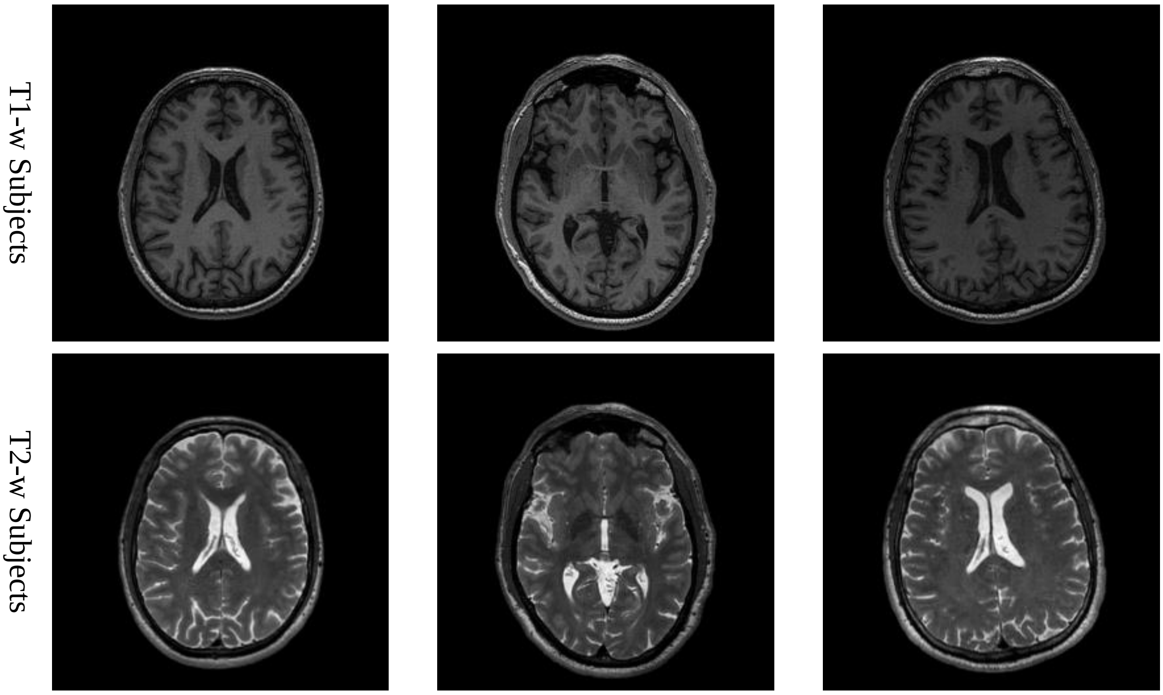


Fig. 1.6 Illustration of the example T1-w and T2-w MR images from the NAMIC dataset. In each panel, the first row shows the T1-w image while the second row shows the corresponding T2-w data.

The second dataset: National Alliance for Medical Image Computing (NAMIC)[2] includes 20 cases. Subjects from the NAMIC dataset were part of a schizophrenic study, specifically, ten are Normal Controls, and ten are Schizophrenic. The data acquisition protocol for each subject has Structural Magnetic Resonance Imaging (sMRI), Diffusion Tensor Imaging (DTI) and Resting State fMRI. In this work, we focus on the structural MRI volumes which are PD-w, T1-w, and T2-w images. Specifically, MR images were acquired with using a 3T GE system at BWH in Boston, MA. The MR image acquisition protocol for each subject has two pulse sequences: contiguous spoiled gradient-recalled acquisition (fastSPGR) (TR=7.4ms, TE=3ms, TI=600, 10 degree flip angle, 25.6cm^2 field of view, matrix=256×256, 1×1×1 mm voxel size); XETA (eXtended Echo Train Acquisition) provides a series of contiguous T2-w data (TR=2500ms, TE=80ms, 25.6 cm^2 field of view, 1 mm slice thickness, 1×1×1 mm voxel size). Some examples sampled from the NAMIC dataset are shown in Fig. 1.6.

[2] http://insight-journal.org/midas/collection/view/190

1.7 Related Work

With the goal to transfer the modality information from the source domain to the target domain, recent developments in cross-modality synthesis, such as texture synthesis [32, 44, 59], face photo-sketch synthesis [42, 158], and multi-modal retrieval [110, 138], have shown promising results. In this thesis, we focus on the problems of image super-resolution and cross-modality synthesis, so only review related methods on these two aspects. To synthesize a target modality image from a source modality image, several approaches have been suggested in the literature with state-of-the-art results [14, 72, 153]. Most of these methods can be broadly referred to as the nature image domain and the medical image domain roughly subdivided based on the type of images.

1.7.1 Nature Image Domain

In the nature image domain, the purpose of image super-resolution (SR) is to reconstruct an high-resolution (HR) image from its low-resolution (LR) counterpart. According to the image priors, image SR methods can be grouped into two main categories: interpolation-based, external or internal data driven learning methods. Interpolation-based SR works, including the classic bilinear [90], bicubic [78], and some follow-up methods [131, 179], interpolate much denser HR grids by the weighted average of the local neighbors. Most modern image SR methods have shifted from interpolation to learning based. These methods focus on learning a compact dictionary or manifold space to relate LR/HR image pairs, and presume that the lost high-frequency (HF) details of LR images can be predicted by learning from either external datasets or internal self-similarity. The external data driven SR approaches [16, 40, 168] exploit a mapping relationship between LR and HR image pairs from a specified external dataset. In the pioneer work of Freeman *et al.* [40], the NN of an LR patch is found, with the corresponding HR patch, and used for estimating HF details in a Markov network. Chang *et al.* [16] projected multiple NNs of the local geometry from the LR feature space onto the HR feature space to estimate the HR embedding. Furthermore, sparse coding-based methods [128, 168] were explored to generate a pair of dictionaries for LR and HR patch pairs to address the image SR problem. Wang *et al.* [155] and Huang *et al.* [60] further suggested modeling the relationship between LR and HR patches in the feature space to relax the strong constraint. Recently, an efficient CNN based approach was proposed in [27], which directly learned an end-to-end mapping between LR and HR images to perform complex nonlinear regression tasks. For internal dataset driven SR methods, this can be built using the similarity searching [125] and/or scale-space pyramid of the given image itself [61].

In parallel, various cross-modality synthesis methods have been proposed for synthesizing unavailable modality data from available source images. One of the well-established modality transformation approaches is the example-based learning method generated by Freeman *et al.* [41]. Given a patch of a test image, several NNs with similar properties are picked from the source image space to reconstruct the target one using Markov random fields. In addition, Hertzmann *et al.* provided image analogies [59], which transfers the texture information from a source modality space onto a target modality space. Gatys *et al.* [44] introduced a CNN algorithm of artistic style, that new images can be generated by performing an image pre-search in high-level image content to match generic feature representations of example images.

1.7.2 Medical Image Domain

In the medical imaging community [127, 150, 153], synthesis algorithms can be summarized into a main family, i.e., example-based methods roughly subdivided in accordance with the size of the training set. Particularly, image SR can be treated as a way of synthesis which attempts to improve image resolution by algorithms instead of carrying out during the acquisition stage. Image SR effectively solved the problem of long acquisition and breath-hold from the requirement of high quality images, therefore the accuracy of clinical diagnosis can be increased while the images are acquired in a reasonable time.

Example-based methods learn the source-target mapping from a very small number of source-target image pairs (e.g. several or even a pair of images) by extracting multiple image patches from the source image and assuming the same sparse codes are shared between source and target modality spaces. One of the well-established cross-modality synthesis approaches in this category is applied to facilitate multi-modal image registration in correlative microscopy [15]. Kroon *et al.* [82] mapped between T1-w and T2-w magnetic resonance images by simply using the peaks in a joint histogram of registered image pairs to transform between source and target image representations. Techniques based on sparse representations have been presented, which separately learn two corresponding dictionaries from registered image pairs and synthesize the target MRI modality data from the patches of the source MRI modality [127]. Specifically, Roy *et al.* [127] used sparse coding for desirable MR contrast synthesis assuming that cross-modality patch pairs have same representations and can be directly used for training dictionaries to estimate the contrast of the target modality. Similar work was also used in [67]. In [4], a canonical correlation analysis-based approach was proposed to yield a feature space that can get underlying common structures of co-registered data for better correlation of dictionary pairs. Recently, Jog *et al.* [72] proposed a nonlinear regression-based image synthesis approach that used registered image pairs to train a random forest regressor for predicting the target from the source image intensity.

Some example-based methods learn the source-target mapping assuming that a large set of source-target modality image pairs (e.g. the whole dataset) is available. These approaches vary on how to generate a model (e.g. learning a dictionary, a manifold or a network) that relates to the number of the patches of the registered image pairs. In measuring the similarity between training and test data of the same modality, Ye *et al.* [172] proposed an iterative patch-based modality propagation approach. For each patch of the test image, a global search was performed comparing the input patch with each patch in the training dataset. The nearest neighbors to the input patch were found in the source domain; the target modality image was synthesized with the corresponding target modality patches. Rather than learning the mapping between both domains in the original data space, coupled dictionary learning [168] can alleviate simple cross-modality heterogeneity in the projected feature space. As an extension, semi-coupled dictionary learning was presented by advancing a linear mapping to model the relationship on the sparse representations from both domains. Burgos *et al.* [14] introduced another framework called pseudo CT synthesis for generating CT-like image from the T1-w or T2-w input using multi-atlas deformable registration and tissue contrast fusion. In [150], a location-sensitive deep network [150] has been put forward to explicitly utilize the voxel image coordinates by incorporating image intensities and spatial information into a deep network for synthesizing purposes. Instead of using coupled image pairs as training data, matching feature representations and learning spatial relations with joint sparse coding [153] has shown great potential in synthesizing images across modalities. To improve the quality of the synthesized images across different modalities, Huang *et al.* [65] proposed to first align weakly-supervised data and then generate super-resolution cross-modality data simultaneously using joint convolutional sparse coding scheme. Inspired by this strategy, we integrate paired and unpaired training data by constructing correspondences across different modalities and leverage weakly-coupled data effectively.

As argued in [153], collecting a large number of multi-modality images is both time-consuming and expensive, and sometimes even impractical in medical imaging. It would be preferable to use an unsupervised approach to deal with input data instead of ensuring data to be coupled invariably. Most of the methods, especially the full-set-based approaches, require considerable amounts of co-registered training data in both source and target domains. Motivated by this and the above works, we propose several more practical cross-modality image synthesis solutions that link source-target domains in either a fully supervised setting or a weakly-coupled fashion, which outperform existing state-of-the-art methods on our experiments.

1.8 Thesis Outline

This dissertation begins with works that address the core algorithmic problem of designing cross-modality synthesis methods to assist comprehensive assessment of complex diseases in either diagnostic examinations or as part of medical research trials. In particular, we develop both supervised and weakly-supervised approaches that process and align the two modalities and train their parameters on datasets of multi-modality brain MRI acquisitions. The rest of this thesis is organized as follows.

In Chapter 2, we consider the problem of collecting features in MRI by their modality-specific properties learned with full supervision. Chapter 3 presents a novel dual convolutional filter learning algorithm for the cross-modality synthesis of brain MRI data, with the key contribution being an algorithm for bringing state-of-the-art 3D method for robust synthesis. Chapter 4 deals with the challenging task of weakly-supervised learning useful for both image super-resolution and cross-modality synthesis. In Chapter 5, we first introduce a geometry regularized joint dictionary learning method in a supervised setting, which is designed to work on brain image for synthesizing the missing/complex modality data. Another part of the model included in Chapter 5 extends the proposed geometrical regularization approach to a weakly-supervised scheme for cross-modality synthesis. Chapter 6 constructs an unsupervised deep learning architecture for automatically synthesize the application/post processing-efficient high-resolution or required modality data with dual learning during training. Finally, Chapter 7 concludes this thesis and discusses future works. The parallel contrast experiments are involved in Appendix A.

1.9 Contributions

In this dissertation, we develop models for the cross-modality synthesis of the three-dimensional brain image. In particular, we develop several learning-based architectures that process and align the two modalities and train models on different public multi-modality brain datasets. Most contributions in this dissertation have first appeared as my publications, which are summarized below[3]:

1. Chapter 2: Simultaneous super-resolution and cross-modality synthesis in magnetic resonance imaging.

2. Chapter 3: Dual convolutional filter learning for super-resolution and cross-modality synthesis in MRI [64].

[3] A more detailed account of contributions appears in Section 7.1. Other publications such as [66] are beyond the scope of this dissertation and thus not discussed here.

3. Chapter 4: Simultaneous super-resolution and cross-modality synthesis of 3D medical images using weakly-supervised joint convolutional sparse coding [65].

4. Chapter 5: Geometry regularized joint dictionary learning for cross-modality image synthesis in magnetic resonance imaging [62]. Cross-Modality Image synthesis via weakly-coupled and geometry co-regularized joint dictionary learning [63].

5. Chapter 6: Task-driven bidirectional fault-aware adversarial networks for three-dimensional brain image analysis.

Chapter 2

Feature-Clustered and Normalized Joint Sparse Representation

Multi-modality Magnetic Resonance Imaging (MRI) has enabled significant progress to both clinical diagnosis and medical research. Applications range from different diagnosis to novel insights into disease mechanisms and phenotypes. However, there exist many practical scenarios where acquiring high-quality multi-modality MRI is restricted, for instance, owing to limited scanning time. This imposes constraints on multi-modality MRI processing tools, e.g. segmentation and registration. Such limitations are not only recurrent in prospective data acquisition, but also when dealing with existing databases with either missing or low quality imaging data. In this chapter, we explore the problem of synthesizing high-resolution images corresponding to one MRI modality from a low-resolution image of another MRI modality of the same subject. This is achieved by introducing the cross-modality dictionary learning scheme and a patch-based globally redundant model based on sparse representations. We use high-frequency multi-modality image features to train dictionary pairs, which are robust, compact, and correlated in this multimodal feature space. A feature clustering step is integrated into the reconstruction framework speeding up the search involved in the reconstruction process. Images are partitioned into a set of overlapping patches to maintain the consistency between neighboring pixels and increase speed further.Extensive experimental validations on two multi-modality databases of real brain MR images show that the proposed method outperforms state-of-the-art algorithms in two challenging tasks: image super-resolution and simultaneous SR and cross-modality synthesis. Our method was assessed on both healthy subjects and patients suffering from schizophrenia with excellent results.

2.1 Introduction

Magnetic Resonance Imaging (MRI) has advanced both clinical diagnosis and biomedical research in the neurosciences. MRI has widely been used given its non-invasiveness and the versatility associated with multi-modality imaging protocols that unravel both brain structure and function. Each MRI sequence (hereafter called an MRI modality) is based upon different image contrast mechanisms that relate to complementary properties of brain tissue structure and function and help to unravel anatomical differences and physiologic alterations of brain tissue in health and disease [121].

Despite these benefits, acquiring a full battery of MRI modalities faces constraints associated with increased scanning costs, limited availability of scanning time, and patient comfort, among others. Also, as MRI technologies improve, enhanced resolution or new contrast mechanisms can be utilized. However, in longitudinal imaging cohorts, its benefits will not be available retrospectively for earlier time points in the study, imposing a natural limitation on the dataset. This brings an additional complexity to image analysis and interpretation as the imaging protocol can change. Finally, many reasons can lead to incomplete records for a subject who took part in a large imaging study owing to imaging artifacts, acquisition errors, and lost or corrupted data sets. In all such scenarios, it would be desirable to have a mechanism to synthesize the high-resolution missing data in a different modality with the available MRI modality. However, most of the existing methods tackling this problem either focuses on image super-resolution (SR) or cross-modality synthesis, but not on solving both problems jointly.

Image SR aims to reconstruct a high-resolution (HR) image from a low-resolution (LR) counterpart. It is an under determined inverse problem since a multiplicity of solutions exist for the LR input. To solve such a problem, solution space is often constrained by involving strong prior information. In the early years of studies, some simple interpolation-based smooth methods [50, 58, 87, 140] were proposed to zoom up LR images. However, Van Ouwerkerk [151] pointed out that such interpolation methods cannot recover detailed information lost in the down-sampling procedure, and even may blur sharp edges. SR techniques were then proposed [16, 47, 117, 124, 128, 133, 168], which take the degradation model (e.g. blurring, noise, and down-sampling effects) into account, to reconstruct the image with much higher accuracy. Such methods estimate the HR image by learning co-occurrence priors between the LR and HR image pairs [168]. For instance, Freeman *et al.* [41] presented a learning-based approach to estimate an HR image from an LR input via Markov Network and Bayesian Belief Propagation. Although the resolution can generally be improved effectively, corners, edges, and ridges are still blurred. Based on such a strategy, Sun *et al.* [141] addressed the above problem by a computationally intensive process of analysis of millions of LR-HR patch pairs. Neighbor Embedding was then proposed for single-image SR [16]. This consists of projecting the local geometry from the LR feature space onto the HR feature space to estimate the HR embedding. Although a small dataset was used in the training process (partly to solve the massive computational load) results were confined to the small number of neighbors. To adequately recover the general image structure and its details, Non-Local Means (NL means) [106, 107] was presented to reconstruct the HR image with noise suppression exploiting image self-similarities. However, for strong denoising levels, images are visually over-smooth. Recently, sparse representations were exploited for solving the SR problem. For example, Yang *et al.* [168] adopted a joint dictionary learning framework for mapping LR and HR image pairs into a common representation space. Rueda *et al.* [128] took advantage of this model and applied it to address the SR problem in brain MRI. A common drawback shared by both methods is that they only consider local image information in the image synthesis leading to suboptimal reconstructions.

In parallel to the SR technique, researchers have been developing methods to solve the problem of cross-modality image synthesis. This problem can be tackled either by transforming MRI intensities across modalities or by synthesizing tissue contrast in the target domain based on patches of the source domain. Histogram matching is a simple way of transforming image intensities from one modality onto another or to normalize histogram ranges across subjects [7, 24, 113, 123, 134]. Applications such as segmentation and registration can benefit from histogram normalization and/or transformation to reduce the dependency of the results to intensity variations across individuals or imaging protocols. Although this method is widely used in neuroimaging (e.g. [7, 24, 113, 123, 134]), it has demonstrated its weakness for converting data with inconsistent intensities and apparent errors [127]. An alternative approach to reconstruct a target MRI modality from a source MRI modality (or more generally, from any other imaging modality) is the example-based image synthesis [127]. In this approach, two dictionaries are independently trained on corresponding patches from registered image pairs of the source and target modalities, respectively. Then the target image is synthesized from the source data based on a reconstruction algorithm that links the patches to reconstruct the source image to the corresponding patches in the target dictionary. Such approaches have also been applied with very promising results to the related problems of label fusion [145] and image hallucination [124]. The procedure to reconstruct the target image imposes that the same code that optimally reconstructs the source patches from the source dictionary must be applied directly to reconstruct the target patches from the target dictionary based on a mapping learned from a set of image pairs. To do so, the most common procedure is to train two dictionaries via random sampling of the registered image patches from two domains and build the correspondence between patches of two modalities. Such methods concatenate both domains according to the intensities of the paired patches, leading to two separate dictionary learning processes in their respective modalities. In this context, the joint representation of the two domains (juxtaposing the two independently-computed codes) is suboptimal regarding a jointly learned code that exploits the cross-modality correlations. In addition, example-based methods rely on the given cross-modality exemplar pairs, and does not capture the rich variability in image texture across a population of subjects. According to the similarity measurement between training and test data of the same modality, Ye *et al.* [172] proposed a patch-based modality propagation method. Through global search, the input patch was compared against the training patches in the dataset. Several nearest neighbors with similar properties were picked from the source domain and corresponding patches in the target modality used for image synthesis. In [14], a pseudo CT synthesis algorithm was proposed, which aims at generating CT-like images from T1-w / T2-w inputs, using multi-atlas registration and tissue contrast fusion. Nguyen *et al.* [150] proposed a location-sensitive deep network method to integrate image intensities and spatial information into a deep network for cross-modality image synthesis. To verify the effectiveness of synthesized data, Tulder *et al.* used restricted Boltzmann machines to learn abstract representations from training data for synthesizing the missing image sequences. More recently, a nonlinear regression-based image synthesis approach [72] was proposed to predict the intensities in the target modality. While training, this method used registered image pairs from source and target modalities to learn a random forest regressor for regressing the target modality data. Besides these methods, Vemulapalli *et al.* proposed an unsupervised approach which relaxed needing registered image pairs during training, to deal with the synthesis problem.

In this chapter, we present a novel MRI Simultaneous Super-Resolution and Cross-Modality Synthesis (SiSCS) method for reconstructing the HR version of the target modality based on an LR image of the source modality while treating each 3D volume as a stack of 2D images. We simultaneously train a cross-modality dictionary pair based on registered patches of the LR source modality and the HR target modality. For an accurate image synthesis, the sparse codes of the LR source modality should be the same as those of the HR ground truth on the premise of high correlation between the paired LR source data and HR target data. We map high-frequency (HF) features of the registered image pairs between source and target modalities into a common feature space to fit the style-specific local structures and resolutions. We introduce patch-based global redundancy, consisting of cross-modal matching and self-similarity, to enhance the quality of image reconstruction based on sparse representations. Prior papers such as [12, 34] and follow-up studies [29, 106, 107, 187] have shown that self-similar image properties were used for enabling exact local image reconstruction. However, classical NL means [165] are computationally expensive. To overcome such problem, we present an integrated clustering algorithm into the original redundancy framework for making the data of the same class correlated, and speeding up the similarity measure from each subclass. In addition, we set patches as the unit to preserve the intrinsic neighbor information of pixels and reduce the computational cost.

In summary, our method offers these four contributions:

1. We normalize the vectors of dictionary pairs in an HF feature space (rather than in the original image space) to a unified range to achieve intensity consistent learning.
2. A novel cross-modality dictionary learning based on a compact representation of HF features in both domains is proposed to derive co-occurrence prior.
3. Simultaneous estimation of the dictionaries corresponding to both modalities, leading to matched representations for a given sparse code.
4. Sparse code pre-clustering provides a globally redundant reconstruction scheme incorporated into the local reconstruction model, enhances the robustness of the synthesis, and speeds up code search.

Extensive experiments on a public dataset of brain MR images show that the proposed method achieves a competitive performance compared to other state-of-the-art algorithms. To the best of our knowledge, this work is the first to undertake SR reconstruction of the specific target MRI modality from an available source MRI LR modality.

2.2 Background

2.2.1 Image Degradation Model

SR image reconstruction, understood as an inverse problem, attempts to recover an HR image in matrix form $\mathbf{X}^H$ from an LR input $\mathbf{X}^L$. A degradation model (Fig. 2.1) is assumed as prior information to solving this inverse problem. In its simplest form, the source LR image $\mathbf{X}^L$ is modeled as a blurred and down-sampled counterpart of its HR image $\mathbf{X}^H$ by:

$$\mathbf{X}^L = \mathscr{S}\mathscr{B}\mathbf{X}^H, \tag{2.1}$$

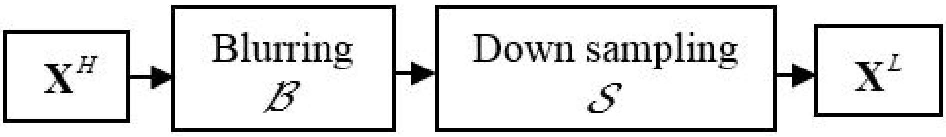


Fig. 2.1 The degradation model.

where $\mathscr{B}$ and $\mathscr{S}$ represent the blurring and down-sampling operators, respectively [151].

2.2.2 Dictionary Learning

Dictionary learning has been successfully applied to a number of problems in image processing, such as image restoration [1, 104, 129], denoising [29, 34, 165], and enhancement [39, 129, 168]. In image reconstruction based on dictionary learning, an image is normally treated as the combination of many patches [1, 34, 104, 129, 168] and denoted as $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N] \in \mathbb{R}^{k \times N}$. An image is approximated as $\mathbf{X} \approx \Phi\mathbf{A}$, where $\mathbf{X}$ is the target matrix being approximated, $\Phi = [\phi_1, \phi_2, ..., \phi_K] \in \mathbb{R}^{k \times K}$ denotes a projection dictionary with K atoms, and $\mathbf{A} = [\alpha_1, \alpha_2, ..., \alpha_N] \in \mathbb{R}^{K \times N}$ is a set of N K-dimensional sparse codes of $\mathbf{X}$ with $\|\mathbf{A}\|_0 \ll K$. Representing Eq. (2.1) for sparse reconstruction of $\mathbf{X}^L$ regarding Φ^L can be achieved by:

$$\mathbf{X}^L \approx \Phi^L \mathbf{A} = \mathscr{B}\mathscr{S}(\Phi^H \mathbf{A}), \tag{2.2}$$

where Φ^L and Φ^H denotes an LR dictionary and an HR dictionary, respectively. For each image, the sparse decomposition is obtained by solving:

$$\min_{\mathbf{A}} \|\mathbf{A}\|_0 \ s.t. \ \mathbf{X} = \Phi\mathbf{A} \ (or \, \|\mathbf{X} - \Phi\mathbf{A}\|_p \leq \varepsilon), \tag{2.3}$$

where $\|\cdot\|_0$ controls the number of non-zero elements in $\mathbf{A}$, and ε is used for managing the reconstruction errors. As shown in [26], the minimization problem as stated in Eq. (2.3) is an NP-hard problem under the l_0-norm with the l_1-norm to obtain a near-optimal solution [20]. The estimation is then accomplished by minimizing a least squares problem with a quadratic constraint, whose Lagrange multiplier formulation is:

$$< \Phi, \mathbf{A} >= \arg\min_{\Phi, \mathbf{A}} \|\mathbf{X} - \Phi\mathbf{A}\|_2^2 + \lambda \|\mathbf{A}\|_1, \tag{2.4}$$

where λ is a regularization factor trading off the parametric sparsity and the reconstruction error of the solution.

2.3 Method

The proposed SiSCS method computes an estimation of an HR version of a target MRI modality based on an LR version of a source MRI modality using jointly learned dictionary. SR reconstruction in this work is inspired in earlier work on brain hallucination [124], with an assumption that an HR image can be reconstructed from the LR input with the help by another

HR image using dictionaries of paired data in a sparse representation framework [76, 127]. In this work, we partition the images in the training database into a set of overlapping image patches. These image patches are built simultaneously on the source and target spaces by registered source-target image pairs. We propose a cross-modality dictionary learning that enforces the computation of joint sparse codes. Instead of working with the original data of the paired patches, we choose an HF representation of the data in the gradient domain, so the sparse codes promote a high correlation between the two modalities regarding the LR and HR, respectively. In brief, given the test image in matrix form $\mathbf{X}^t$ (with modality $\mathscr{M}_1$), the proposed method will synthesize an SR image $\mathbf{Y}^t$ with modality $\mathscr{M}_2$ from $\mathbf{X}^t$ through a patch-based global redundant reconstruction model regarding the learned cross-modality dictionary pair. The entire framework of SiSCS model is summarized in Fig. 2.2.

2.3.1 Data Description

Let $\mathscr{X} = \{\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_m\}$ be m training images of modality $\mathscr{M}_1$ in the source domain, and $\mathscr{Y} = \{\mathbf{Y}_1, \mathbf{Y}_2, ..., \mathbf{Y}_m\}$ be m training images of modality $\mathscr{M}_2$ in the target domain. We denote cross-modality image pairs as $\{\mathbf{X}_i, \mathbf{Y}_i\}$, while $\mathbf{X}_i$ and $\mathbf{Y}_i$ are registered. In this work, we consider the LR input and HR output and define the observed LR counterparts based on the HR images in $\mathscr{X}$ as Eq. (2.1). $\mathscr{X} = \{\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_m\}$ is then updated as $\mathscr{X}^L = \{\mathbf{X}_1^L, \mathbf{X}_2^L, ..., \mathbf{X}_m^L\}$, and cross-modality image pairs can be rewritten as $\{\mathbf{X}_i^L, \mathbf{Y}_i\}$. After that, we build our algorithm based on these data.

2.3.2 Gradient Feature Representation

A mapping is constructed between each pair of LR and HR images based on HF edges and texture features. This feature scheme is reasonable from the perceptual viewpoint since humans are more sensitive to changes in HF content [16]. It has been shown that HF feature representations of LR images are arguably the most important for forecasting the missing HF components of HR images [168]. Such kind of feature representation also makes the sparse codes of paired data that possess the same information close to each other [16, 41, 127]. To achieve this, researchers have suggested using a high-pass filter [41]. In this context, we define a feature operator $\mathscr{F}$ to perform feature extraction for the LR image set. For this purpose, we follow [16] and adopt the first-order and second-order gradients to represent features of each LR image $\mathbf{X}_i^L$. The derivatives are then defined as [16, 168]

$$\begin{aligned} f_1^1 &= [-1,0,1] \quad , \quad f_1^2 = [-1,0,1]^T \\ f_2^1 &= [-2,-1,0,1,2] \quad , \quad f_2^2 = [-2,-1,0,1,2]^T . \end{aligned}$$

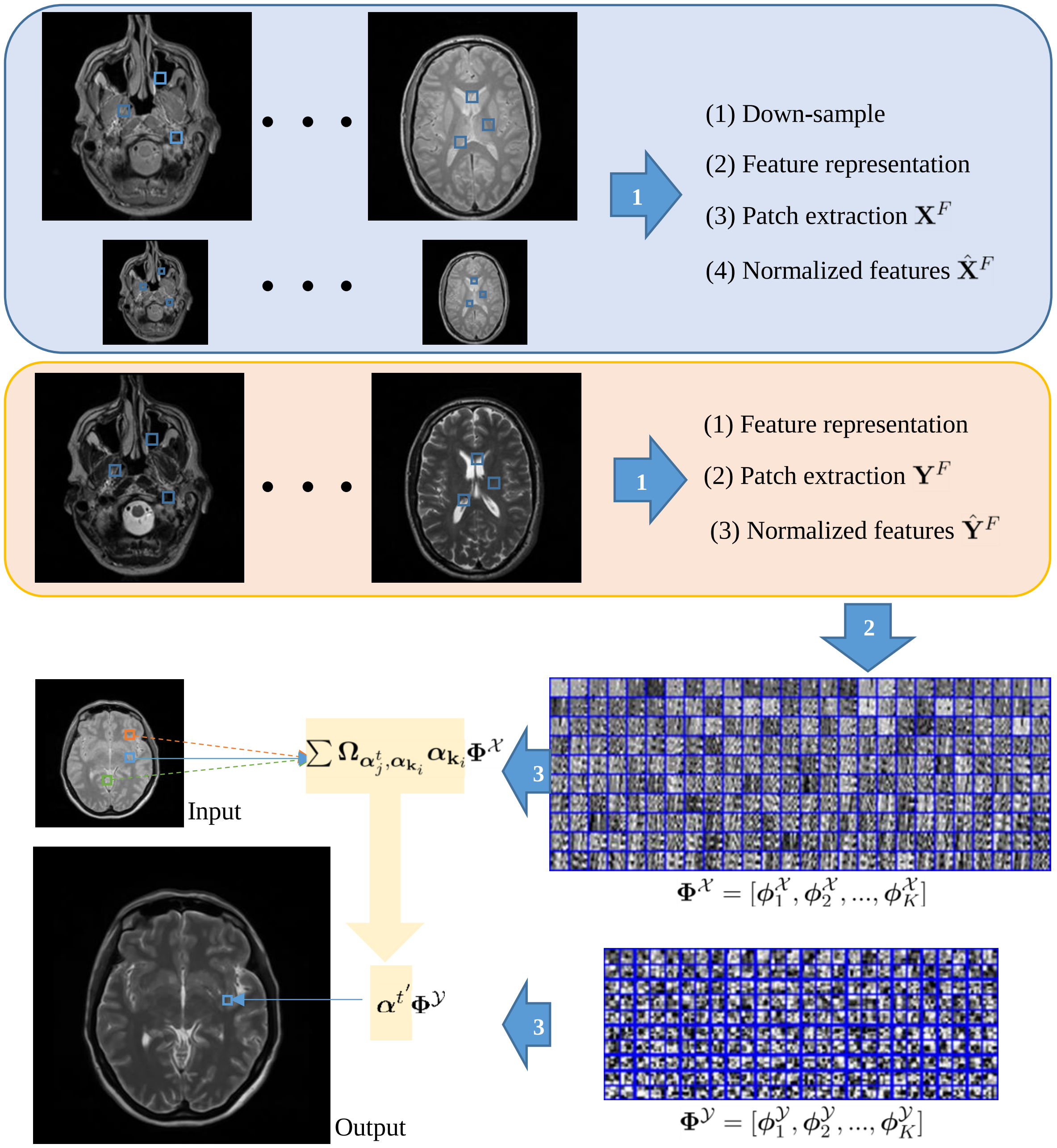

(1) Down-sample
(2) Feature representation
(3) Patch extraction $\mathbf{X}^F$
(4) Normalized features $\hat{\mathbf{X}}^F$
1
(1) Feature representation
(2) Patch extraction $\mathbf{Y}^F$
(3) Normalized features $\hat{\mathbf{Y}}^F$
1
2
$\sum \boldsymbol{\Omega}_{\alpha_j^t,\alpha_{\mathbf{k}_i}} \alpha_{\mathbf{k}_i} \boldsymbol{\Phi}^{\mathcal{X}}$
3
Input
$\boldsymbol{\Phi}^{\mathcal{X}} = [\phi_1^{\mathcal{X}}, \phi_2^{\mathcal{X}}, ..., \phi_K^{\mathcal{X}}]$
$\alpha^{t'} \boldsymbol{\Phi}^{\mathcal{Y}}$
3
Output
$\boldsymbol{\Phi}^{\mathcal{Y}} = [\phi_1^{\mathcal{Y}}, \phi_2^{\mathcal{Y}}, ..., \phi_K^{\mathcal{Y}}]$

Fig. 2.2 Illustration of the SiSCS model. Step 1: Feature collection. Step 2: Cross-modality dictionary learning. Step 3: Globally redundant synthesis.

where each LR image results in four filtered images involving horizontal and vertical gradients for both orders by $\mathscr{F} * \mathbf{X}^L$, with $*$ respecting the convolution operator and $\mathscr{F}$ takes the form of one the following operator: $f_1^1, f_1^2, f_2^1, f_2^2$. We then denote the features of LR images as $\mathbf{X}_i^F = \mathscr{F} * \mathbf{X}_i^L$. On the other hand, for the HR image set, we capture their HF features through directly removing the corresponding low-frequency information, which can be done by subtracting the mean value of HR data for $\mathbf{Y}_i$ [16], i.e. $\mathbf{Y}_i^F = \mathbf{Y}_i - mean(\mathbf{Y}_i)$. Further, images are treated as the collection of n patches and denoted as the matrices $\mathbf{X} = \left[\mathbf{x}_1^L, \mathbf{x}_2^L, ..., \mathbf{x}_n^L\right]$ and $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_n]$, and the corresponding HF features can be represented as $\mathbf{X}^F = \left[\mathbf{x}_1^F, \mathbf{x}_2^F, ..., \mathbf{x}_n^F\right]$ and $\mathbf{Y}^F = \left[\mathbf{y}_1^F, \mathbf{y}_2^F, ..., \mathbf{y}_n^F\right]$ in the source and target domains, respectively.

2.3.3 Cross-Modality Dictionary Learning

Following the dictionary learning procedure described in Section II, instead of considering the relationship between two sets of training data, we can learn two independent dictionaries [124] regarding the source domain and the target domain:

$$\begin{aligned}\Phi^{\mathscr{X}} &= \arg \min_{\Phi^{\mathscr{X}}, \mathbf{A}^{\mathscr{X}}} \left\| \mathbf{X}^F - \Phi^{\mathscr{X}} \mathbf{A}^{\mathscr{X}} \right\|_2^2 + \lambda \left\| \mathbf{A}^{\mathscr{X}} \right\|_1, \\ \Phi^{\mathscr{Y}} &= \arg \min_{\Phi^{\mathscr{Y}}, \mathbf{A}^{\mathscr{Y}}} \left\| \mathbf{Y}^F - \Phi^{\mathscr{Y}} \mathbf{A}^{\mathscr{Y}} \right\|_2^2 + \lambda \left\| \mathbf{A}^{\mathscr{Y}} \right\|_1.\end{aligned} \tag{2.5}$$

However, such a strategy is time-consuming and results in two sets of independent sparse codes. To solve a similar problem, Yang *et al.* [168] explored an image SR method that uses joint dictionary learning to correlate the sparse codes of LR data with those corresponding HR data. This is done by mapping LR and HR patch pairs into a common space to enforce the sparse codes of paired data possess the same value. Based on this method, we develop a cross-modality dictionary learning algorithm using the features of registered patch pairs to build the mapping relationship for highly dissimilar training data. To proceed with the synthesis, a dictionary pair $\Phi^{\mathscr{X}}$ and $\Phi^{\mathscr{Y}}$ should be simultaneously trained from data relating both modalities and resolutions. For doing this, we first capture the HF features of both sets and then project them into a common space to achieve an effective correlation. Once the LR and HR patch pairs are incorporated in the feature space, we proceed with the joint dictionary learning. However, such a strategy fails to consider different modalities involving inconsistent intensity scales in the feature space. To solve this problem, we introduce a normalization function so we can handle dissimilar features within the same range. The maximum l_2-norm values are then computed for both feature sets:

$$\sigma^{\mathscr{X}} = \max\left\{\|\mathbf{x}_i^F\|_2\right\}, \sigma^{\mathscr{Y}} = \max\left\{\|\mathbf{y}_i^F\|_2\right\}. \tag{2.6}$$

Once $\sigma^{\mathscr{X}}$ and $\sigma^{\mathscr{Y}}$ are obtained, we use them for intensity normalization of all patch features, i.e.

$$\hat{\mathbf{x}}_i^F = \frac{\mathbf{x}_i^F}{\sigma^{\mathscr{X}}}, \hat{\mathbf{y}}_i^F = \frac{\mathbf{y}_i^F}{\sigma^{\mathscr{Y}}}. \tag{2.7}$$

To maximize the correlation between normalized feature pairs in both modalities, we map them into a common high-dimensional space and propose a cross-modality dictionary learning method to simultaneously train two dictionaries for both datasets, leading to

$$\begin{aligned} &\arg\min_{\Phi^{\mathscr{X}},\Phi^{\mathscr{Y}},\mathbf{A}} \frac{1}{P}\left\|\hat{\mathbf{X}}^F - \Phi^{\mathscr{X}}\mathbf{A}\right\|_2^2 + \frac{1}{Q}\left\|\hat{\mathbf{Y}}^F - \Phi^{\mathscr{Y}}\mathbf{A}\right\|_2^2 \\ &+\lambda(\frac{1}{P}+\frac{1}{Q})\left\|\mathbf{A}\right\|_1 \;\; s.t. \; \left\|\Phi_i^{\mathscr{X}}\right\|_2^2 \leq 1, \left\|\Phi_i^{\mathscr{Y}}\right\|_2^2 \leq 1, \end{aligned} \tag{2.8}$$

where $\frac{1}{P}$ and $\frac{1}{Q}$ are the regularization parameters for balancing two error terms, and P and Q represent the sizes of LR and HR patches, respectively. The above formulation is convex regarding each dictionary (or sparse codes) assuming the other one fixed. Constructing $\Phi^{\mathscr{X}}$ and $\Phi^{\mathscr{Y}}$ is achieved by alternating the computation of the sparse codes and the update of the dictionary pairs. We summarize the training part of our SiSCS method in Algorithm 1.

Algorithm 1: SiSCS Training

```
Input: Training data $\mathbf{X}$ and $\mathbf{Y}$, parameters $\lambda$, $\mu$, $h$, $\gamma$.
Down sample and blur $\mathbf{X}$ by Eq. (2.1) to obtain $\mathbf{X}^L$.
Extract HF features and treat images as patches: $\mathbf{X}^F$, $\mathbf{Y}^F$.
Normalize patches by Eq. (2.7).
Initialize $\Phi_0^{\mathscr{X}}$, $\Phi_0^{\mathscr{Y}}$, $\mathbf{A}_0$.
while not converged do
    Update $\mathbf{A}_{i+1}$ by sparse coding in Eq. (2.8) with $\Phi_i^{\mathscr{X}}$ and $\Phi_i^{\mathscr{Y}}$ fixed.
    Update $\Phi_{i+1}^{\mathscr{X}}$ and $\Phi_{i+1}^{\mathscr{Y}}$ by dictionary learning in Eq. (2.8) with $\mathbf{A}_{i+1}$.
end
Output: $\Phi^{\mathscr{X}}$, $\Phi^{\mathscr{Y}}$.
```

2.3.4 Clustering-based Globally Redundant Codes

Once $\Phi^{\mathscr{X}}$ and $\Phi^{\mathscr{Y}}$ have been obtained from Eq. (2.8), we seek to reconstruct a test image $\mathbf{X}^t$ by sparsely representing the normalized features of $\mathbf{X}^t$ and $\Phi^{\mathscr{X}}$. This is done by solving Eq. (2.4) as

$$\arg\min_{\mathbf{A}^t} \left\|\hat{\mathbf{X}}^t - \Phi^{\mathscr{X}}\mathbf{A}^t\right\|_2^2 + \lambda \left\|\mathbf{A}^t\right\|_1, \tag{2.9}$$

where each patch of $\mathbf{X}^t$ is treated as its feature representation and normalized following Eq. (2.7) regarding LR and $\mathscr{M}_1$ modality by $\hat{\mathbf{X}}^t = \frac{\mathscr{F}*\mathbf{X}^t}{\sigma^{\mathscr{X}}}$, and $\mathbf{A}^t$ indicates the sparse coefficients of $\hat{\mathbf{X}}^t$. The estimated sparse codes can be directly used to synthesize the image $\mathbf{Y}^t$ of our desired modality $\mathscr{M}_2$ and HR by a linear combination of elements in the dictionary $\Phi^{\mathscr{Y}}$, namely $\mathbf{Y}^t = \Phi^{\mathscr{Y}}\mathbf{A}^t$.

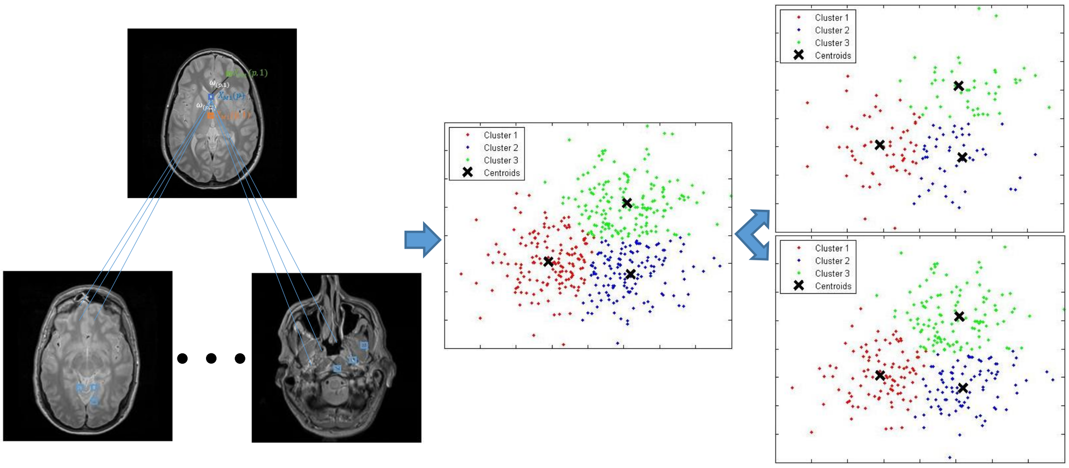


Fig. 2.3 Example of the clustering-based global redundancy model including cross-modal matching and self-similarity. For each reference patch in the test image, groups of similar patches including cross-modal matching and self-similarity can be found based on K-means clustering (we collect three clusters here).

Integrating nonlocal reconstruction was successfully explored in [12, 106, 29]. Nonlocal reconstruction method recognizes that images often display repetitive patterns across the image field, and that at each location the local texture resembles a weighted combination of the local textures at other locations [12]. We then assume there exist patches in $\mathbf{X}^t$ and $\mathbf{X}$ that resemble the j-th patch $\mathbf{X}_j^t$ of the test image. Groups of similar patches based on self- and cross-modal similarities (i.e., matching local self-similarity and global similarity across different modalities' data) are identified. Then non-local means (NL means) [12, 107] is applied to synthesize each target patch, which is reconstructed as a weighted average of all similar patches. Each neighboring patch is weighed inversely proportionally to its distance to the reference patch in the source image [107]. The patch similarity calculations and global search involved in establishing the set of similar patches is computationally intensive. To speed up computing the distance between the reference patch and each patch in the training database, we perform a two-stage search that eliminates grossly dissimilar patches first, and then refines with a local search. This multi-level search is more robust to noise and also addresses the problem of searches leading to very few retrievals due to less repetitive patterns. The first level search is carried out using K-means clustering using as input the sparse codes of the source patch and based on the Euclidean distance (cf. Fig. 2.3).

Let Ψ be the collection of the normalized HF features collected from $\hat{\mathbf{X}}^t$ and $\hat{\mathbf{X}}^F$. Then, we assume that they provide with s observations $\{\mathbf{k}_1, \mathbf{k}_2, ..., \mathbf{k}_s\}$ leading to s coefficients $\{\alpha_{\mathbf{k}_1}, \alpha_{\mathbf{k}_2}, ..., \alpha_{\mathbf{k}_s}\}$ from which we wish to generate K clusters $\{\psi_1, \psi_2, ..., \psi_K\}$. The clusters are computed as [53]:

$$\begin{aligned} &\arg\min_{\{\psi_c\}_{c=1}^{K}} \sum_{c=1}^{K} \sum_{\mathbf{k}_i \in \psi_c} \|\alpha_{\mathbf{k}_i} - \delta_c\|^2, \\ &\delta_c = \frac{1}{\psi_c} \sum_{\alpha_{\mathbf{k}_i} \in \psi_c} \alpha_{\mathbf{k}_i}. \end{aligned} \tag{2.10}$$

where δ_c is the mean vector for c-th cluster $\{\psi_c\}_{c=1}^{K}$. We pool the patches (other than the one to be synthesized) from the reference image with those in the training set as this has particular advantages when the reference image is abnormal or when the database is small. With the experiments reported with the IXI database, we found this is used in less than 2.76% of the subjects. In those cases, the non-local self-similarity has a stronger influence than the cross-subject self-similarity. This clustering yields a representative NL mean patch. We estimate the sparse codes for the j-th patch of the test image as the weighted average of the sparse codes $\alpha_{\mathbf{x}}$ associated with the corresponding cluster ψ_c via

$$\hat{\alpha}_j^t = \sum_{\alpha_{\mathbf{k}_i} \in \psi_c} \Omega_{\alpha_j^t, \alpha_{\mathbf{k}_i}} \alpha_{\mathbf{k}_i}, \tag{2.11}$$

where $\hat{\alpha}_j^t$ is the optimized sparse codes, $\alpha_{\mathbf{k}_i}$ denotes the sparse codes of $\mathbf{k}_i$ within the corresponding cluster, and $\Omega_{\alpha_j^t, \alpha_{\mathbf{k}_i}}$ is the weight for computing the level of similarity to be inversely proportional to the Euclidean distance between α_j^t and $\alpha_{\mathbf{k}_i}$, where

$$\Omega_{\alpha_j^t, \alpha_{\mathbf{k}_i}} = \frac{1}{\mu} \exp\left\{ -\frac{\left\|\alpha_j^t - \alpha_{\mathbf{k}_i}\right\|_2^2}{h^2} \right\}, \tag{2.12}$$

with $\Omega_{\alpha_j^t, \alpha_{\mathbf{k}_i}}$ satisfying $0 \leq \Omega_{\alpha_j^t, \alpha_{\mathbf{k}_i}} \leq 1$ and $\sum \Omega_{\alpha_j^t, \alpha_{\mathbf{k}_i}} = 1$, μ being a normalization constant, and h being a scalar. Similarly to the NL means method, the coefficient with higher similarity to α_j^t will have a larger weight in average (an example is shown in Fig. 2.3). Vectors within the cluster ψ_c contains not only test items but also training components. The conventional nonlocal method (e.g. NL means) cannot express the complex structures in MR images. In this scenario, our global redundancy approach can efficiently adapt to different structures. Therefore, the local sparse representation model in Eq. (2.9) that meets the complementary function to those of Eq. (2.11) will be modified as:

$$\arg\min_{\mathbf{A}^t} \left\|\hat{\mathbf{X}}^t - \Phi^{\mathscr{X}} \mathbf{A}^t\right\|_2^2 + \lambda \left\|\mathbf{A}^t\right\|_1 + \gamma \left\|\mathbf{A}^t - \hat{\mathbf{A}}^t\right\|_2. \tag{2.13}$$

Algorithm 2: SiSCS Synthesis

```
Input: Test image $\mathbf{X}^t$, dictionary pairs $\Phi^{\mathscr{X}}$, $\Phi^{\mathscr{Y}}$.
Extract features, divide patches and normalize: $\hat{\mathbf{X}}^t$.
Initialize $\mathbf{A}_0^t$.
while not converged do
    Update $\mathbf{A}_{i+1}^t$ by Eq. (2.9).
    Perform clustering by Eq. (2.10).
    Update $\mathbf{A}_{i+1}^t$ using Eq. (2.13).
end
Compute $\mathbf{Y}_{i+1}^t \leftarrow \Phi^{\mathscr{Y}} \mathbf{A}_{i+1}^t$.
Output: Synthesized image $\mathbf{Y}^t$.
```

where γ is a tunable regularization parameter. Finally, we update the synthesized image via $\mathbf{Y}^t = \Phi^{\mathscr{Y}} \mathbf{A}^t$. As with most of the super-resolution methods [151, 168], the contents of a LR test image besides the HF components are firstly preserved by extracting the features of each patch and then added the subtracted mean values back into the reconstructed SR version. Considering the domain-specific (i.e., source domain and target domain) information, we use the proposed globally redundant information to replace the original mean values from each patch of the test image. Then, we generate the target image by adding the pseudo mean values into the obtained HF components. The pseudo code for cross-modality synthesis is shown in Algorithm 2.

2.4 Experiments

To evaluate the performance of the proposed method, two scenarios were considered: (1) MR image super-resolution; (2) Simultaneous SR and cross-modality synthesis. We evaluated our model on two datasets: IXI[1] (containing 578 $256 \times 256 \times p$ p = 112∼136 MR healthy subjects) and NAMIC[2] (including 19 $128 \times 128 \times 88$ subjects, ten are normal controls and nine are schizophrenic). In our experiment, we applied leave-one-out cross-validation where removing the testing image from the entire dataset and learn on the remaining ones. For the experimental settings, we first sliced each 3D volume as the 2D stacks and then treated the 2D slices as many patches of 5×5 pixels size. We randomly sampled 100,000 patch pairs for training the cross-modality dictionary pair. The relevant parameters h, γ were set to be 16 and 0.1, respectively. We took the factor of dictionary size and sparsity regularization parameter into consideration and fixed the dictionary size to 1024 and λ =0.15 based on the quantitative analysis in Section 2.4.1 and 2.4.2. The parameter K of the K-means was fixed

[1] http://brain-development.org/ixi-dataset/

[2] http://hdl.handle.net/1926/1687

Table 2.1 Effects of Dictionary Size on SR Reconstruction and Simultaneous Super-Resolution and Cross-Modality Synthesis using All Slices of the IXI dataset.

	Dictionary Size	256	512	1024	2048
Cost	(min)	**8.51**	12.96	18.68	28.77
SR	PSNR(dB)	45.29	45.35	45.43	**45.62**
	SSIM	0.9865	0.9867	0.9867	**0.9872**
Synthesis	PSNR(dB)	39.54	39.55	39.57	**40.00**
	SSIM	0.8995	0.8995	0.8996	**0.8997**

to 10 to guarantee each cluster had enough candidates. Finally, we adopt the widely used Peak Signal to Noise Ratio (PSNR) in decibels (dB) and Structural Similarity Index (SSIM) [159] for illustrating the effectiveness of different methods. PSNR is employed to measure the quantitative evaluation of reconstruction of lossy compression codecs, which is defined as:

$$\mathrm{PSNR}(s,t) = 10\log_{10}\left(\frac{\mathrm{MAX}^2 mn}{\sum_i^m \sum_j^n [s(i,j)-t(i,j)]^2}\right), \tag{2.14}$$

where MAX denotes the maximum pixel value, m and n are the sizes of the synthesized image s and its ground truth t, and $s(i,j)$ and $t(i,j)$ represent the pixels of s and t at positions i and j, respectively. SSIM is a metric for measuring the perceived visual image quality [159]. SSIM is calculated:

$$\mathrm{SSIM}(s,t) = \frac{(2\mu_s\mu_t + c_1)(2\sigma_{st} + c_2)}{(\mu_s^2 + \mu_t^2 + c_1)(\sigma_s^2 + \sigma_t^2 + c_2)}, \tag{2.15}$$

where μ_s and μ_t are the mean values in s and t; σ_s and σ_t are their standard deviations; σ_{st} is the covariance of s and t; and $c_1 = (k_1L)^2$ and $c_2 = (k_2L)^2$ with L, the image intensity dynamic range, and $k_1 = 0.01$, $k_2 = 0.03$ [159].

2.4.1 Dictionary Size

Larger dictionaries can yield higher accuracy at the price of more calculations. In this experiment, we randomly picked 10 PD-w subjects to test the influence of four dictionary sizes (*viz.* 256, 512, 1024, and 2048) on both SR and simultaneous SR and cross-modality synthesis (PD-w→T2-w). Table 2.1 specifies relevant times for training dictionaries of different sizes, and the averaged PSNRs and SSIMs for image reconstructions using different dictionaries. From Table 2.1, we can see that a larger dictionary contributes a better estimation with larger PSNRs and SSIMs at a higher computation cost. We selected the size 1024 to yield a good image quality within acceptable computational time.

Table 2.2 Performance Measures of SR Resolution and Simultaneous Super-Resolution and Cross-Modality Synthesis for Different Sparsity Values using All slices of the IXI dataset.

	λ	0.10	0.15	0.25	0.35	0.45	0.55	0.65	0.75	0.85
Cost	(min)	12.55	11.06	10.97	10.34	9.17	8.13	8.10	**8.02**	8.25
SR	PSNR(dB)	47.41	**49.85**	49.80	46.47	40.82	36.93	36.92	36.90	36.90
	SSIM	0.9935	**0.9962**	0.9960	0.9932	0.9831	0.9429	0.9429	0.9428	0.9428
SiSCS	PSNR(dB)	39.36	**39.39**	39.32	37.62	35.46	34.95	34.95	34.94	34.93
	SSIM	0.9066	**0.9077**	0.9076	0.9055	0.8667	0.8500	0.8500	0.8499	0.8498

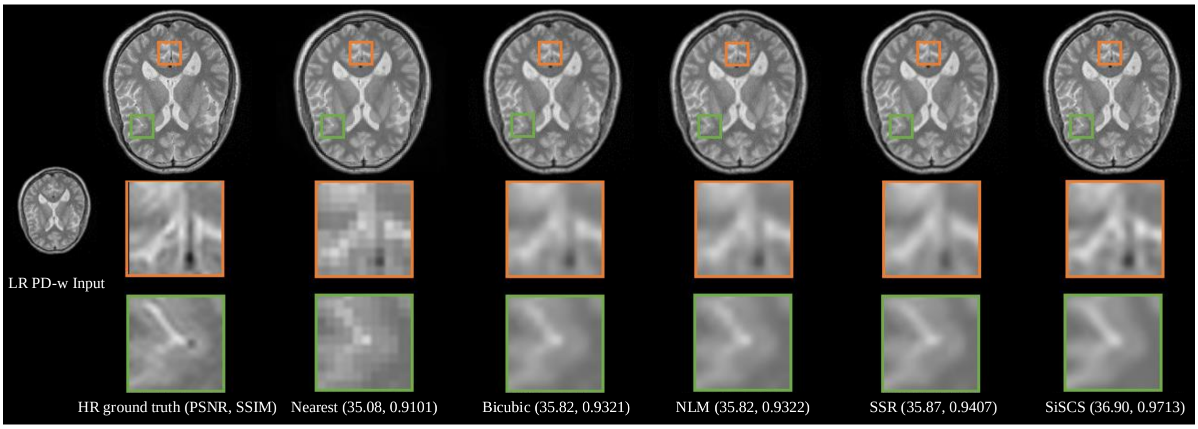


Fig. 2.4 Comparison of the SR results with ground truth.

2.4.2 Sparsity

In Eq. (2.4), λ plays an important role in the sparse representation as it is used for controlling the sparsity of the results. Empirically, λ is suitable from 0 to 0.85 [140, 127] for maintaining the model stability. In this section, we assess how λ influences the results through quantifiably measuring the PSNRs and SSIMs of the reconstructed image for different λs. To evaluate this, we utilized the same test data reported in the Section 2.4.1 and fixed the dictionary size to 1024. The experimental results are listed in Table 2.2. As shown, $\lambda \in [0.15, 0.45]$ yielded better performance, especially when $\lambda = 0.15$, the results on both scenarios achieve the highest PSNRs and SSIMs among all reconstructions. To comprehensively analyze the most suitable sparsity value for our algorithm, we computed the elapsed time for $\lambda \in [0.10, 0.85]$ and show the results in Table 2.2. As λ increased, the computational cost decreased, and the quality of reconstruction declined. Therefore, we chose no larger value; rather, a smaller λ was selected for achieving better results. We finally chose a sparsity parameter of 0.15.

2.4.3 MRI Super-Resolution

First, we evaluated the performance of our clustering-based global redundancy framework for MR image SR on all PD-w subjects of IXI dataset. Generally, LR images can be generated by first blurring the HR images with a 2D Gaussian kernel with standard deviation 1.6 and then down-sampling the blurred images by scaling factor 2 in both horizontal and vertical directions. To ensure the one-to-one correspondence for each extracted LR-HR patch pair, we further up-sampled the LR images by factor of 2 using bi-cubic interpolation (making the SR problem more challenging) and then extracted patches from them. This ensures that samplings from the same locations of both domains indicate the same tissue information. Based on the extracted patch pairs, we can train the corresponding dictionaries. Finally, we inputted an LR counterpart of the test image for reconstructing its HR image via the proposed model with a scaling factor of 2. To show the performance of our approach, we compared our results with these methods: nearest neighbor interpolation (Nearest), bi-cubic interpolation (Bi-cubic), non-local MRI up-sampling (NLM) [107], and Single-image SR (SSR) of brain MRI [128]. For all experiments, we tuned parameters in the characteristics of each method and demonstrated their best results among overall records by PSNRs and SSIMs.

Fig. 2.4 presents a comparison between the SR reconstructed PD-w slices based on different methods. Top row shows the axial views of the SR results for a subject. In the second and third rows we zoom in two specific regions for better visualization. The last row provides PSNRs and SSIMs for each listed method. The proposed method outperforms all other methods displaying the highest PSNR and SSIM. Although NLM yields a cleaner image with noise lower than bi-cubic interpolation, its effectiveness is nearly the same as bi-cubic. Fig. 2.8 provides quantitative results on all PD-w subjects of IXI dataset. Our method achieved the highest PSNR and SSIM compared to other methods.

2.4.4 Simultaneous Super-Resolution & Cross-Modality Synthesis

We then addressed the problem of simultaneous super-resolution and cross-modality synthesis by evaluating our SiSCS model on both IXI and NAMIC datasets. First, we used PD-w and T2-w subjects from IXI for synthesizing HR T2-w image considering LR PD-w acquisition from the same subject and *vice versa*. Second, generating HR T2-w image from LR PD-w input based on the pre-processed data (i.e., performing skull strapping and bias corrections[3]) and *vice versa*. Third, we considered the generation of T1-w image based on T2-w input and *vice versa*. We conducted the first two sets of experiments on the IXI dataset, while the third

[3] Following [153, 172], all the experiments data were skull stripped, linear registered and/or inhomogeneity corrected.

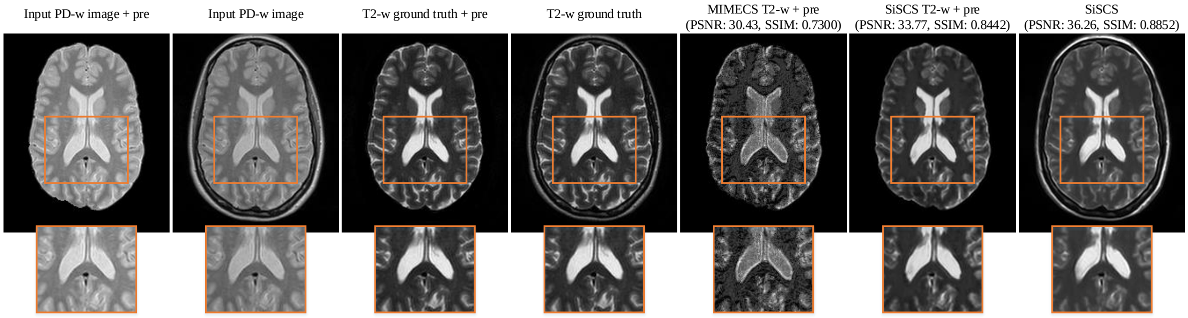


Fig. 2.5 Axial views of synthesized HR T2-w examples based on the LR PD-w inputs using different methods (zoom in for details).

one was explored on the NAMIC dataset. The representative and state-of-the-art synthesis methods including MR image exampled-based contrast synthesis (MIMECS) [127] approach, Vemulapalli's supervised method (V-s) [153] and Vemulapalli's unsupervised method (V-us) [153] were employed to compare with our SiSCS. However, Vemulapalli's methods is limited by the specific cross-modality synthesis used in the NAMIC dataset. Therefore, original data (without degradation processing) were used in all Vemulapalli's methods. All existing synthesis approaches must pre-process the data first. In our algorithm, such pre-processing is unnecessary and can be exchangeable which can be validated in the first set of experiments. The benefits of performing pre-processing are reflected in the reduction of the interference by non-desired tissue compartments, as the skull. However, such processes also bring problems, for instance, the lack of tissue contrast due to the imprecise skull stripping.

For the first two sets of experiments, we evaluated our algorithm and compared with MIMECS while displaying results in Fig. 2.5 for visual inspection. For each scenario, we applied the proposed method on IXI from the PD-w and T2-w subjects. Our algorithm performs consistently across the whole dataset, reaching the best performance for almost all subjects. We evaluated SiSCS and relevant methods in the third scenario, allowing us to comprehensively compare the performance of the proposed method in both healthy and pathological cases with the recently published algorithms. The advantage of SiSCS over other methods was shown in Fig. 2.6 and the close-up views of the selected parts are also provided for better visualization. The overall performance comparison is given in Fig. 2.9. From Fig. 2.9, we can see that SiSCS is always better than MIMECS and Vemulapalli's approaches. This demonstrates the effectiveness of our simultaneous SR and cross-modality technique.

The following experiments show synthesized images of schizophrenic patients. We carry out simultaneous synthesis and SR in two different learning scenarios: a) dictionary learning based on healthy subjects (denoted by SiSCS-H), and b) dictionary learning based on both healthy and schizophrenic (denoted by SiSCS) cases. In both scenarios we report synthesis

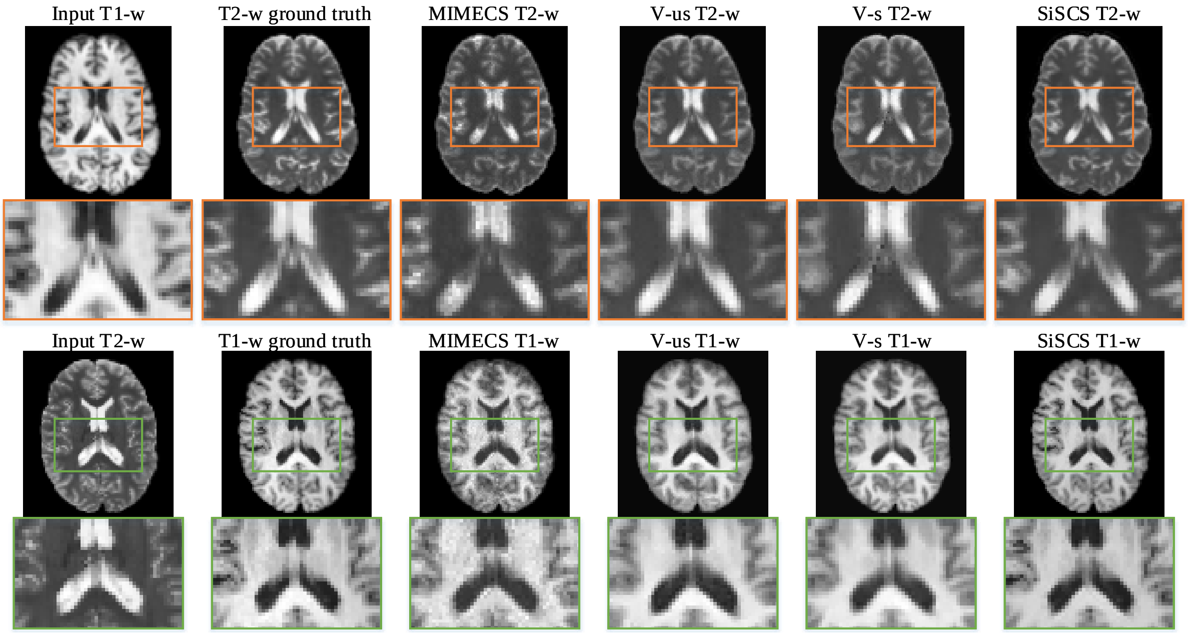


Fig. 2.6 Visual comparison of synthesized results using different methods on the NAMIC dataset (zoom in for details).

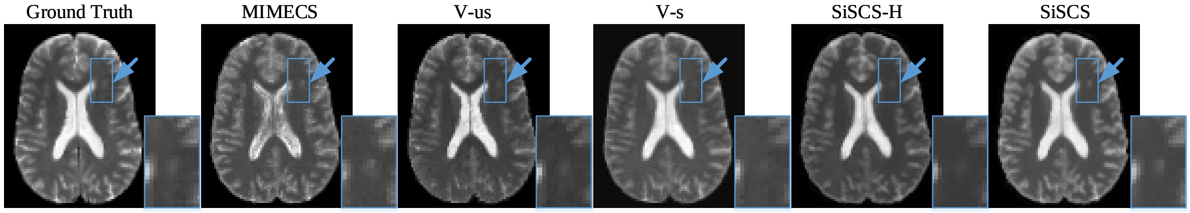


Fig. 2.7 Synthesis result of a pathological case comparison between SiSCS and other stat-of-the-art methods.

results corresponding to schizophrenia cases only. Fig. 2.7 provides visual results of various synthesis methods of an illustrative patient. Table 2.3 summarizes key quantitative performance metrics over the total set of 9 schizophrenic subjects. Both visual and quantitative results show that, compared to earlier methods, our approach provides the best results. Our experiments also show that SiSCS-H is outperformed by SiSCS trained on both control and pathologic cases and tested using cross-validation.

2.5 Conclusions

We present a novel approach to simultaneous super-resolution and cross-modality synthesis (SiSCS) in brain MRI. SiSCS first learns a cross-modality dictionary in a high-frequency space.

Table 2.3 Average Assessment Measures for Image Synthesis of Nine Pathological Cases

Case		MIMECS	V-us	V-s	SiSCS-H	SiSCS
T1->T2	PSNR(dB)	22.95	23.87	26.69	24.86	**27.03**
	SSIM	0.8698	0.8701	0.8895	0.8712	**0.8912**
T2->T1	PSNR(dB)	27.38	27.47	29.12	27.91	**30.01**
	SSIM	0.9000	0.9002	0.9087	0.9003	**0.9177**

Then, SiSCS reconstructs the target image using a patch-based cross-modal estimation model with a nonlocal sparse image representation. We conducted extensive experimental assessment of our technique in both health and schizophrenic subjects. Across experiments, both on PSNR and SSIM metrics, SiSCS outperformed three major contending techniques. Experiments showed consistent out performance across super-resolution and joint super-resolution and cross-modality synthesis, respectively. In our experiments we showed that dictionary learning for synthesis of schizophrenic images requires that pathological sets are included.

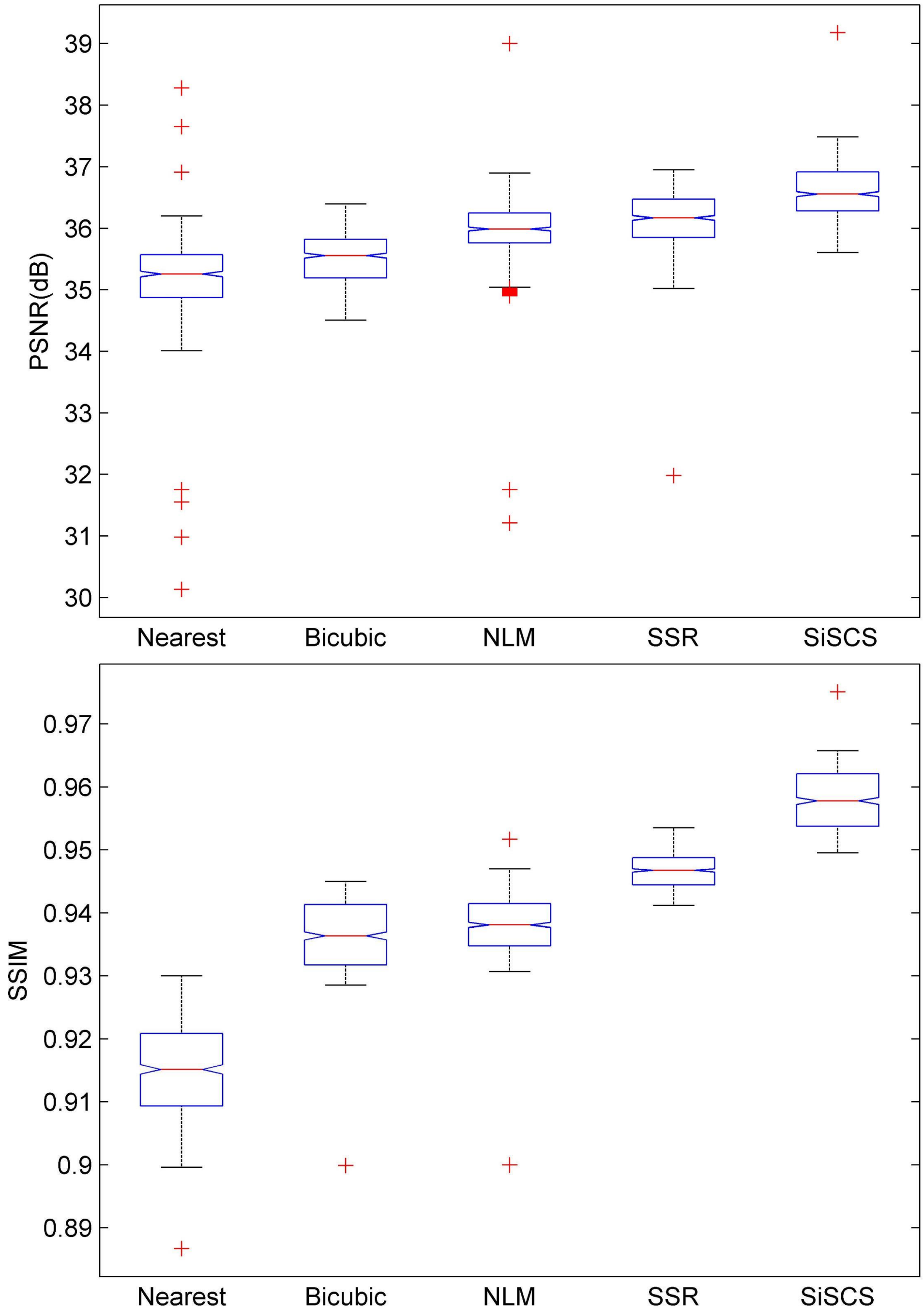


Fig. 2.8 Boxplots of error measures (PSNRs, SSIMs) for SR reconstructions with different methods.

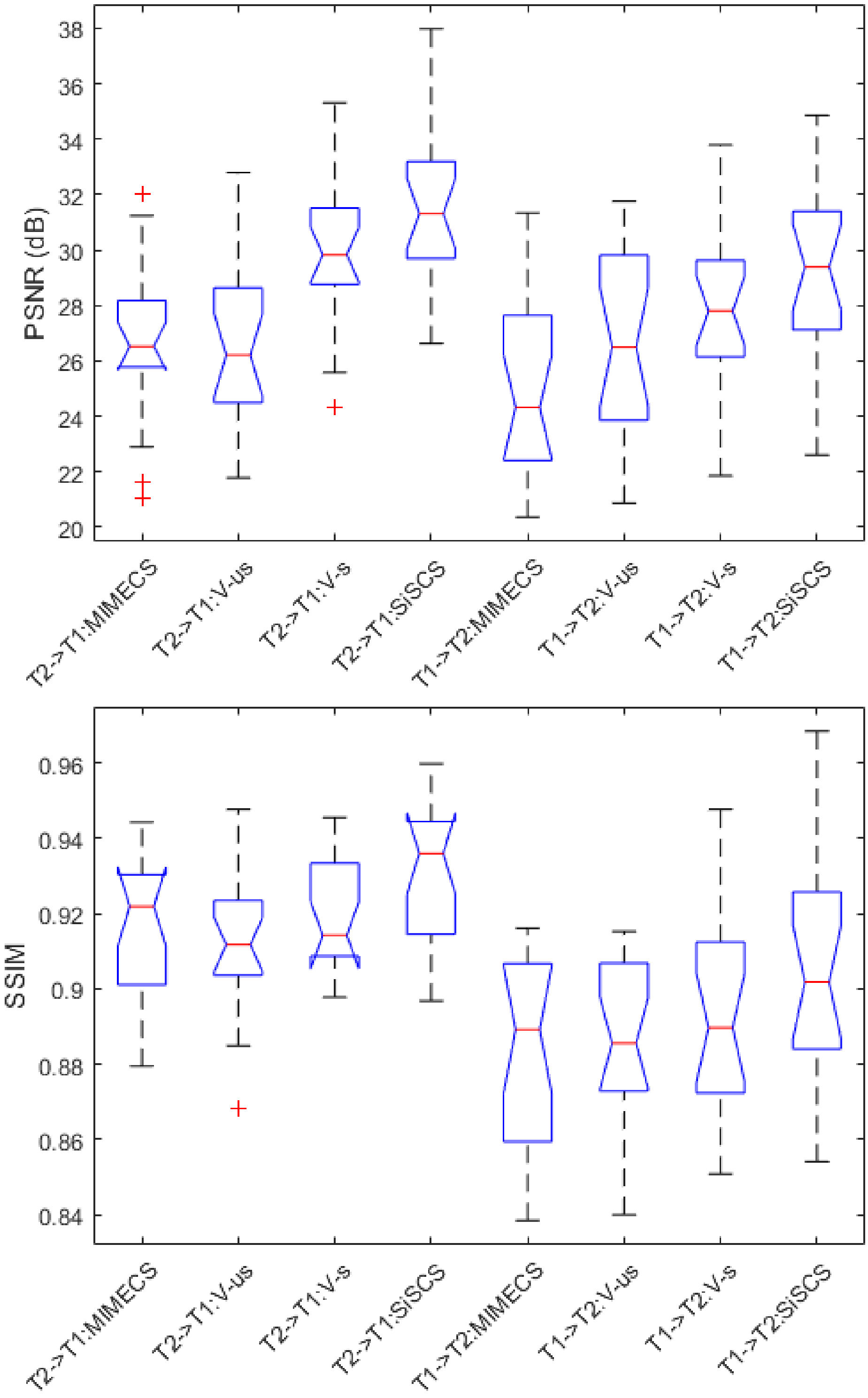


Fig. 2.9 Synthesis performance comparison between SiSCS and other stat-of-the-art methods on the NAMIC dataset.

Chapter 3

Dual Convolutional Filter Learning

Cross-modal image synthesis is a topical problem in medical image computing. Existing methods for image synthesis are either tailored to a specific application, require large scale training sets, or are based on partitioning images into overlapping patches. In this chapter, we propose a novel Dual cOnvolutional filTer lEarning (DOTE) approach to overcome the drawbacks of these approaches. We construct a closed loop joint filter learning strategy that generates informative feedback for model self-optimization. Our method can leverage data more efficiently thus reducing the size of the required training set. We extensively evaluate DOTE in two challenging tasks: image super-resolution and cross-modality synthesis. The experimental results demonstrate superior performance of our method over other state-of-the-art methods.

3.1 Introduction

In medical image analysis, it is sometimes convenient or necessary to infer an image from one modality or resolution from another image modality or resolution for better disease visualization, prediction and detection purposes. A major challenge of cross-modality image segmentation or registration comes from the differences in tissue appearance or spatial resolution in images arising from different physical acquisition principles or parameters, which translates into the difficulty to represent and relate these images. Some existing methods tackle this problem by learning from a large amount of registered images and constraining pairwise solutions in a common space. In general, one would desire to have high-resolution (HR) three-dimensional Magnetic Resonance Imaging (MRI) with near isotropic voxel resolution as opposed to the more common image stacks of multiple 2D slices for accurate quantitative image analysis and diagnosis. Multi-modality imaging can generate tissue contrast arising from various anatomical or functional features that present complementary information about the underlying organ. Acquiring low-resolution (LR) single-modality images, however, is not uncommon.

To solve the above problems, super-resolution (SR) [148, 168] reconstruction is carried out for recovering an HR image from its LR counterpart, and cross-modality synthesis (CMS) [153] is proposed for synthesizing target modality data from available source modality images. Generally, these methods have explored image priors from either internal similarities of image itself [125] or external data support [176], to construct the relationship between two modalities. Although these methods achieve remarkable results, most of them suffer from the fundamental limitations associated with large scale pairwise training sets or patch-based overlapping mechanism. Specifically, a large amount of multi-modal images is often required to learn a sufficiently expressive dictionaries/networks. However, this is impractical since collecting medical images is very costly and limited by many factors. On the other side, patch-based methods are subjected to inconsistencies introduced during the fusion process that takes place in areas where patches overlap.

To deal with the bottlenecks of training data and patch-based implementation, we develop a dual convolutional filter learning (DOTE) method with an application to neuroimaging that investigates data (in both source and target modalities from the same set of subjects) in a more effective way, and solves image SR and CMS problems respectively. The contributions of this work are mainly in four aspects:

1. We present a unified model (DOTE) for any cross-modality image synthesis problem.
2. The proposed method can efficiently reduce the amount of training data needed from the model, by generating abundant feedbacks from dual mapping functions during the training process.
3. Our method integrates feature learning and mapping relation in a closed loop for self-optimization. Local neighbors are preserved intrinsically by directly working on the whole images.
4. We evaluate DOTE on two datasets in comparison with stat-of-the-art methods. Experimental results demonstrate superior performance of DOTE over these approaches.

3.2 Background

3.2.1 Convolutional Sparse Coding

Convolutional Sparse Coding (CSC) remedies a fundamental drawback of conventional patch-based sparse representation methods by modeling shift invariance for consistent approximation of local neighbors on whole images. Instead of decomposing the vector as the multiplication of dictionary atoms and the coded coefficients, CSC provides a more elegant way to model local interactions. That is, by representing an image as the summation of convolutions of the sparsely distributed feature maps and the corresponding filters. Concretely, given an $m \times n$ image $\mathbf{x}$ in vector form, the problem of learning a set of vectorized filters for sparse feature maps can be solved by minimizing the objective function that combines the convolutional least-squares term and the l_1-norm penalty on the representations:

$$\begin{gathered} \arg\min_{\mathbf{f},\mathbf{s}} \frac{1}{2} \left\| \mathbf{x} - \sum_{k=1}^{K} \mathbf{f}_k * \mathbf{s}_k \right\|_2^2 + \lambda \sum_{k=1}^{K} \|\mathbf{s}_k\|_1 \\ s.t.\ \|\mathbf{f}_k\|_2^2 \leq 1 \ \forall k = \{1, \ldots, K\}, \end{gathered} \tag{3.1}$$

where $\mathbf{f}_k \in \mathbf{F} = \left[\mathbf{f}_1^T, \ldots, \mathbf{f}_K^T\right]^T$ is the k-th $d \times d$ filter, $*$ denotes the 2D convolution operator, $\mathbf{s}_k \in \mathbf{S} = \left[\mathbf{s}_1^T, \ldots, \mathbf{s}_K^T\right]^T$ refers to the sparse feature map corresponding to $\mathbf{f}_k$ with size $(m+d-1) \times (n+d-1)$ to approximate $\mathbf{x}$, and λ is a regularization parameter. The problem in Eq. (3.1) can be efficiently and explicitly solved in the Fourier domain, derived within an Alternating Direction Method of Multipliers (ADMM) framework [10].

3.2.2 Dual Learning

Dual Learning (DL) [55] is a new learning paradigm that translates the input model by forming a closed loop between source and target domains to generate informative feedbacks. Specifically, for any dual tasks (e.g., $A \leftrightarrow B$) DL strategy appoints $A \rightarrow B$ as the primary task and the other $A \leftarrow B$ as the dual task, and forces them learning from each other to produce the pseudo-input A'. It can achieve the comparable performance through iteratively updating and minimizing the reconstruction error $A - A'$ that helps maximize the use of data. Therefore, making the learning-based methods less dependent on the large numbers of training data.

3.2.3 Problem Formulation

The cross-modality image synthesis problem can be formulated as: given an 3D image $\mathbf{X}$ of modality $\mathscr{M}_1$, the task is to infer from $\mathbf{X}$ a target 3D image $\mathbf{Y}$ that approximates to the ground truth of modality $\mathscr{M}_2$. Let $\mathscr{X} = [\mathbf{X}_1, \ldots, \mathbf{X}_C] \in \mathbb{R}^{m \times n \times z \times C}$ be a set of images of modality $\mathscr{M}_1$ in the source domain, and $\mathscr{Y} = [\mathbf{Y}_2, \ldots, \mathbf{Y}_C] \in \mathbb{R}^{m \times n \times z \times C}$ be a set of images of modality $\mathscr{M}_2$ in the target domain. m, n are the dimensions of axial view of the image, and z denotes the size of image along the z-axis, while C is the number of elements in the training sets. Each pair of $\{\mathbf{X}_i, \mathbf{Y}_i\}$ $\forall i = \{1, \ldots, C\}$ are registered. To bridge image appearances across different

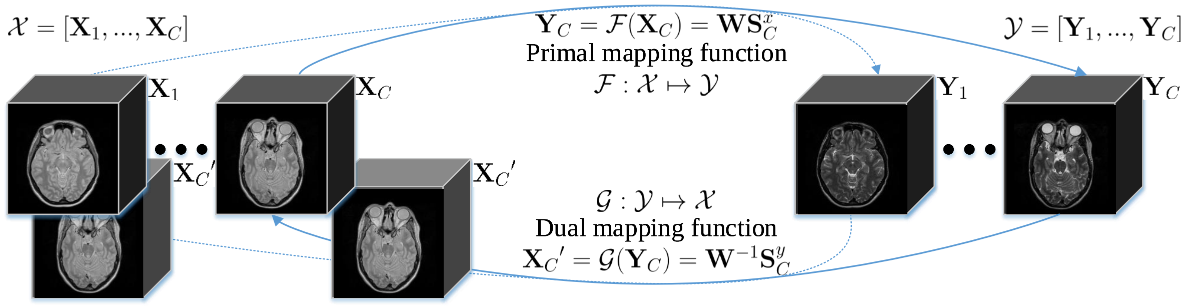

$\mathcal{X} = [\mathbf{X}_1, ..., \mathbf{X}_C]$
$\mathbf{X}_1$
$\mathbf{X}_C$
$\mathbf{X}_C{}'$
$\mathbf{X}_C{}'$
$\mathbf{Y}_C = \mathcal{F}(\mathbf{X}_C) = \mathbf{W}\mathbf{S}^x_C$
Primal mapping function
$\mathcal{F} : \mathcal{X} \mapsto \mathcal{Y}$
$\mathcal{Y} = [\mathbf{Y}_1, ..., \mathbf{Y}_C]$
$\mathbf{Y}_1$
$\mathbf{Y}_C$
$\mathcal{G} : \mathcal{Y} \mapsto \mathcal{X}$
Dual mapping function
$\mathbf{X}_C{}' = \mathcal{G}(\mathbf{Y}_C) = \mathbf{W}^{-1}\mathbf{S}^y_C$

Fig. 3.1 Flowchart of the proposed method for MRI cross-modality synthesis.

modalities while preserving the intrinsic local interactions (i.e., intra-domain consistency), we propose a method based on CSC to jointly learn a pair of filters $\mathbf{F}^x$ and $\mathbf{F}^Y$. Moreover, inspired by the DL strategy, we form a closed loop between both domains and assume that there exists a primal mapping function $\mathscr{F}(\cdot)$ from $\mathscr{X}$ to $\mathscr{Y}$ for relating and predicting from one another. We also assume there exists a dual mapping function $\mathscr{G}(\cdot)$ from $\mathscr{Y}$ to $\mathscr{X}$ to generate feedbacks for model self-optimization. Experimentally, we investigate human brain MRI and apply our method to two cross-modality synthesis tasks, i.e., image SR and CMS. An overview of our method is depicted in Fig. 3.1.

Notation

Matrices and 3D images are written in bold uppercase (e.g., image $\mathbf{X}$), vectors and vectorized 2D images in bold lowercase (e.g., filter $\mathbf{f}$) and scalars in lowercase (e.g., element k).

3.3 Method

3.3.1 Dual Convolutional Filter Learning

Inspired by CSC (cf. Sec. 3.2) and the benefits of conventional coupled sparsity, we propose a dual convolutional filter learning (DOTE) model, which extends the original CSC formulation into a DL strategy and joint representation into a unified framework. More specifically, given $\mathscr{X}$ together with the corresponding $\mathscr{Y}$ for training, in order to facilitate a joint mapping, we associate the sparse feature maps of each registered data pair $\{\mathbf{X}_i, \mathbf{Y}_i\}_{i=1}^{C}$ by constructing a forward mapping function $\mathscr{F} : \mathscr{X} \mapsto \mathscr{Y}$ with $\mathbf{Y} = \mathscr{F}(\mathbf{X})$. Since such cross-modality synthesis problem satisfies a dual-learning mechanism, we further leverage the duality of the bidirectional transformation between the two domains. That is, by establishing a dual mapping function $\mathscr{G} : \mathscr{Y} \mapsto \mathscr{X}$ with $\mathbf{X} = \mathscr{G}(\mathbf{Y})$. Incorporating feature maps representing and the above closed-loop mapping functions, we can thus derive the following objective function:

$$\begin{aligned} \arg\min_{\mathbf{F}^x,\mathbf{F}^y,\mathbf{S}^x,\mathbf{S}^y,\mathbf{W}} & \frac{1}{2}\left\|\mathbf{X}-\sum_{k=1}^{K}\mathbf{F}_k^x * \mathbf{S}_k^x\right\|_F^2 + \frac{1}{2}\left\|\mathbf{Y}-\sum_{k=1}^{K}\mathbf{F}_k^y * \mathbf{S}_k^y\right\|_F^2 + \gamma\sum_{k=1}^{K}\|\mathbf{W}_k\|_F^2 \\ +\lambda\left(\sum_{k=1}^{K}\|\mathbf{S}_k^x\|_1 + \sum_{k=1}^{K}\|\mathbf{S}_k^y\|_1\right) & + \beta\left(\sum_{k=1}^{K}\|\mathbf{S}_k^y - \mathbf{W}_k\mathbf{S}_k^x\|_F^2 + \sum_{k=1}^{K}\|\mathbf{S}_k^x - \mathbf{W}_k^{-1}\mathbf{S}_k^y\|_F^2\right) \\ & s.t. \ \|\mathbf{f}_k^x\|_2^2 \le 1 \ \|\mathbf{f}_k^y\|_2^2 \le 1 \ \forall k = \{1,\ldots,K\}. \end{aligned} \tag{3.2}$$

where $\mathbf{S}_k^x$ and $\mathbf{S}_k^y$ take the role of the k-th sparse feature maps that approximate data $\mathbf{X}$ and $\mathbf{Y}$ when convolved with the k-th filters $\mathbf{F}_k^x$ and $\mathbf{F}_k^y$ of a fixed spatial support, $k = 1,\ldots,K$. $\|\cdot\|_F$ is a Frobenius norm chosen to induce the convolutional least squares approximation, and $*$ is represented as a 3D convolution operator, while λ, β, γ are the regularization parameters. Particularly, dual mapping functions $\mathscr{F}\left(\mathbf{S}_k^x,\mathbf{W}_k\right) = \mathbf{W}_k\mathbf{S}_k^x$ and $\mathscr{G}\left(\mathbf{S}_k^y,\mathbf{W}_k^{-1}\right) = \mathbf{W}_k^{-1}\mathbf{S}_k^y$ are used to relate the sparse feature maps of $\mathbf{X}$ and $\mathbf{Y}$ over $\mathbf{F}^x$ and $\mathbf{F}^y$. They are done by solving two sets of least squares terms (i.e., $\sum_{k=1}^{K}(\left\|\mathbf{S}_k^y - \mathbf{W}_k\mathbf{S}_k^x\right\|_F^2 + \sum_{k=1}^{K}\left\|\mathbf{S}_k^x - \mathbf{W}_k^{-1}\mathbf{S}_k^y\right\|_F^2)$ with respect to the linear projections.

3.3.2 Optimization

Similar to classical dictionary learning methods, the objective function in Eq. (3.2) is not simultaneously convex with respect to the learned filter pairs,the sparse feature maps and the mapping. Instead, we divide the proposed method into three sub-problems: learning $\mathbf{S}^x$, $\mathbf{S}^y$, training $\mathbf{F}^x$, $\mathbf{F}^y$, and updating $\mathbf{W}$.

Computing sparse feature maps

We first initialize the filters $\mathbf{F}^x$, $\mathbf{F}^y$ as two random matrices and the mapping $\mathbf{W}$ as an identity matrix, then fix them for calculating the solutions of sparse feature maps $\mathbf{S}^x$, $\mathbf{S}^y$. As a result, the problem of Eq. (3.2) can be converted into two optimization sub-problems. Unfortunately, this cannot be solved under l_1 penalty without breaking rotation invariance. The resulting alternating algorithms [10] by introducing two auxiliary variables $\mathbf{U}$ and $\mathbf{V}$ enforce the constraint inherent in the splitting. In this work, we follow [10] and solve the convolution subproblems in the

Fourier domain within an ADMM optimization strategy:

$$
\begin{aligned}
&\min_{\mathbf{S}^x} \frac{1}{2}\left\|\hat{\mathbf{X}} - \sum_{k=1}^{K} \hat{\mathbf{F}}_k^x \odot \hat{\mathbf{S}}_k^x\right\|_F^2 + \lambda \sum_{k=1}^{K} \|\mathbf{U}_k^x\|_1 + \beta \sum_{k=1}^{K} \|\hat{\mathbf{S}}_k^y - \mathbf{W}_k \hat{\mathbf{S}}_k^x\|_F^2 \\
&\qquad s.t.\ \|\mathbf{V}_k^x\|_2^2 \leq 1,\ \mathbf{V}_k^x = \mathbf{V}\Phi^T \hat{\mathbf{F}}_k^x,\ \mathbf{U}_k^x = \mathbf{S}_k^x\ \forall k = \{1,\ldots,K\}, \\
&\min_{\mathbf{S}^y} \frac{1}{2}\left\|\hat{\mathbf{Y}} - \sum_{k=1}^{K} \hat{\mathbf{F}}_k^y \odot \hat{\mathbf{S}}_k^y\right\|_F^2 + \lambda \sum_{k=1}^{K} \left\|\mathbf{U}_k^y\right\|_1 + \beta \sum_{k=1}^{K} \|\hat{\mathbf{S}}_k^x - \mathbf{W}_k^{-1} \hat{\mathbf{S}}_k^y\|_F^2 \\
&\qquad s.t.\ \left\|\mathbf{V}_k^y\right\|_2^2 \leq 1,\ \mathbf{V}_k^y = \mathbf{V}\Phi^T \hat{\mathbf{F}}_k^y,\ \mathbf{U}_k^y = \mathbf{S}_k^y\ \forall k = \{1,\ldots,K\},
\end{aligned}
\tag{3.3}
$$

where ˆ applied to any symbol denotes the frequency representations (i.e., Discrete Fourier Transform (DFT)). For instance, $\hat{\mathbf{X}} \leftarrow f(\mathbf{X})$ where $f(\cdot)$ is the Fourier transform operator. $\odot$ represents the component-wise product. Φ^T is the inverse DFT matrix, and $\mathbf{V}$ projects a filter onto the small spatial support. The auxiliary variables $\mathbf{U}_k^x$, $\mathbf{U}_k^y$, $\mathbf{V}_k^x$ and $\mathbf{V}_k^y$ relax each of the CSC problems under dual mapping constraint by leading to several subproblem decompositions.

Learning convolutional filters

Like when solving for sparse feature maps, filter pairs can be learned similarly by setting $\mathbf{S}^x$, $\mathbf{S}^y$ and $\mathbf{W}$ fixed, and then learning $\mathbf{F}^x$ and $\mathbf{F}^y$ by minimizing

$$
\begin{aligned}
&\min_{\mathbf{F}^x,\mathbf{F}^y} \frac{1}{2}\left\|\hat{\mathbf{X}} - \sum_{k=1}^{K} \hat{\mathbf{F}}_k^x \odot \hat{\mathbf{S}}_k^x\right\|_F^2 + \frac{1}{2}\left\|\hat{\mathbf{Y}} - \sum_{k=1}^{K} \hat{\mathbf{F}}_k^y \odot \hat{\mathbf{S}}_k^y\right\|_F^2 \\
&\qquad s.t.\ \|\mathbf{f}_k^x\|_2^2 \leq 1, \left\|\mathbf{f}_k^y\right\|_2^2 \leq 1\ \forall k = \{1,\ldots,K\},
\end{aligned}
\tag{3.4}
$$

Eq. (3.4) can be solved by a one-by-one update strategy [155] through an augmented Lagrangian method [10].

Updating mapping

With fixed $\mathbf{F}^x$, $\mathbf{F}^y$, $\mathbf{S}^x$ and $\mathbf{S}^y$, we solve the following ridge regression problem for updating mapping $\mathbf{W}$:

$$
\min_{\mathbf{W}} \sum_{k=1}^{K} \|\mathbf{S}_k^y - \mathbf{W}_k \mathbf{S}_k^x\|_F^2 + \|\mathbf{S}_k^x - \mathbf{W}_k^{-1} \mathbf{S}_k^y\|_F^2 + \left(\frac{\gamma}{\beta}\right) \sum_{k=1}^{K} \|\mathbf{W}_k\|_F^2 . \tag{3.5}
$$

Particularly, the primal mapping function $\left\|\mathbf{S}_k^y - \mathbf{W}_k \mathbf{S}_k^x\right\|_F^2$ constructs an intrinsic mapping while the corresponding dual mapping function $\left\|\mathbf{S}_k^x - \mathbf{W}_k^{-1} \mathbf{S}_k^y\right\|_F^2$ is utilized to give feedbacks and further optimize the relationship between $\mathbf{S}_k^x$ and $\mathbf{S}_k^y$. Ideally (as the final solution), $\mathbf{S}_k^y = \mathbf{W}_k\mathbf{S}_k^x$, such that the problem in Eq. (3.5) is reduced to $\min_{\mathbf{W}_k} \sum_{k=1}^{K} \left\|\mathbf{S}_k^y - \mathbf{W}_k\mathbf{S}_k^x\right\|_F^2 + \left(\frac{\gamma}{\beta}\right) \sum_{k=1}^{K} \|\mathbf{W}_k\|_F^2$ with the solution $\mathbf{W} = \mathbf{S}_k^y\mathbf{S}_k^{xT}(\mathbf{S}_k^x\mathbf{S}_k^{xT} + \frac{\gamma}{\beta}\mathbf{I})^{-1}$, where $\mathbf{I}$ is an identity matrix. We summarize the proposed DOTE method in the following Algorithm 3.

Algorithm 3: DOTE algorithm

```
Input: Training data $\mathbf{X}$ and $\mathbf{Y}$, parameters $\lambda$, $\gamma$, $\beta$.
Initialize $\mathbf{F}_0^x$, $\mathbf{F}_0^y$, $\mathbf{S}_0^x$, $\mathbf{S}_0^y$, $\mathbf{W}_0$, $\mathbf{U}_0^x$, $\mathbf{U}_0^y$, $\mathbf{V}_0^x$, $\mathbf{V}_0^y$.
Perform FFT $\mathbf{S}_0^x \to \hat{\mathbf{S}}_0^x$, $\mathbf{S}_0^y \to \hat{\mathbf{S}}_0^y$, $\mathbf{F}_0^x \to \hat{\mathbf{F}}_0^x$, $\mathbf{F}_0^y \to \hat{\mathbf{F}}_0^y$, $\mathbf{U}_0^x \to \hat{\mathbf{U}}_0^x$, $\mathbf{U}_0^y \to \hat{\mathbf{U}}_0^y$, $\mathbf{V}_0^x \to \hat{\mathbf{V}}_0^x$,
    $\mathbf{V}_0^y \to \hat{\mathbf{V}}_0^y$.
Let $\hat{\mathbf{S}}_0^y \leftarrow \mathbf{W}\hat{\mathbf{S}}_0^x$.
while not converged do
    Solve for $\hat{\mathbf{S}}_{k+1}^x$, $\hat{\mathbf{S}}_{k+1}^y$, $\hat{\mathbf{U}}_{k+1}^x$ and $\hat{\mathbf{U}}_{k+1}^y$ using (3.3) with fixed filters and $\mathbf{W}_k$.
    Train $\hat{\mathbf{F}}_{k+1}^x$, $\hat{\mathbf{F}}_{k+1}^y$, $\hat{\mathbf{V}}_{k+1}^x$ and $\hat{\mathbf{V}}_{k+1}^y$ by (3.4) with fixed feature maps and $\mathbf{W}_k$.
    Update $\mathbf{W}_{k+1}$ by (3.5).
    Inverse FFT $\hat{\mathbf{F}}_{k+1}^x \to \mathbf{F}_{k+1}^x$, $\hat{\mathbf{F}}_{k+1}^y \to \mathbf{F}_{k+1}^y$.
end
Output: $\mathbf{F}^x$, $\mathbf{F}^y$, $\mathbf{W}$.
```

3.3.3 Synthesis

Once the optimization is completed, we can obtain the learned filters $\mathbf{F}^x$, $\mathbf{F}^y$ and the mapping $\mathbf{W}$. We then apply the proposed model to synthesize images across different modalities (i.e., LR → HR and $\mathcal{M}_1 \to \mathcal{M}_2$, respectively). Given a test image $\mathbf{X}^t$, we compute the sparse feature maps $\mathbf{S}^{tx}$ related to $\mathbf{F}^x$ by solving a single CSC problem like Eq. (3.1): $\mathbf{S}^{tx} = \arg\min_{\mathbf{S}^{tx}} \frac{1}{2}\left\|\mathbf{X}^t - \sum_{k=1}^{K} \mathbf{F}_k^x * \mathbf{S}_k^{tx}\right\|_2^2 + \lambda \sum_{k=1}^{K} \left\|\mathbf{S}_k^{tx}\right\|_1$. After that, we can synthesize the target modality image of $\mathbf{X}^t$ by the sum of K target feature maps $\mathbf{S}_k^{ty} = \mathbf{W}\mathbf{S}_k^{tx}$ convolved with $\mathbf{F}_k^y$, i.e., $\mathbf{Y}^t = \sum_{k=1}^{K} \mathbf{F}_k^y \mathbf{S}_k^{ty}$.

3.4 Experimental Results

3.4.1 Experimental Setup

The proposed DOTE is validated on two datasets: IXI[1] (including 578 $256 \times 256 \times p$ $p = 112 \sim 136$ MR healthy subjects) and NAMIC[2] (involving 20 $128 \times 128 \times 88$ subjects). In our

[1] http://brain-development.org/ixi-dataset/

[2] http://hdl.handle.net/1926/1687

avg.	ScSR	Zeyde	NLSR	ANR	A+	CSC-SR	DOTE
PSNR	29.98	33.10	33.97	35.23	35.72	36.18	**37.07**
SSIM	0.9265	0.9502	0.9548	0.9568	0.9600	0.9651	**0.9701**

Table 3.1 Quantitative evaluation: DOTE vs. other SR methods.

avg.	$\text{DOTE}_{\text{nodual}}\frac{1}{4}$	$\text{DOTE}_{\text{nodual}}\frac{1}{2}$	$\text{DOTE}_{\text{nodual}}\frac{3}{4}$	DOTE $\frac{1}{4}$	DOTE $\frac{1}{2}$	DOTE $\frac{3}{4}$
PSNR	31.23	33.17	36.09	**36.56**	**36.68**	**37.07**
SSIM	0.9354	0.9523	0.9581	**0.9687**	**0.9690**	**0.9701**

Table 3.2 Quantitative evaluation: DOTE vs. $\text{DOTE}_{\text{nodual}}$.

experiments, we perform 4-fold cross-validation for testing. That is, selecting 144 subjects from IXI and 5 subjects from NAMIC, respectively, as our test data. Following [155, 168], the regularization parameters λ, β, and γ are empirically set to be 0.05, 0.10, 0.15, respectively. The number of filters is set as 800 according to [51]. Convergence towards primal feasible solution is proved in [10] by first converting Eq. (3.2) into two optimization sub-problems that involve two proxies $\mathbf{U}$, $\mathbf{V}$ and then solving them alternatively. DOTE converges after ca. 10 iterations. For the evaluation criteria, we adopt PSNR and SSIM index to objectively assess the quality of our results.

MRI Super-Resolution

As we introduced in Section 3.1, we first address image SR as one of cross-modality image synthesis. In this scenario, we investigate the T2-w images of the IXI dataset for evaluating and comparing DOTE with ScSR [168], A+ [148], NLSR [125], Zeyde [176], ANR [147], and CSC-SR [51]. Generally, LR images are generated by down-sampling HR ground-truth images using bicubic interpolation. We perform image SR with scaling factor 2, and show visual results in Fig. 3.2. The quantitative results are reported in Fig. 3.3, while the average PSNRs and SSIMs for all 144 test subjects are shown in Table 3.1. The proposed model achieves the best PSNRs and SSIMs. Moreover, to validate our argument that DL-based self-optimization strategy is beneficial and requires less training data, we compare $\text{DOTE}_{\text{nodual}}$ (removing dual mapping term) and DOTE under different training data size (i.e., $\frac{1}{4}, \frac{1}{2}, \frac{3}{4}$ of the original dataset). The results are listed in Table 3.2. From Table 3.2, we see that DOTE is always better than $\text{DOTE}_{\text{nodual}}$ especially with few training samples.

Metric(avg.)	NAMIC					
	T1− >T2			T2− >T1		
	MIMECS	Vemulapalli	DOTE	MIMECS	Vemulapalli	DOTE
PSNR	24.98	27.22	**29.83**	27.13	28.95	**32.03**
SSIM	0.8821	0.8981	**0.9013**	0.9198	0.9273	**0.9301**

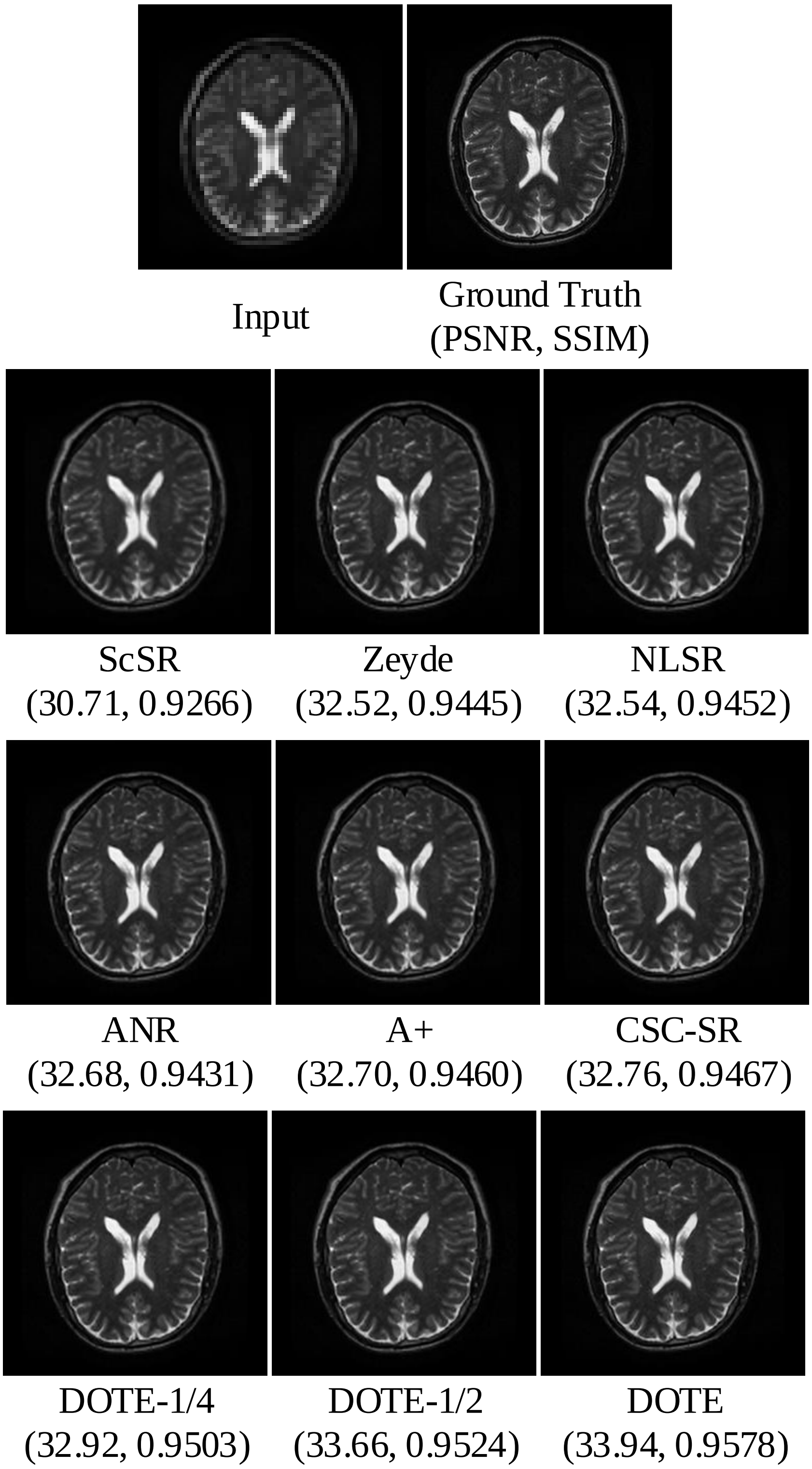
Table 3.3 CMS results: DOTE vs. other synthesis methods on the NAMIC dataset.

Cross-Modality Synthesis

For the problem of CMS, we evaluate DOTE and the relevant algorithms on both datasets involving four groups of experiments: (1) synthesizing T2-w image from PD-w acquisition and (2) *vice versa*; (3) generating T1-w image from T2-w input, and (4) *vice versa*. We conduct (1-2) experiments on the IXI dataset, while (3-4) are explored on the NAMIC dataset. The representative and state-of-the-art CMS methods, including Vemulapalli's method [153] and MIMECS [127] are employed to compare with our DOTE approach. We demonstrate visual and quantitative results in Fig. 3.4, Fig. 3.5 and Table. 3.3, respectively. Our algorithm yields the best results against MIMECS and Vemulapalli for two datasets validating our claim of being able to synthesize better results through the expanded dual optimization.

3.5 Conclusions

We presented a dual convolutional filter learning (DOTE) method which directly decomposes the whole image based on CSC, such that local neighbors are preserved consistently. The proposed dual mapping functions integrated with joint learning model form a closed loop that leverages the training data more efficiently and keeps a very stable mapping between image modalities. We applied DOTE to both image SR and CMS problems. Extensive results showed that our method outperforms other state-of-the-art approaches. Future work could concentrate on extending DOTE to higher-order imaging modalities like diffusion tensor MRI and to other modalities beyond MRI.



Input

Ground Truth
(PSNR, SSIM)

ScSR
(30.71, 0.9266)

Zeyde
(32.52, 0.9445)

NLSR
(32.54, 0.9452)

ANR
(32.68, 0.9431)

A+
(32.70, 0.9460)

CSC-SR
(32.76, 0.9467)

DOTE-1/4
(32.92, 0.9503)

DOTE-1/2
(33.66, 0.9524)

DOTE
(33.94, 0.9578)

Fig. 3.2 Example SR results and the corresponding PSNRs and SSIMs.

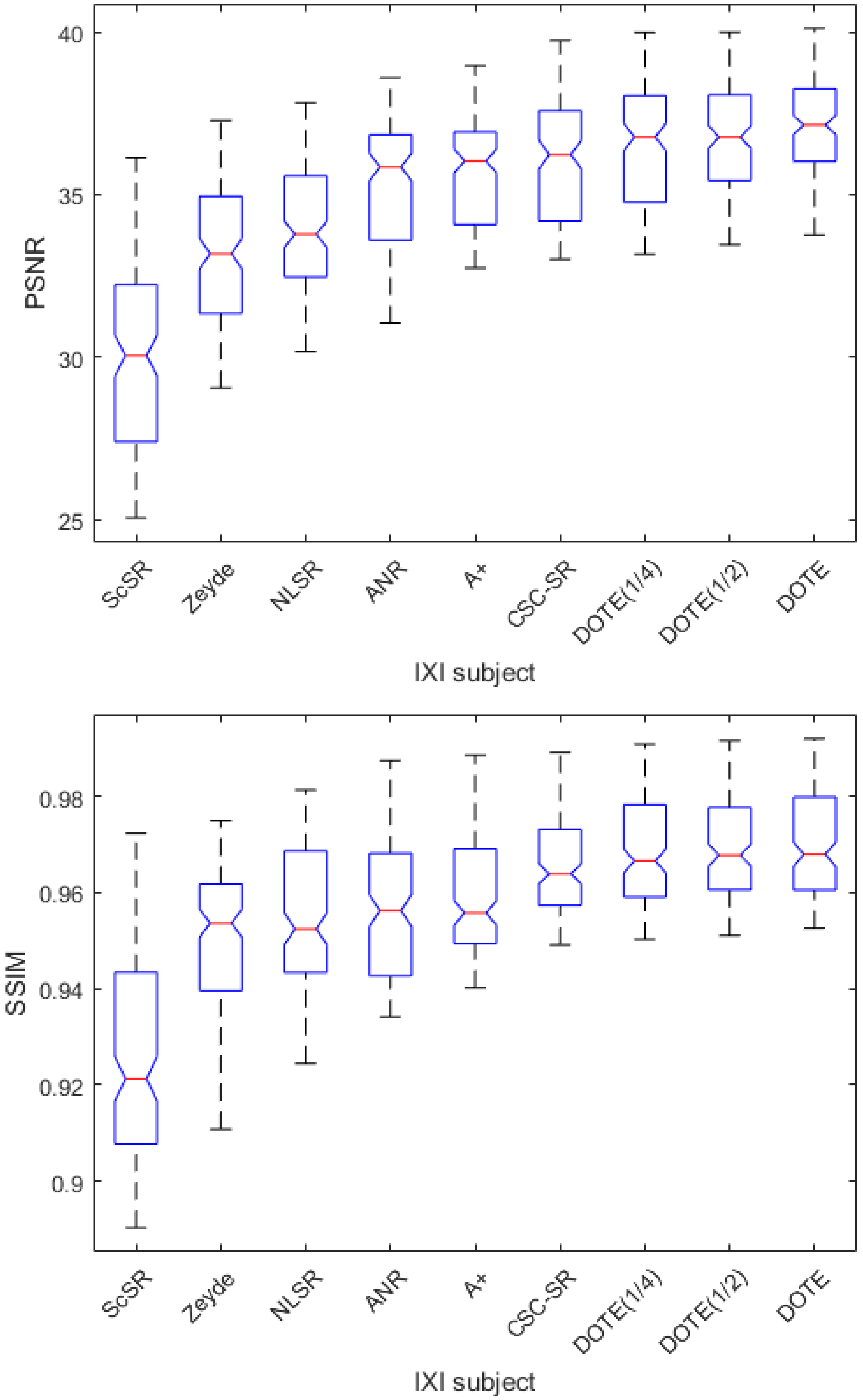


Fig. 3.3 Error measures of SR results on the IXI dataset.

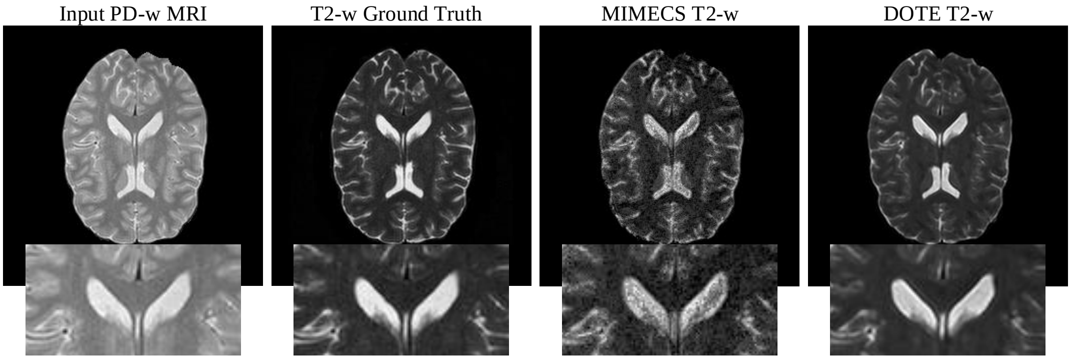


Fig. 3.4 Visual comparison of synthesized results using MIMECS and DOTE.

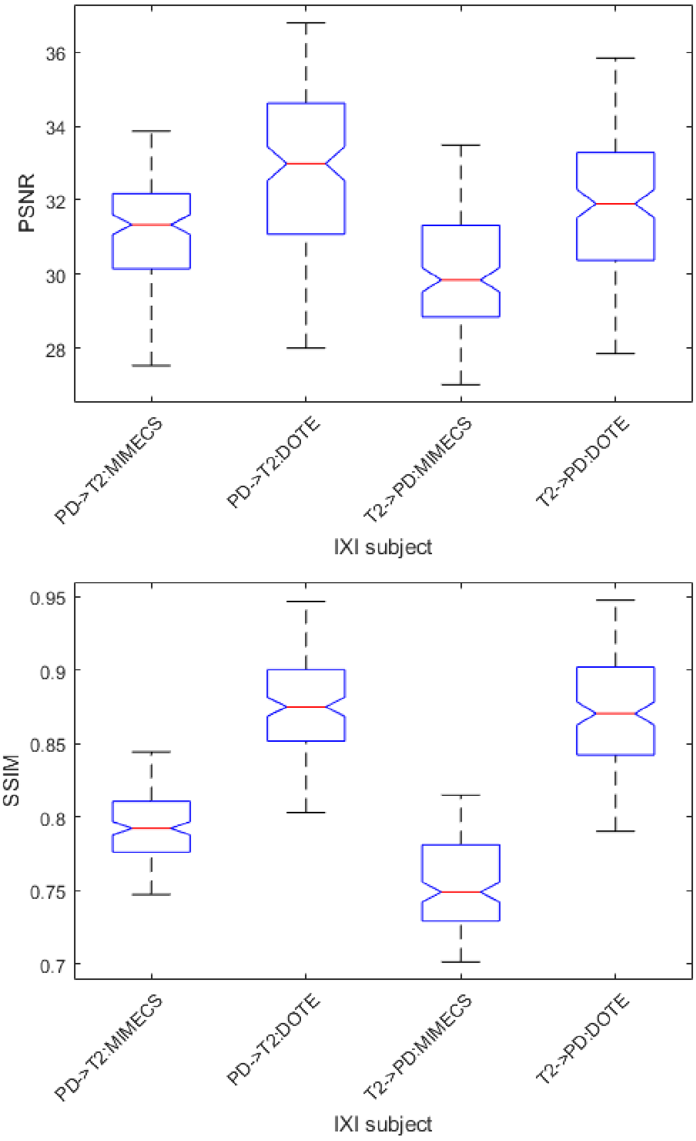


Fig. 3.5 CMS results: DOTE vs. MIMECS on the IXI dataset.

Chapter 4

Weakly-Supervised Joint Convolutional Sparse Coding

Magnetic Resonance Imaging (MRI) offers high-resolution *in vivo* imaging and rich functional and anatomical multimodality tissue contrast. In practice, however, there are challenges associated with considerations of scanning costs, patient comfort, and scanning time that constrain how much data can be acquired in clinical or research studies. In this chapter, we explore the possibility of generating high-resolution and multimodal images from low-resolution single-modality imagery. We propose the weakly-supervised joint convolutional sparse coding to simultaneously solve the problems of super-resolution (SR) and cross-modality image synthesis. The learning process requires only a few registered multimodal image pairs as the training set. Additionally, the quality of the joint dictionary learning can be improved using a larger set of unpaired images[1]. To combine unpaired data from different image resolutions/modalities, a hetero-domain image alignment term is proposed. Local image neighborhoods are naturally preserved by operating on the whole image domain (as opposed to image patches) and using joint convolutional sparse coding. The paired images are enhanced in the joint learning process with unpaired data and an additional maximum mean discrepancy term, which minimizes the dissimilarity between their feature distributions. Experiments show that the proposed method outperforms state-of-the-art techniques on both SR reconstruction and simultaneous SR and cross-modality synthesis.

[1] Unpaired data/images: acquisitions are from different subjects without registration. Paired data/images: acquisitions of the same subject obtained from different modalities are registered.

4.1 Introduction

With the rapid progress in Magnetic Resonance Imaging (MRI), there are a multitude of mechanisms to generate tissue contrast that are associated with various anatomical or functional features. However, the acquisition of a complete multimodal set of high-resolution images faces constraints associated with scanning costs, scanner availability, scanning time, and patient comfort. In addition, long-term longitudinal studies such as ADNI [111] imply that changes exist in the scanner or acquisition protocol over time. In these situations, it is not uncommon to have images of the same subject but obtained from different sources, or to be confronted with missing or corrupted data from earlier time points. In addition, high-resolution (HR) 3D medical imaging usually requires long breath-hold and repetition times, which lead to long-term scanning times that are challenging or unfeasible in clinical routine. Acquiring low-resolution (LR) images and/or skipping some imaging modalities altogether from the acquisition are then not uncommon. In all such scenarios, it is highly desirable to be able to generate HR data from the desired target modality from the given LR modality data.

The relevant literature in this area can be divided into either super-resolution (SR) reconstruction from single/multiple image modalities or cross-modality (image) synthesis (CMS). On the one hand, SR is typically concerned with achieving improved visual quality or overcoming the resolution limits of the acquired image data. Such a problem is generally under-determined and ill-posed, hence, the solution is not unique. To mitigate this fact, the solution space needs to be constrained by incorporating strong priors. Prior information comes in the form of smoothness assumptions as in, for example, interpolation-based SR [78, 131]. State-of-the-art methods mostly adopt either external data or internal data to guide the learning algorithms [125, 144]. On the other hand, due to variations in optimal image representations across modalities, the learned image model from one modality data may not be the optimal model for a different modality. How to reveal the relationship between different representations of the underlying image information is a major research issue to be explored. In order to synthesize one modality from another, recent methods in CMS proposed utilizing non-parametric methods like nearest neighbor (NN) search [41], nonlinear regression forests [73], coupled dictionary learning [127], and convolutional neural network (CNN) [44], to name a few. Although these algorithms achieve remarkable results, most of them suffer from the fundamental limitations associated with supervised learning and/or patch-based synthesis. Supervised approaches require a large number of training image pairs, which is impractical in many medical imaging applications. Patch-based synthesis suffers from inconsistencies introduced during the fusion process that takes place in areas where patches overlap.

In this chapter, we propose a weakly-supervised convolutional sparse coding method with an application to neuroimaging that utilizes a small set of registered multimodal image pairs and solves the SR and CMS problems simultaneously. Rather than factorizing each patch into a linear combination of patches drawn from a dictionary built under sparsity constraints (sparse coding), or requiring a training set with fully registered multimodal image pairs, or requiring the same sparse code to be used for both modalities involved, we generate a unified learning model that automatically learns a joint representation for heterogeneous data (e.g., different resolutions, modalities and relative poses). This representation is learned in a common feature space that preserves the local consistency of the images. Specifically, we utilize the co-occurrence of texture features across both domains. A manifold ranking method picks features of the target domain from the most similar subjects in the source domain. Once the correspondence between images in different domains is established, we directly work on a whole image representation that intrinsically respects local neighborhoods. Furthermore, a mapping function is learned that links the representations between the two modalities involved. We call the proposed method WEakly-supErvised joiNt convolutIonal sparsE coding (WEENIE), and perform extensive experiments to verify its performance.

The main contributions of this chapter are as follows:

1. This is the first attempt to jointly solve the SR and CMS problems in 3D medical imaging using weakly-supervised joint convolutional sparse coding.

2. To exploit unpaired images from different domains during the learning phase, a hetero-domain image alignment term is proposed, which allows identifying correspondences across source and target domains and is invariant to pose transformations.

3. To map LR and HR cross-modality image pairs, joint learning based on convolutional sparse coding is proposed that includes a maximum mean discrepancy term.

4. Finally, extensive experimental results show that the proposed model yields better performance than state-of-the-art methods in both reconstruction error and visual quality assessment measures.

4.2 Preliminaries

Convolutional Sparse Coding (CSC) was introduced in the context of modeling receptive fields preciously, and later generalized to image processing, in which the representation of an entire image is computed by the sum of a set convolutions with dictionary filters. The goal of CSC is to remedy the shortcoming of conventional patch-based sparse coding methods by removing shift variations for consistent approximation of local neighbors on whole images. Concretely, given the vectorized image $\mathbf{x}$, the problem of generating a set of vectorized filters for sparse feature maps is solved by minimizing the objective function that combines the squared reconstruction error and the l_1-norm penalty on the representations:

$$\begin{aligned} &\arg\min_{\mathbf{f},\mathbf{z}} \frac{1}{2}\left\|\mathbf{x} - \sum_{k=1}^{K} \mathbf{f}_k * \mathbf{z}_k\right\|_2^2 + \lambda \sum_{k=1}^{K} \|\mathbf{z}_k\|_1 \\ &s.t.\ \|\mathbf{f}_k\|_2^2 \leq 1\ \forall k = \{1,...,K\}, \end{aligned} \tag{4.1}$$

where $\mathbf{x}$ is an $m \times n$ image in vector form, $\mathbf{f}_k$ refers to the k-th $d \times d$ filter in vector form, $\mathbf{z}_k$ is the sparse feature map corresponding to $\mathbf{f}_k$ with size $(m+d-1) \times (n+d-1)$ to approximate $\mathbf{x}$, λ controls the l_1 penalty, and $*$ denotes the 2D convolution operator. $\mathbf{f} = \left[\mathbf{f}_1^T,...,\mathbf{f}_K^T\right]^T$ and $\mathbf{z} = \left[\mathbf{z}_1^T,...,\mathbf{z}_K^T\right]^T$ are K filters and feature maps stacked as the single column vector, respectively. Here, the inequality constraint on each column of vectorized $\mathbf{f}_k$ prevents the filter from absorbing all the energy of the system.

Similar to the original sparse coding problem, Zeiler *et al.* [175] proposed to solve the CSC in Eq. (4.1) through alternatively optimizing one variable while fixing the other one in the spatial domain. Advances in recent fast convolutional sparse coding (FCSC) [10] have shown that feature learning can be efficiently and explicitly solved by incorporating CSC within an alternating direction method of multipliers (ADMMs) framework in the Fourier domain.

4.2.1 Problem Formulation

The simultaneous SR and cross-modality synthesis problem can be formulated as: given a three-dimensional LR image $\mathbf{X}$ of modality $\mathscr{M}_1$, the task is to infer from $\mathbf{X}$ a target 3D image $\mathbf{Y}$ that is as similar as possible to the HR ground truth of desirable modality $\mathscr{M}_2$. Suppose that we are given a group of LR images of modality $\mathscr{M}_1$, i.e., $\mathscr{X} = [\mathbf{X}_1,...,\mathbf{X}_P] \in \mathbb{R}^{m\times n\times t\times P}$, and a set of HR images of modality $\mathscr{M}_2$, i.e., $\mathscr{Y} = [\mathbf{Y}_2,...,\mathbf{Y}_Q] \in \mathbb{R}^{m\times n\times t\times Q}$. P and Q are the numbers of samples in the training sets, and m, n denote the dimensions of axial view of each image, while t is the size of the image along the z-axis. Moreover, in both training sets, subjects of source modality $\mathscr{M}_1$ are mostly different from target modality $\mathscr{M}_2$, that is, we are working with a small number of paired data while most of them are unpaired. Therefore, the difficulties of this problem vary with hetero-domain images, e.g., resolutions and modalities, and how well the two domains fit. To bridge image appearances across heterogeneous representations, we propose a method for automatically establishing a one-to-one correlation between data in $\mathscr{X}$ and $\mathscr{Y}$ firstly, then employ the aligned data to jointly learn a pair of filters, while assuming that there exists a mapping function $\mathscr{F}(\cdot)$ for associating and predicting cross-modality data in the projected common feature space. Particularly, we want to synthesize MRI of human brains in this chapter. An overview of our proposed work is depicted in Fig. 4.1.

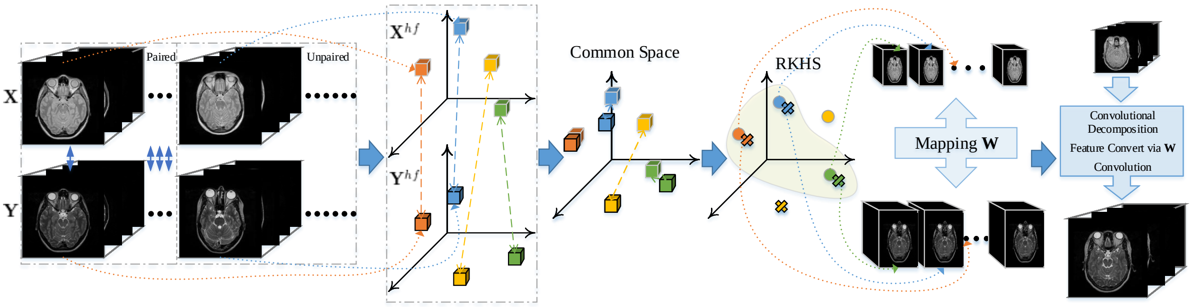


Fig. 4.1 Flowchart of the proposed method (WEENIE) for simultaneous SR and cross-modality synthesis.

4.2.2 Notation

For simplicity, we denote matrices and 3D images as upper-case bold (e.g., image $\mathbf{X}$), vectors and vectorized 2D images as lower-case bold (e.g., filter $\mathbf{f}$), and scalars as lower-case (e.g., the number of filter k). Image with modality $\mathscr{M}_1$ called source modality belongs to the source domain, and with modality $\mathscr{M}_2$ called target modality belongs to the target domain.

4.3 Weakly-Supervised Joint Convolutional Sparse Coding

4.3.1 Hetero-Domain Image Alignment

The design of an alignment $\mathscr{A}(\cdot)$ from $\mathscr{X}$ to $\mathscr{Y}$ requires a combination of extracting common components from LR/HR images and some measures of correlation between both modalities. In SR literature, common components are usually accomplished by extracting high-frequency (HF) edges and texture features from LR/HR images, respectively [16, 168]. In this chapter, we adopt first- and second-order derivatives involving horizontal and vertical gradients as the features for LR images by $\mathbf{X}_p^{hf} = \mathscr{G} * \mathbf{X}_p$. $\mathscr{G} = \begin{bmatrix} \mathbf{G}_1^1, \mathbf{G}_1^2 \\ \mathbf{G}_2^1, \mathbf{G}_2^2 \end{bmatrix}$, and each gradient $\mathbf{G}$ has the same length of z-axis as input image while $\mathbf{g}_1^1 = [-1, 0, 1]$, $\mathbf{g}_1^2 = {\mathbf{g}_1^1}^T$, and $\mathbf{g}_2^1 = [-2, -1, 0, 1, 2]$, $\mathbf{g}_2^2 = {\mathbf{g}_2^1}^T$. For HR images, HF features are obtained through directly subtracting mean value, i.e., $\mathbf{Y}_p^{hf} = \mathbf{Y}_p - mean(\mathbf{Y}_p)$. To define the hetero-domain image alignment term $\mathscr{A}(\cdot)$, we assume that the intrinsic structures of brain MRI of a subject across image modalities are also similar in the HF space since images of different modalities are more likely to be described differently by features. When HF features of both domains are obtained, it is possible to build a way for cross-modality data alignment (in particular, a unilateral cross-modality matching can be thought as a special case in [66]). To this end, we define a subject-specific transformation matrix $\mathbb{A}$ as

$$\mathbb{A} = \begin{bmatrix} K(\mathbf{X}_1^{hf}, \mathbf{Y}_1^{hf}) & \cdots & K(\mathbf{X}_1^{hf}, \mathbf{Y}_Q^{hf}) \\ \vdots & \ddots & \vdots \\ K(\mathbf{X}_P^{hf}, \mathbf{Y}_1^{hf}) & \cdots & K(\mathbf{X}_P^{hf}, \mathbf{Y}_Q^{hf}) \end{bmatrix}, \tag{4.2}$$

where $K(\mathbf{X}_p^{hf}, \mathbf{Y}_q^{hf})$ is used for measuring the distances between each pair of HF data in $\mathscr{X}$ and $\mathscr{Y}$ computed by the Gaussian kernel as

$$K(\mathbf{X}_p^{hf}, \mathbf{Y}_q^{hf}) = \frac{1}{(\sqrt{2\pi}\sigma)^3} e^{-\frac{|\mathbf{X}_p^{hf} - \mathbf{Y}_q^{hf}|^2}{2\sigma^2}}, \tag{4.3}$$

where σ determines the width of Gaussian kernel. In order to establish a one-to-one correspondence across different domains, for each element of $\mathscr{X}$, the most relevant image with maximum K from $\mathscr{Y}$ is preserved while discarding the rest of the elements:

$$\mathbb{A} = \begin{bmatrix} \max\left(K(1,:)\right) & & \\ & \ddots & \\ & & \max\left(K(P,:)\right) \end{bmatrix}, \tag{4.4}$$

where $\max\left(K(p,:)\right)$ denotes the maximum element of the p-th row of $\mathbb{A}$. We further set $\max\left(K(p,:)\right)$ to 1, and all the blank elements to 0. Therefore, $\mathbb{A}$ is a binary matrix. Since $\mathbb{A}$ is calculated in a subject-specific manner, each subject of $\mathscr{X}$ can only be connected to one target of the most similar brain structures. Hence, images under a hetero-domain can be treated as being the registered pairs, i.e., $\mathscr{P}_i = \{\mathbf{X}_i, \mathbf{Y}_i\}_{i=1}^{P}$, by constructing virtual correspondence: $\mathscr{A}(\mathscr{X}, \mathscr{Y}) = \left\|\mathbf{X}^{hf} - \mathbb{A}\mathbf{Y}^{hf}\right\|_2^2$.

4.3.2 Objective Function

For image modality transformation, coupled sparse coding [70, 168] has important advantages, such as reliability of correspondence dictionary pair learning and less memory cost. However, the arbitrarily aligned bases related to the small part of images may lead to shifted versions of the same structures or inconsistent representations based on the overlapped patches. CSC [175] was then proposed to generate a global decomposition framework based on the whole image for solving the above problem. In spired by CSC and the benefits of coupled sparsity [70], we introduce a joint convolutional sparse coding method in a weakly-supervised setting for hetero-domain images. The small number of originally registered pairs are used to carry the intrinsic relationship between $\mathscr{X}$ and $\mathscr{Y}$ while the majority of unpaired data are introduced to exploit and enhance the diversity of the original learning system.

Assume that the aforementioned alignment approach leads to a perfect correspondence across $\mathscr{X}$ and $\mathscr{Y}$, such that each aligned pair of images possesses approximately identical (or the same for co-registered data) information. Moreover, to facilitate image mappings in a joint manner, we require sparse feature maps of each pair of corresponding source and target images to be associated. That is, suppose that there exists a mapping function $\mathscr{F}(\cdot)$, where the feature maps of LR $\mathscr{M}_1$ modality images can be converted to their HR $\mathscr{M}_2$ versions. Given $\mathscr{X}$ and $\mathscr{Y}$, we propose to learn a pair of filters with corresponding feature maps and a mapping function pre-processed with the aligned term $\left\|\mathbf{X}^{hf} - \mathbb{A}\mathbf{Y}^{hf}\right\|_2^2$ by

$$\begin{aligned} \arg\min_{\mathbf{F}^x,\mathbf{F}^y,\mathbf{Z}^x,\mathbf{Z}^y,\mathbf{W}} \frac{1}{2}\left\|\mathbf{X} - \sum_{k=1}^{K} \mathbf{F}_k^x * \mathbf{Z}_k^x\right\|_F^2 \\ +\frac{1}{2}\left\|\mathbf{Y} - \sum_{k=1}^{K} \mathbf{F}_k^y * \mathbf{Z}_k^y\right\|_F^2 + \beta\sum_{k=1}^{K} \|\mathbf{Z}_k^y - \mathbf{W}_k\mathbf{Z}_k^x\|_F^2 \\ +\lambda\left(\sum_{k=1}^{K} \|\mathbf{Z}_k^x\|_1 + \sum_{k=1}^{K} \|\mathbf{Z}_k^y\|_1\right) + \gamma\sum_{k=1}^{K} \|\mathbf{W}_k\|_F^2 \\ s.t.\ \|\mathbf{f}_k^x\|_2^2 \le 1, \|\mathbf{f}_k^y\|_2^2 \le 1\,\forall k, \end{aligned} \tag{4.5}$$

where $\mathbf{Z}_k^x$ and $\mathbf{Z}_k^y$ are the k-th sparse feature maps that estimate the aligned data terms $\mathbf{X}$ and $\mathbf{Y}$ when convolved with the k-th filters $\mathbf{F}_k^x$ and $\mathbf{F}_k^y$ of a fixed spatial support, $\forall k = \{1,...,K\}$. Concretely, $\mathbf{X}$ denotes the aligned image from $\mathscr{P}$ with LR and $\mathscr{M}_1$ modality; $\mathbf{Y}$ denotes the aligned image from $\mathscr{P}$ containing HR and $\mathscr{M}_2$ modality. A convolution operation is represented as $*$ operator, and $\|\cdot\|_F$ denotes a Frobenius norm chosen to induce the convolutional least squares approximate solution. $\mathbf{F}^x$ and $\mathbf{F}^y$ are adopted to list all K filters, while $\mathbf{Z}^x$ and $\mathbf{Z}^y$ represent corresponding K feature maps for source and target domains, respectively. $\mathscr{A}(\mathscr{X},\mathscr{Y})$ is combined to enforce the correspondence for unpaired auxiliary subjects. The mapping function $\mathscr{F}\left(\mathbf{Z}_k^x, \mathbf{W}_k\right) = \mathbf{W}_k\mathbf{Z}_k^x$ is modeled as a linear projection $\mathbf{W}_k$ of $\mathbf{Z}_k^x$ and $\mathbf{Z}_k^y$ by solving a set of the least squares problem (i.e., $\min_{\mathbf{W}} \sum_{k=1}^{K} \left\|\mathbf{Z}_k^y - \mathbf{W}_k\mathbf{Z}_k^x\right\|_F^2$). Parameters λ, β and γ balance sparsity, feature representation and association mapping.

It is worth noting that $\mathscr{P}_i = \{\mathbf{X}_i, \mathbf{Y}_i\}$ may not be perfect since HF feature alignment in Eq. (4.4) is not good enough for very heterogeneous domain adaptation by matching the first- and second-order derivatives of $\mathscr{X}$ and means of $\mathscr{Y}$, which leads to suboptimal filter pairs and inaccurate results. To overcome such a problem, we need additional constraints to ensure the correctness of registered image pairs produced by the alignment. Generally, when feature difference is substantially large, there always exists some subjects of the source domain that are not particularly related to target ones even in the HF subspace. Thus, a registered subject pairs' divergence assessment procedure should be cooperated with the aforementioned joint learning model to handle this difficult setting. Recent works [19, 96, 183] have performed instance/domain adaptation via measuring data distribution divergence using the maximum mean discrepancy (MMD) criterion. We follow such an idea and employ the empirical MMD as the nonparametric distribution measure to handle the hetero-domain image pair mismatch problem in the reproducing kernel Hilbert space (RKHS). This is done by minimizing the difference between distributions of aligned subjects while keeping dissimilar 'registered' pairs (i.e., discrepant distributions) apart in the sparse feature map space:

$$\begin{aligned}&\frac{1}{P}\sum_{i=1}^{P}\sum_{k=1}^{K}\|\mathbf{W}_k(i)\mathbf{Z}_k^x(i)-\mathbf{Z}_k^y(i)\|^2_{\mathscr{H}}\\ =&\sum_{k=1}^{K}(\mathbf{W}_k\mathbf{Z}_k^x)^T M_i\mathbf{Z}_k^y = Tr(\sum_{k=1}^{K}\mathbf{Z}_k^y\mathbf{M}(\mathbf{W}_k\mathbf{Z}_k^x)^T),\end{aligned} \tag{4.6}$$

where $\mathscr{H}$ indicates RKHS space, $\mathbf{Z}_k^x(i)$ and $\mathbf{Z}_k^y(i)$ are the paired sparse feature maps for $\mathscr{P}_i = \{\mathbf{X}_i, \mathbf{Y}_i\}$ with $i = 1,...P$, M_i is the i-th element of $\mathbf{M}$ while $\mathbf{M}$ denotes the MMD matrix and can be computed as follows

$$M_i = \left\{\begin{array}{cc}\frac{1}{P}, & \mathbf{Z}_k^x(i),\mathbf{Z}_k^y(i)\in\mathscr{P}_i\\ -\frac{1}{P^2}, & \text{otherwise.}\end{array}\right., \tag{4.7}$$

By regularizing Eq. (4.5) with Eq. (4.6), filter pairs $\mathbf{F}_k^x$ and $\mathbf{F}_k^y$ are refined and the distributions of real aligned subject pairs are drawn close under the new feature maps. Putting the above together, we obtain the objective function with the pre-alignment (i.e., $\|\mathbf{X}^{hf}-\mathbb{A}\mathbf{Y}^{hf}\|_2^2$):

$$\begin{aligned}\arg\min_{\mathbf{F}^x,\mathbf{F}^y,\mathbf{Z}^x,\mathbf{Z}^y,\mathbf{W}}\ &\frac{1}{2}\left\|\mathbf{X}-\sum_{k=1}^{K}\mathbf{F}_k^x*\mathbf{Z}_k^x\right\|_F^2+\gamma\sum_{k=1}^{K}\|\mathbf{W}_k\|_F^2\\ &+\frac{1}{2}\left\|\mathbf{Y}-\sum_{k=1}^{K}\mathbf{F}_k^y*\mathbf{Z}_k^y\right\|_F^2+\beta\sum_{k=1}^{K}\|\mathbf{Z}_k^y-\mathbf{W}_k\mathbf{Z}_k^x\|_F^2\\ &+\lambda\left(\sum_{k=1}^{K}\|\mathbf{Z}_k^x\|_1+\sum_{k=1}^{K}\|\mathbf{Z}_k^y\|_1\right)+Tr(\sum_{k=1}^{K}\mathbf{Z}_k^y\mathbf{M}(\mathbf{W}_k\mathbf{Z}_k^x)^T)\\ &s.t.\ \|\mathbf{f}_k^x\|_2^2\le 1,\|\mathbf{f}_k^y\|_2^2\le 1\,\forall k.\end{aligned} \tag{4.8}$$

4.3.3 Optimization

We propose a three-step optimization strategy for efficiently tackling the objective function in Eq. (4.8) (termed (WEENIE), summarized in Algorithm 4) considering that such multivariables and unified framework cannot be jointly convex to **F**, **Z**, and **W**. Instead, it is convex with respect to each of them while fixing the remaining variables.

```
Algorithm 4: WEENIE Algorithm
Input: Training data $\mathbf{X}$ and $\mathbf{Y}$, parameters $\lambda$, $\gamma$, $\sigma$.
Initialize $\mathbf{F}_0^x$, $\mathbf{F}_0^y$, $\mathbf{Z}_0^x$, $\mathbf{Z}_0^y$, $\mathbf{U}_0^x$, $\mathbf{U}_0^y$, $\mathbf{S}_0^x$, $\mathbf{S}_0^y$, $\mathbf{W}_0$.
Perform FFT $\mathbf{Z}_0^x \rightarrow \hat{\mathbf{Z}}_0^x$, $\mathbf{Z}_0^y \rightarrow \hat{\mathbf{Z}}_0^y$, $\mathbf{F}_0^x \rightarrow \hat{\mathbf{F}}_0^x$, $\mathbf{F}_0^y \rightarrow \hat{\mathbf{F}}_0^y$, $\mathbf{U}_0^x \rightarrow \hat{\mathbf{U}}_0^x$, $\mathbf{U}_0^y \rightarrow \hat{\mathbf{U}}_0^y$, $\mathbf{S}_0^x \rightarrow \hat{\mathbf{S}}_0^x$,
  $\mathbf{S}_0^y \rightarrow \hat{\mathbf{S}}_0^y$.
Let $\hat{\mathbf{Z}}_0^y \leftarrow \mathbf{W}\hat{\mathbf{Z}}_0^x$.
while not converged do
    Fix other variables, update $\hat{\mathbf{Z}}_{k+1}^x$, $\hat{\mathbf{Z}}_{k+1}^y$ and $\hat{\mathbf{U}}_{k+1}^x$, $\hat{\mathbf{U}}_{k+1}^y$ by (4.9).
    Fix other variables, update $\hat{\mathbf{F}}_{k+1}^x$, $\hat{\mathbf{F}}_{k+1}^y$ and $\hat{\mathbf{S}}_{k+1}^x$, $\hat{\mathbf{S}}_{k+1}^y$ by (4.10) with $\hat{\mathbf{Z}}_{k+1}^x$, $\hat{\mathbf{Z}}_{k+1}^y$,
      $\hat{\mathbf{U}}_{k+1}^x$, $\hat{\mathbf{U}}_{k+1}^y$ and $\mathbf{W}_k$.
    Fix other variables, update $\mathbf{W}_k$ by (4.11) with $\hat{\mathbf{Z}}_{k+1}^x$, $\hat{\mathbf{Z}}_{k+1}^y$, $\hat{\mathbf{U}}_{k+1}^x$, $\hat{\mathbf{U}}_{k+1}^y$, $\hat{\mathbf{F}}_{k+1}^x$, $\hat{\mathbf{F}}_{k+1}^y$,
      and $\hat{\mathbf{S}}_{k+1}^x$, $\hat{\mathbf{S}}_{k+1}^y$.
    Inverse FFT $\hat{\mathbf{F}}_{k+1}^x \rightarrow \mathbf{F}_{k+1}^x$, $\hat{\mathbf{F}}_{k+1}^y \rightarrow \mathbf{F}_{k+1}^y$.
end
Output: $\mathbf{F}^x$, $\mathbf{F}^y$, $\mathbf{W}$.
```

Computing Convolutional Sparse Coding

Optimization involving only sparse feature maps $\mathbf{Z}^x$ and $\mathbf{Z}^y$ is solved by initialization of filters $\mathbf{F}^x$, $\mathbf{F}^y$ and mapping function **W** (**W** is initialized as an identity matrix). Besides the original CSC formulation, we have additional terms associated with data alignment and divergence reducing in the common feature space. Eq. (4.8) is firstly converted to two regularized sub-CSC problems. Unfortunately, each of the problems constrained with an l_1 penalty term cannot be directly solved, which is not rotation invariant. Recent approaches [10, 57] have been proposed to work around this problem on the theoretical derivation by introducing two auxiliary variables **U** and **S** to enforce the constraint inherent in the splitting. To facilitate component-wise multiplications, we exploit the convolution subproblem [10] in the Fourier domain[2] derived

[2] Fast Fourier transform (FFT) is utilized to solve the relevant linear system and demonstrated substantially better asymptotic performance than processed in the spatial domain.

within the ADMMs framework:

$$
\begin{aligned}
\min_{\mathbf{Z}^x} \frac{1}{2} \left\| \hat{\mathbf{X}} - \sum_{k=1}^{K} \hat{\mathbf{F}}_k^x \odot \hat{\mathbf{Z}}_k^x \right\|_F^2 + \left\| \mathbf{X}^{hf} - \mathbb{A}\mathbf{Y}^{hf} \right\|_2^2 \\
+ Tr(\sum_{k=1}^{K} \hat{\mathbf{Z}}_k^y \mathbf{M} (\mathbf{W}_k \hat{\mathbf{Z}}_k^x)^T) + \beta \sum_{k=1}^{K} \| \hat{\mathbf{Z}}_k^y - \mathbf{W}_k \hat{\mathbf{Z}}_k^x \|_F^2 \\
+ \lambda \sum_{k=1}^{K} \| \mathbf{U}_k^x \|_1 \ s.t. \ \| \mathbf{S}_k^x \|_2^2 \leq 1, \mathbf{S}_k^x = \Phi^T \hat{\mathbf{F}}_k^x, \mathbf{U}_k^x = \mathbf{Z}_k^x \ \forall k, \\
\\
\min_{\mathbf{Z}^y} \frac{1}{2} \left\| \hat{\mathbf{Y}} - \sum_{k=1}^{K} \hat{\mathbf{F}}_k^y \odot \hat{\mathbf{Z}}_k^y \right\|_F^2 + \left\| \mathbf{X}^{hf} - \mathbb{A}\mathbf{Y}^{hf} \right\|_2^2 \\
+ Tr(\sum_{k=1}^{K} \hat{\mathbf{Z}}_k^y \mathbf{M} (\mathbf{W}_k \hat{\mathbf{Z}}_k^x)^T) + \beta \sum_{k=1}^{K} \| \hat{\mathbf{Z}}_k^x - \mathbf{W}_k \hat{\mathbf{Z}}_k^y \|_F^2 \\
+ \lambda \sum_{k=1}^{K} \| \mathbf{U}_k^y \|_1 \ s.t. \ \| \mathbf{S}_k^y \|_2^2 \leq 1, \mathbf{S}_k^y = \Phi^T \hat{\mathbf{F}}_k^y, \mathbf{U}_k^y = \mathbf{Z}_k^y \ \forall k,
\end{aligned}
\tag{4.9}
$$

where ˆ applied to any symbol indicates the discrete Fourier transform (DFT), for example $\hat{\mathbf{X}} \leftarrow f(\mathbf{X})$, and $f(\cdot)$ denotes the Fourier transform operator. $\odot$ represents the Hadamard product (i.e., component-wise product), Φ^T is the inverse DFT matrix, and s projects a filter onto a small spatial support. By utilizing slack variables $\mathbf{U}_k^x$, $\mathbf{U}_k^y$ and $\mathbf{S}_k^x$, $\mathbf{S}_k^y$, the loss function can be treated as the sum of multiple subproblems and with the addition of equality constraints.

Training Filters

Similar to theoretical CSC methods, we alternatively optimize the convolutional least squares term for the basis function pairs $\mathbf{F}^x$ and $\mathbf{F}^y$ followed by an l_1-regularized least squares term for the corresponding sparse feature maps $\mathbf{Z}^x$ and $\mathbf{Z}^y$. Like the subproblem of solving feature maps, filter pairs can be learned in a similar fashion. With $\hat{\mathbf{Z}}_k^x$, $\hat{\mathbf{Z}}_k^y$ and $\mathbf{W}_k$ fixed, we can update the corresponding filter pairs $\hat{\mathbf{F}}_k^x$, and $\hat{\mathbf{F}}_k^y$ as

$$
\begin{aligned}
\min_{\mathbf{F}^x, \mathbf{F}^y} \frac{1}{2} \left\| \hat{\mathbf{X}} - \sum_{k=1}^{K} \hat{\mathbf{F}}_k^x \odot \hat{\mathbf{Z}}_k^x \right\|_F^2 + \frac{1}{2} \left\| \hat{\mathbf{Y}} - \sum_{k=1}^{K} \hat{\mathbf{F}}_k^y \odot \hat{\mathbf{Z}}_k^y \right\|_F^2 \\
s.t. \ \| \mathbf{f}_k^x \|_2^2 \leq 1, \| \mathbf{f}_k^y \|_2^2 \leq 1 \ \forall k,
\end{aligned}
\tag{4.10}
$$

The optimization with respect to Eq. (4.10) can be solved by a one-by-one update strategy [155] through an augmented Lagrangian method [10].

Algorithm 5: SRCMS

```
Input: Test image $\mathbf{X}^t$, filter pairs $\mathbf{F}^x$ and $\mathbf{F}^y$, mapping $\mathbf{W}$.
Initialize $\mathbf{Z}_0^t$.
Let $\hat{\mathbf{Z}}_0^t \leftarrow \mathbf{W}\mathbf{Z}_0^t$, $\mathbf{Y}_0^t \leftarrow \mathbf{F}^y\mathbf{W}\mathbf{Z}_0^t$.
while not converged do
    Update $\mathbf{Z}_{k+1}^t$ and $\hat{\mathbf{Z}}_{k+1}^t$ by (4.9) with $\mathbf{Y}_k^t$, and $\mathbf{W}$.
    Update $\mathbf{Y}_{k+1}^t \leftarrow \mathbf{W}\hat{\mathbf{Z}}_{k+1}^t$.
end
Synthesize $\mathbf{Y}^t$ by (4.12).
Output: Synthesized image $\mathbf{Y}^t$.
```

Learning Mapping Function

Finally, $\mathbf{W}_k$ can be learned by fixing $\mathbf{F}_k^x$, $\mathbf{F}_k^y$, and $\mathbf{Z}_k^x$, $\mathbf{Z}_k^y$:

$$\begin{aligned}\min_{\mathbf{W}} \sum_{k=1}^{K} \|\mathbf{Z}_k^y - \mathbf{W}_k\mathbf{Z}_k^x\|_F^2 + \left(\frac{\gamma}{\beta}\right) \sum_{k=1}^{K} \|\mathbf{W}_k\|_F^2 \\ + Tr(\sum_{k=1}^{K} \mathbf{Z}_k^y \mathbf{M}(\mathbf{W}_k\mathbf{Z}_k^x)^T),\end{aligned} \tag{4.11}$$

where Eq. (4.11) is a ridge regression problem with a regularization term. We simplify the regularization term $\mathscr{R}(tr) = Tr(\sum_{k=1}^{K} \mathbf{Z}_k^y\mathbf{M}(\mathbf{W}_k\mathbf{Z}_k^x)^T)$ and analytically derive the solution as $\mathbf{W} = (\mathbf{Z}_k^y\mathbf{Z}_k^{xT} - \mathscr{R}(tr))(\mathbf{Z}_k^x\mathbf{Z}_k^{xT} + \frac{\gamma}{\beta}\mathbf{I})^{-1}$, where $\mathbf{I}$ is an identity matrix.

4.3.4 Synthesis

Once the training stage is completed, generating a set of filter pairs $\mathbf{F}^x$, $\mathbf{F}^y$ and the mapping $\mathbf{W}$, for a given test image $\mathbf{X}^t$ in domain $\mathscr{X}$, we can synthesize its desirable HR version of style $\mathscr{Y}$. This is done by computing the sparse feature maps $\mathbf{Z}^t$ of $\mathbf{X}^t$ with respect to a set of filters $\mathbf{F}^x$, and associating $\mathbf{Z}^t$ to the expected feature maps $\hat{\mathbf{Z}}^t$ via $\mathbf{W}$, i.e., $\hat{\mathbf{Z}}^t \approx \mathbf{W}\mathbf{Z}^t$. Therefore, the desirable HR $\mathscr{M}_2$ modality image is then obtained by the sum of K converted sparse feature maps $\hat{\mathbf{Z}}_k^t$ convolved with desired filters $\mathbf{F}_k^y$ (termed (SRCMS) summarized in Algorithm 5):

$$\mathbf{Y}^t = \sum_{k=1}^{K} \mathbf{F}_k^y\mathbf{W}_k\mathbf{Z}_k^t = \sum_{k=1}^{K} \mathbf{F}_k^y\hat{\mathbf{Z}}_k^t. \tag{4.12}$$

4.4 Experimental Results

We conduct the experiments using two datasets, i.e., IXI[3] and NAMIC brain mutlimodality[4] datasets. Following [51, 155, 168], LR counterparts are directly down-sampled from their HR ground truths with rate $1/2$ by bicubic interpolation, boundaries are padded (with eight pixels) to avoid the boundary effect of Fourier domain implementation. The regularization parameters σ, λ, β, and γ are empirically set to be 1, 0.05, 0.1, 0.15, respectively. Optimization variables $\mathbf{F}$, $\mathbf{S}$, $\mathbf{Z}$, and $\mathbf{U}$ are randomly initialized with Gaussian noise considering [10]. Generally, a larger number of filters leads to better results. To balance between computation complexity and result quality, we learn 800 filters following [51]. In our experiments, we perform a more challenging division by applying half of the dataset (processed to be weakly co-registered data) for training while the remaining for testing. To the best of our knowledge, there is no previous work specially designed for SR and cross-modality synthesis simultaneously by learning from the weakly-supervised data. Thus, we extend the range of existing works as the baselines for fair comparison, which can be divided into two categories as follows: (1) brain MRI SR; (2) SR and cross-modality synthesis (one-by-one strategy in comparison models). For the evaluation criteria, we adopt the widely used PSNR and SSIM [159] indices to objectively assess the quality of the synthesized images.

Experimental Data: The IXI dataset consists of 578 $256 \times 256 \times n$ $n = 112 \sim 136$ MR healthy subjects collected at three hospitals with different mechanisms (i.e., Philips 3T system, Philips 1.5T system, and GE 3T system). Here, we utilize 180 Proton Density-weighted (PD-w) MRI subjects for image SR, while applying both PD-w and registered T2-weighted (T2-w) MRI scans of all subjects for major SRCMS. Further, we conduct SRCMS experiments on the processed NAMIC dataset, which consists of 20 $128 \times 128 \times 88$ subjects in both T1-weighted (T1-w) and T2-w modalities. As mentioned, we leave half of the dataset out for cross-validation. We randomly select 30 registered subject pairs for IXI, and 3 registered subject pairs for NAMIC, respectively, from the half of the corresponding dataset for training purposes, and process the reminding training data to be unpaired. Particularly, all the existing methods with respect to cross-modality synthesis in brain imaging request a pre-processing, i.e., skull stripping and/or bias corrections, as done in [127, 153]. We follow such processes and further validate whether pre-processing (especially skull stripping) is always helpful for brain image synthesis.

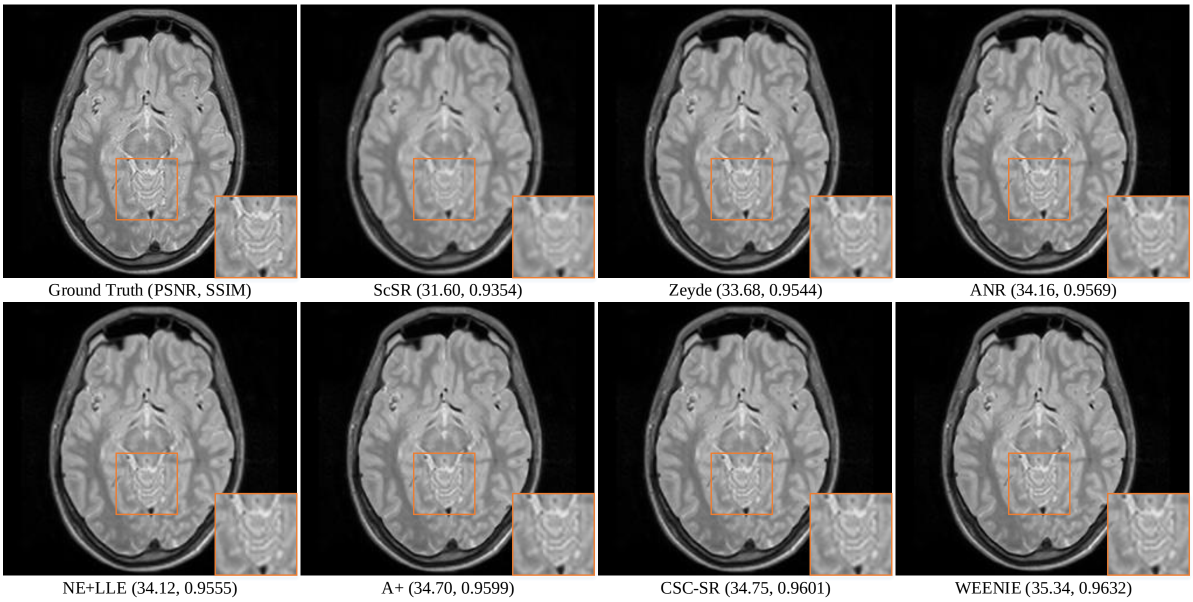

Ground Truth (PSNR, SSIM)
ScSR (31.60, 0.9354)
Zeyde (33.68, 0.9544)
ANR (34.16, 0.9569)
NE+LLE (34.12, 0.9555)
A+ (34.70, 0.9599)
CSC-SR (34.75, 0.9601)
WEENIE (35.34, 0.9632)

Fig. 4.2 Example SR results and corresponding PSNRs, SSIMs (zoom in for details).

4.4.1 Brain MRI Super-Resolution

For the problem of image SR, we focus on the PD-w subjects of the IXI dataset to compare the proposed WEENIE model with several state-of-the-art SR approaches: sparse coding-based SR method (ScSR) [168], anchored neighborhood regression method (ANR) [147], neighbor embedding + locally linear embedding method (NE+LLE) [16], Zeyde's method [176], convolutional sparse coding-based SR method (CSC-SR) [51], and adjusted anchored neighborhood regression method (A+) [148]. We perform image SR with scaling factor 2, and show visual results on an example slice in Fig. 4.2. The quantitative results for different methods are shown in Fig. 4.3, and the average PSNR and SSIM for all 95 test subjects are listed in Table 4.1. The proposed method, in the case of brain image SR, obtains the best PSNR and SSIM values. The improvements show that the MMD regularized joint learning property on CSC has more influence than the classic sparse coding-based methods as well as the state-of-the-arts. It states that using MMD combined with the joint CSC indeed improves the representation power of the learned filter pairs.

4.4.2 Simultaneous Super-Resolution and Cross-Modality Synthesis

To comprehensively test the robustness of the proposed WEENIE method, we perform SRCMS on both datasets involving six groups of experiments: (1) synthesizing SR T2-w image from

[3] http://brain-development.org/ixi-dataset/
[4] http://hdl.handle.net/1926/1687

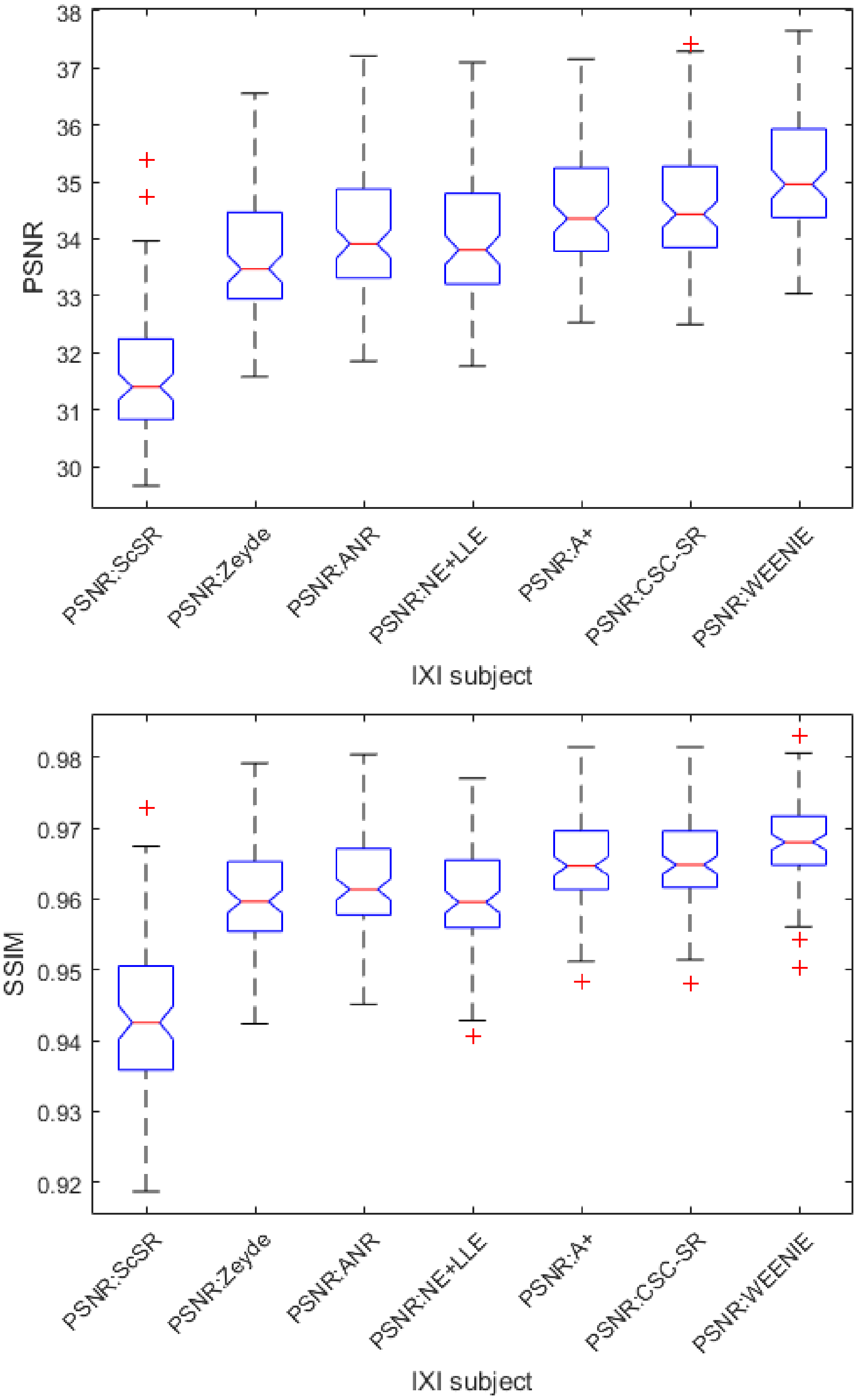


Fig. 4.3 Performance comparisons of different SR approaches.

Metric(avg.)	ScSR [168]	Zeyde [176]	ANR [147]	NE+LLE [16]	A+ [148]	CSC-SR [51]	WEENIE
PSNR(dB)	31.63	33.68	34.09	34.00	34.55	34.60	**35.13**
SSIM	0.9654	0.9623	0.9433	0.9623	0.9591	0.9604	**0.9681**

Table 4.1 Quantitative evaluation (PSNR and SSIM): WEENIE vs. other SR methods on 95 subjects of the IXI dataset.

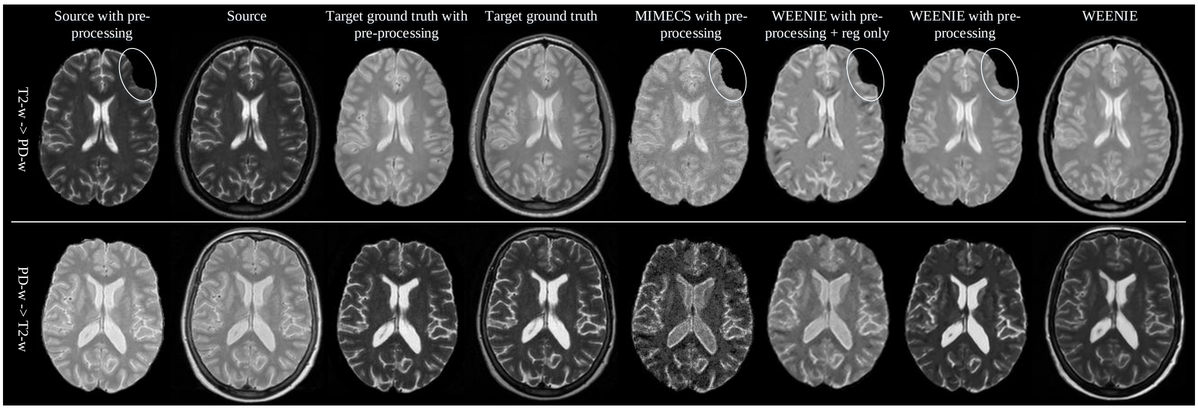

Source with pre-processing
Source
Target ground truth with pre-processing
Target ground truth
MIMECS with pre-processing
WEENIE with pre-processing + reg only
WEENIE with pre-processing
WEENIE
T2-w -> PD-w
PD-w -> T2-w

Fig. 4.4 Visual comparison of synthesized results using different methods.

LR PD-w acquisition and (2) *vice versa*; (3) generating SR T2-w image from LR PD-w input based on pre-processed data (i.e., skull strapping and bias corrections) and (4) *vice versa*; (5) synthesizing SR T1-w image from LR T2-w subject and (6) *vice versa*. The first four sets of experiments are conducted on the IXI dataset while the last two cases are evaluated on the NAMIC dataset. The state-of-the-art synthesis methods include Vemulapalli's supervised approach (V-S) [153], Vemulapalli's unsupervised approach (V-US) [153] and MR image exampled-based contrast synthesis (MIMECS) [127] approach. However, Vemulapalli's methods cannot be applied for our problem, because they only contain the cross-modality synthesis stage used in the NAMIC dataset. Original data (without degradation processing) are used in all Vemulapalli's methods. MIMECS takes image SR into mind and adopts two independent steps (i.e. synthesis+SR) to solve the problem. We compare our results on only using registered image pairs denoted by WEENIE(reg) (that can directly substantiate the benefits of involving unpaired data) and the results using all training images with/without preprocessing for the proposed method against MIMECS, V-US and V-S in above six cases and demonstrate examples in Fig. 4.4 for visual inspection. The advantage of our method over the MIMECS shows, e.g., in white matter structures, as well as in the overall intensity profile. We show the quantitative results in Fig. 4.5, and summarize the averaged values in Table 4.2 and Table 4.3, respectively. It can be seen that the performance of our algorithm is consistent across two whole datasets, reaching the best PSNR and SSIM for almost all subjects.

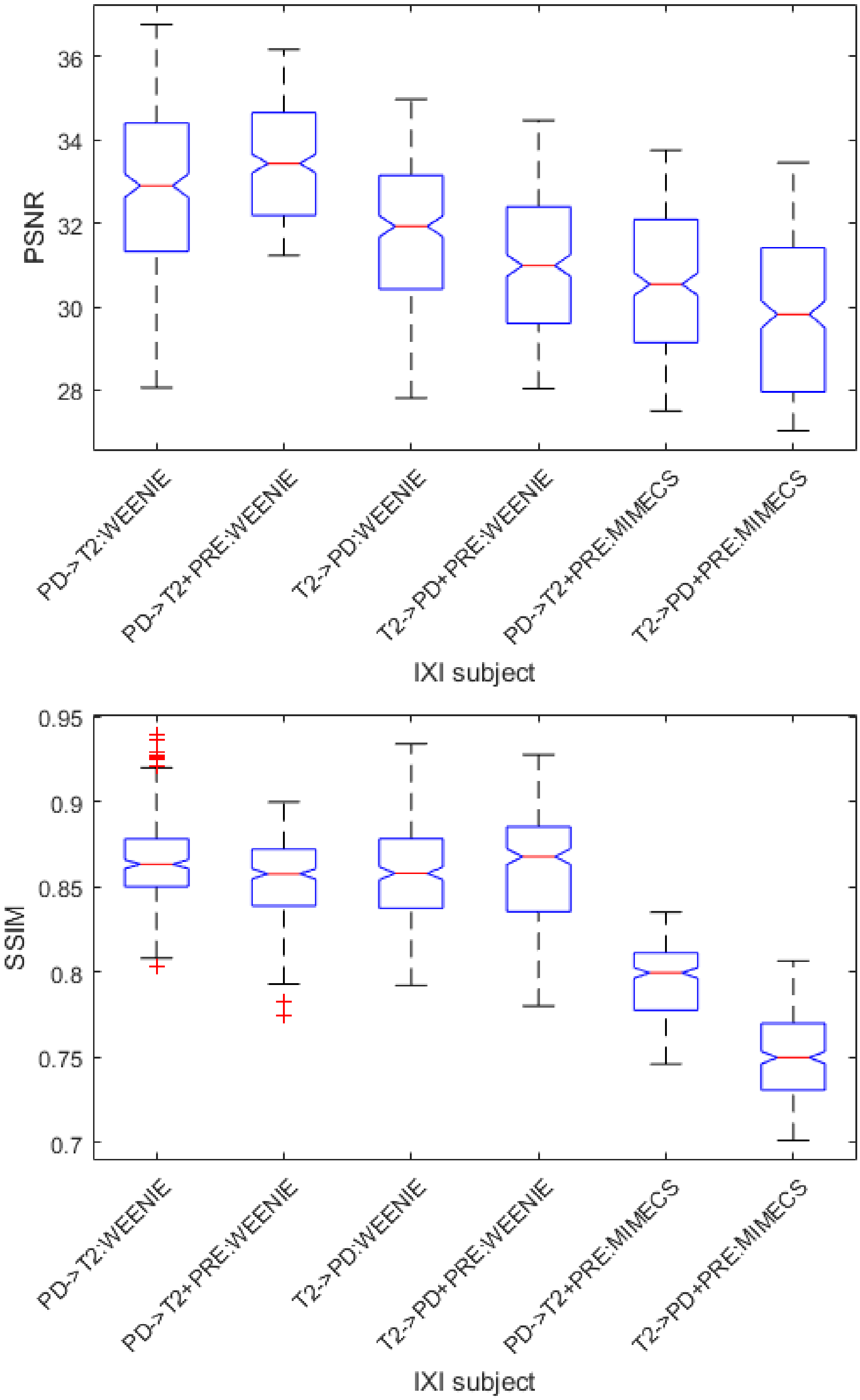


Fig. 4.5 SRCMS results: WEENIE vs. MIMECS on IXI dataset.

Metric(avg.)	IXI							
	PD→T2	T2→PD	PD→T2+PRE			T2→PD+PRE		
	WEENIE		MIMECS	WEENIE(reg)	WEENIE	MIMECS	WEENIE(reg)	WEENIE
PSNR(dB)	37.77	31.77	30.60	30.93	**33.43**	29.85	30.29	**31.00**
SSIM	0.8634	0.8575	0.7944	0.8004	**0.8552**	0.7503	0.7612	**0.8595**

Table 4.2 Quantitative evaluation (PSNR and SSIM): WEENIE vs. other synthesis methods on the IXI dataset.

Metric(avg.)	NAMIC							
	T1→T2				T2→T1			
	MIMECS	Ve-US	Ve-S	WEENIE	MIMECS	Ve-US	Ve-S	WEENIE
PSNR(dB)	24.36	26.51	27.14	**27.30**	27.26	27.81	29.04	**30.35**
SSIM	0.8771	0.8874	0.8934	**0.8983**	0.9166	0.9130	0.9173	**0.9270**

Table 4.3 Quantitative evaluation (PSNR and SSIM): WEENIE vs. other synthesis methods on the NAMIC dataset.

4.5 Conclusions

In this paper, we proposed a novel weakly-supervised joint convolutional sparse coding (WEENIE) method for simultaneous super-resolution and cross-modality synthesis (SRCMS) in 3D MRI. Different from conventional joint learning approaches based on sparse representation in supervised setting, WEENIE only requires a small set of registered image pairs and automatically aligns the correspondence for auxiliary unpaired images to span the diversities of the original learning system. By means of the designed hetero-domain alignment term, a set of filter pairs and the mapping function were jointly optimized in a common feature space. Furthermore, we integrated our model with a divergence minimization term to enhance robustness. With the benefit of consistency prior, WEENIE directly employs the whole image, which naturally captures the correlation between local neighborhoods. As a result, the proposed method can be applied to both brain image SR and SRCMS problems. Extensive results showed that WEENIE can achieve superior performance against state-of-the-art methods.

Chapter 5

Geometry Constrained Dictionary Learning

5.1 Geometry Regularized Joint Dictionary Learning

Multi-sequence MRI protocols are used in comprehensive examinations of various pathologies in both clinical diagnosis and medical research. Various MRI techniques provide complementary information about living tissue. However, a comprehensive examination covering all modalities is rarely achieved due to considerations of cost, patient comfort, and scanner time availability. This may lead to incomplete records owing to image artifacts or corrupted or lost data. In this chapter, we explore the problem of synthesizing images for one MRI modality from an image of another MRI modality of the same subject using a novel geometry regularized joint dictionary learning framework for non-local patch reconstruction. Firstly, we learn a cross-modality joint dictionary from a multi-modality image database. Training image pairs are first co-registered. A cross-modality dictionary pair is then jointly learned by minimizing the cross-modality divergence via a Maximum Mean Discrepancy term in the objective function of the learning scheme. This guarantees that the distribution of both image modalities is taken jointly into account when building the resulting sparse representation. In addition, in order to preserve intrinsic geometrical structure of the synthesized image patches, we further introduced a graph Laplacian regularization term into the objective function. Finally, we present a patch-based non-local reconstruction scheme, providing further fidelity of the synthesized images. Experimental results demonstrate that our method achieves significant performance gains over previously published techniques.

5.1.1 Introduction

Magnetic Resonance Imaging (MRI) is a versatile and noninvasive imaging technique extensively used in neuroimaging studies. MRI comes in many different flavors (viz. MRI sequences, or henceforth also referred as MRI modalities[1]), each providing diverse and complementary image contrast mechanisms unraveling structural and functional information about brain tissue. Multi-modality MRI are nowadays very common in many pharmaceutical clinical trials, in research studies of neurosciences, or in population imaging cohorts targeted to understand neurodegeneration and cognitive decline. The acquisitions of a full battery of all these MR images can face constraints associated with their cost, limited availability of scanning time, patient com-fort or safety considerations. Moreover, in large scale studies it is not uncommon to face incomplete datasets due to the presence of imaging artifacts, acquisition errors or corrupted data. While many such studies use imputation techniques to compensate for these latter issues, this is usually only at the level of the derived imaging biomarkers and not of the data itself. Finally, in longitudinal imaging studies where image data is collected over several years, evolution of imaging technology may lead to the appearance of new MRI sequences that are added to an existing imaging protocol at some point in time but for which are not available as part of the imaging battery acquired at earlier time points. In these and other applications, it would be desirable to have a methodology that is able to synthesize the unavailable data from the available MRI studies. The assumption here is that the synthesis ability comes from the cross-modality correspondences of sparse codes obtained during training, and can be used to encoding missing MRI. The degree to which this hypothesis is valid will have to be scrutinized in each application but is worth exploring.

To cope with this problem, several methods were proposed through either trans-forming MRI intensities or reconstructing tissue contrasts to obtain the missing MRI data. Histogram matching is the most common approach within this group. Although this technique is widely used in neuroimaging, it has been pointed out its inefficacy for multi-modality image synthesis due to the lack of specificity for certain ratios of tissue types [127]. On the other hand, techniques based on sparse representations have been presented, which separately learn two corresponding dictionaries from co-registered image pairs and synthesize a desired MRI modality data from the patches of the available MRI modality [127]. These approaches, however, boil down to an example-based synthesis strategy, which does not fully exploit the available training data to its fullest. In contrast, here, we establish fundamental connections with trans-fer learning (*a.k.a.* domain adaptation) used in many fields, *e.g.* [132, 118]. Such methods can

[1] Here, we use the word modality in the sense of a specific kind of MRI sequence. Note that the proposed technique would equally be applicable when the protocol involves different imaging modalities in a more classical sense (*e.g.* MRI, CT, US, SPECT, and PET).

successfully solve the above problem by learning a paired dictionary from both modalities while assuming each co-registered image pair with a nearly identical distribution [127]. However, this assumption cannot be fully satisfied in practice since cross-modality data may have very different feature distributions in different spaces.

In this chapter, we propose a novel geometry regularized joint dictionary learning method for synthesizing any unavailable MRI from available MRI data. This chapter offers the following three contributions:

1. We address cross-modality MRI synthesis by jointly learning a cross-modality dictionary that penalizes differences in the statistical distribution of the sparse codes in both domains rather than directly imposing the same code to both domains as done before. This is achieved by incorporating a new term in the computation of the joint sparse codes using the Maximum Mean Discrepancy measure;

2. We exploit the geometrical information underlying the input data and incorporate this new term into the cross-modality joint dictionary learning optimization;

3. A non-local reconstruction framework that provides a more expressive and compact patch representation is adopted to synthesize the corresponding patch from a different MRI protocol.

To the best of our knowledge, this is the first time that joint dictionary learning is computed by minimizing the discrepancy between the statistical distributions of the sparse codes of the involved MRI modalities while preserving the intrinsic geometrical structure of the image. In the remainder of this chapter, we first define the cross-modality synthesis problem, and then introduce our proposed method in Section 2. The experimental results are demonstrated in Section 3. Finally, we discuss the results and conclude the chapter in Section 4.

5.1.2 Method

In this section, we propose cross-modality image synthesis via geometry regularized joint dictionary learning for effectively minimizing the cross-modality discrepancy. This consists of an extension of the conventional dictionary learning by jointly learning from the data of two modalities at the same time while minimizing the sparse codes divergence between the different modalities.

Problem Definition

Let $\mathbf{L}^{M_k} = \left\{ \mathbf{I}_i^{M_k} \right\}_{i=1}^{m}$ be a library of m subjects imaged with k modalities each (k = 1 or 2), with $\mathbf{I}_i$ being the training image of the i-th sample. Each pair of images in both libraries, *i.e.* $\left\{\mathbf{I}_i^{M_1}, \mathbf{I}_i^{M_2}\right\}$ is assumed co-registered. Further, images are treated as the combination of many patches and denoted as $\mathbf{X}^{M_k} = \left\{\mathbf{x}_i^{M_k}\right\}_{i=1}^{n} \in \mathbb{R}^{s \times n}$, where s is the size of a vectorized patch, and n represents the number of training patches for both modalities. We denote the test image in the same way by a matrix $\mathbf{Y} = \{\mathbf{y}_l\}_{l=1}^{c} \in \mathbb{R}^{s \times c}$, where c is the number of patches in the test image. All of the elements in $\mathbf{Y}$ are considered with either modality $M1$ or modality $M2$.

Problem: Given $\mathbf{X}^{M_1}$ and $\mathbf{X}^{M_2}$, our goal is to learn a pair of dictionaries $\left\{\mathbf{D}^{M_1}, \mathbf{D}^{M_1}\right\}$ and the unified sparse codes $\mathbf{A}^u$ minimizing the cross-modality discrepancy of $\mathbf{A}^{M_1}$ and $\mathbf{A}^{M_2}$, where $\mathbf{A}^{M_k}$ is the sparse codes matrix of $\mathbf{X}^{M_k}$.

Dictionary Learning

Let $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^{n} \in \mathbb{R}^{s \times n}$ be a training data matrix with n input items sampled in the s-dimensional space, $\mathbf{D} = \{\mathbf{d}_i\}_{i=1}^{K} \in \mathbb{R}^{s \times K}$ be a projection dictionary with K atoms, where $K > s$ to make the dictionary overcomplete. Learning D from a sparse representation of X can be formulated as:

$$\min_{\mathbf{D},\mathbf{A}} \|\mathbf{X} - \mathbf{D}\mathbf{A}\|_F^2 + \lambda \|\mathbf{A}\|_0, \tag{5.1}$$

where $\mathbf{A} = \{\alpha_i\}_{i=1}^{n} \in \mathbb{R}^{K \times n}$ is a set of n K-dimensional sparse codes of $\mathbf{X}$, $\|\cdot\|_F$ is the Frobenius norm, $\|\cdot\|_0$ is l_0-norm, which fixes the number of non-zero elements of $\mathbf{A}$, and λ denotes a regularization parameter to trade off the sparsity and the reconstruction error. As shown in [26], the minimization problem as stated in 5.1 is an NP-hard problem under the l_0-norm constraint. An alternative solution is to relax the l_0-norm constraint with the l_1-norm constraint to obtain a near-optimum result [20].

Geometry Regularized Joint Dictionary Learning

Following the dictionary learning procedure described in Sec. 5.1.2, instead of transferring the estimated sparse codes from the first domain to the other [127, 168], we can learn the dictionaries of both domains independently

$$\begin{aligned} &\min_{\mathbf{D}^{M_1},\mathbf{A}^{M_1}} \left\|\mathbf{X}^{M_1} - \mathbf{D}^{M_1}\mathbf{A}^{M_1}\right\|_F^2 + \lambda_1 \left\|\mathbf{A}^{M_1}\right\|_1, \\ &\min_{\mathbf{D}^{M_2},\mathbf{A}^{M_2}} \left\|\mathbf{X}^{M_2} - \mathbf{D}^{M_2}\mathbf{A}^{M_2}\right\|_F^2 + \lambda_2 \left\|\mathbf{A}^{M_2}\right\|_1. \end{aligned} \tag{5.2}$$

However, such a strategy is time-consuming and results in two sets of independent sparse codes that do not necessarily satisfy the assumption of high-correlation between both modalities to reconstruct M_2-like images from M_1-like ones. To solve a similar problem, Yang *et al.* [168]

proposed an image super-resolution approach that uses coupled dictionary learning. Their method maps image pairs (*e.g.* low and high resolution or, here, two different modalities) into a common space, which enforces the sparse codes of paired data possess the same values. Instead of directly imposing the same sparse codes across each pair, our work allows the codes to be different for each modality, and fosters the most similar distributions across them. This is done by measuring the distribution divergence for the co-registered image pairs over the empirical Maximum Mean Discrepancy (MMD), which is then minimized and incorporated into the dictionary learning problem.

Maximum Mean Discrepancy Regularization

We seek that the probability distributions of the codes associated to cross-modality patch pairs is identical when computing the optimal sparse representation. To this effect, the MMD [48] is used. The MMD is a nonparametric statistic utilized to assess whether two samples are drawn from the same distribution. In our case, the two samples correspond to the sparse codes of the training set for the two modalities involved. The MMD is calculated as the largest difference in the expected mean value of the K-dimensional codes for both modalities. To compute the MMD, we follow [48, 98, 139] to estimate the largest difference in expectations over functions in the unit ball of a reproducing kernel Hilbert space:

$$\begin{aligned} \text{MMD} &= \left\| \frac{1}{n}\sum_{i=1}^{n} \alpha_i - \frac{1}{n}\sum_{j=n+1}^{2n} \alpha_j \right\|_{\mathscr{H}}^2 \\ &= Tr(\mathbf{A}^u \mathbf{M} \mathbf{A}^{u^T}), \end{aligned} \tag{5.3}$$

where $\mathbf{A}^u$ represents the unified sparse codes, $\mathbf{A}^{u^T}$ is the transposed matrix of $\mathbf{A}^u$, and $\mathbf{M}$ denotes the MMD matrix defined as:

$$M_{i,j} = \begin{cases} \frac{1}{n^2}, & \mathbf{x}_i, \mathbf{x}_j \in \mathbf{X}^{M_1} \ or \ \mathbf{x}_i, \mathbf{x}_j \in \mathbf{X}^{M_2} \\ -\frac{1}{n^2}, & otherwise. \end{cases} \tag{5.4}$$

Graph Laplacian Regularization

During dictionary learning, high-level patch semantics are captured in each dictionary atom. However, this process fails to introduce any prior knowledge on the geometrical structure within patches. Instead, by introducing a Graph Laplacian (GL) term [184], we can preserve the local manifold structure of the sparse graph and better capture the intrinsic geometrical properties of the entire data space. Given $\{\mathbf{X}^{M_1}, \mathbf{X}^{M2}\} \in \mathbb{R}^{s \times 2n}$, a q-nearest neighbor graph

$\mathscr{G}$ with 2n vertices can be constructed. The weight matrix of $\mathscr{G}$ is $\mathbf{W} \in \mathbb{R}^{2n \times 2n}$, defined as the matrix with elements $\mathbf{w}_{i,j}$= 1 if and only if for any two data points $\mathbf{x}_i$, $\mathbf{x}_j$, $\mathbf{x}_i$ is among the q-nearest neighbors of $\mathbf{x}_j$ or *vice versa* ($\mathbf{w}_{i,j}$ = 0, otherwise). Let $\Phi = diag(\phi_1, \phi_2, ..., \phi_{2n})$ be the diagonal degree matrix with elements $\phi_i = \sum_{j=1}^{2n} \mathbf{w}_{i,j}$. The GL term, incorporated into the sparse representation as a regularization criterion [184], imposes that the obtained sparse codes vary smoothly along the geodesics of the manifold that is captured by the graph. The GL matrix is then defined as $\mathbf{G} = \Phi - \mathbf{W}$. In order to preserve the geometrical structure in $\mathscr{G}$, we map $\mathscr{G}$ to the unified coefficients $\mathbf{A}^u$ by:

$$\begin{aligned} &\frac{1}{2}\sum_{i,j=1}^{2n} \|\alpha_i - \alpha_j\|_2^2 \mathbf{w}_{i,j} \\ &= \sum_{i=1}^{2n} \alpha_i \alpha_i^T \phi_{ii} - \sum_{i,j=1}^{2n} \alpha_j \alpha_i^T \mathbf{w}_{i,j} \\ &= Tr(\mathbf{A}^u \mathbf{G} \mathbf{A}^{u^T}). \end{aligned} \tag{5.5}$$

Objective Function

To maximize the correlation between patch pairs in both modalities, we map them into a common higher-dimensional space that meets two complementary objectives to those of Eq. (5.2), *viz.* the MMD and GL terms. Therefore, our geometry regularized joint dictionary learning objective function becomes:

$$\begin{aligned} \min_{\mathbf{D}^{M_1}, \mathbf{D}^{M_2}, \mathbf{A}^u} \frac{1}{2}(\|\mathbf{X}^{M_1} - \mathbf{D}^{M_1}\mathbf{A}^u\|_F^2 + \|\mathbf{X}^{M_2} - \mathbf{D}^{M_2}\mathbf{A}^u\|_F^2) \\ + Tr(\mathbf{A}^u(\gamma\mathbf{M} + \delta\mathbf{G})\mathbf{A}^{u^T}) + \lambda \|\mathbf{A}^u\|_1 . \end{aligned} \tag{5.6}$$

where γ and δ are the regularization parameters for trading off the effect of the MMD and GL terms, respectively.

Image Synthesis via Nonlocal Reconstruction

Once the cross-modality dictionary pairs have been computed by solving Eq. (5.6), we seek to reconstruct a test image $Y \in \mathbb{R}^{s \times c}$ by, first, sparsely representing Y with respect to $\mathbf{D}^{M_1} \in \mathbb{R}^{s \times K}$ by solving Eq. (5.1) with l_1-norm as:

$$\mathbf{A}^t = \arg\min_{\mathbf{A}^t} \|\mathbf{Y} - \mathbf{D}^{M_1}\mathbf{A}^t\|_F^2 + \lambda \|\mathbf{A}^t\|_1 , \tag{5.7}$$

where $\mathbf{A}^t \in \mathbb{R}^{K \times c}$ denotes the sparse codes of $\mathbf{Y}$. The estimated coefficients can be directly used (or "transferred") to synthesize the image $\hat{\mathbf{Y}}$ of our target modality M_2 by a linear combination of elements in the dictionary $\mathbf{D}^{M_2}$, namely, $\hat{\mathbf{Y}} = \mathbf{D}^{M_2}\mathbf{A}^t$.

To achieve richer synthesis ability, in this chapter, we improve the sparse representation performance through an optimized nonlocal reconstruction model. To faithfully synthesize the desired image, we enforce the sparse coefficients $\mathbf{A}^t$ as close as possible to the target codes. That is, by groups of similar patches being encoded onto subsets of the dictionary that are similar, the estimated sparse codes vary smoothly as the patches themselves vary. This makes the whole reconstruction scheme more robust to the influence of patch noise and more accurate. To this end, we adopt the representative non-local means [12] in the sparse representation model by modifying Eq. (5.7) as

$$\hat{\mathbf{A}}^t = \arg\min_{\beta^t} \left\|\mathbf{Y} - \mathbf{D}^{M_1}\beta^t\right\|_F^2 + \lambda \left\|\beta^t\right\|_1. \tag{5.8}$$

where $\beta^t = \sum\sum_{p\in\Omega_i} \mu_{i,p}\alpha^t_{i,p}$ and α^t_i indicates the sparse codes of $\mathbf{y}_i$. For each $\mathbf{y}_i$, we express its similar patch set as Ω_i, and define p as a random element within Ω_i. Also, we define $\mu_{i,p}$ as the weight for computing the level of similarity between $\mathbf{y}_i$ and $\mathbf{y}_q$, where $\mu_{i,p} = \frac{1}{\mathscr{C}} exp\left\{-\frac{\|\mathbf{y}_i - \mathbf{y}_{i,p}\|_2^2}{h^2}\right\}$ with $\mathscr{C}$ being the normalization constant and h being a scalar (note that $\mu_{i,p}$ satisfies $0 \leq \mu_{i,p} \leq 1$ and $\sum_{p\in\Omega_i} \mu_{i,p} = 1$). Then, we can update the synthesized image via $\hat{\mathbf{Y}} = \mathbf{D}^{M_2}\hat{\mathbf{A}}^t$.

5.1.3 Experiments

In this section, we show extensive experiments for the MRI cross-modality synthesis problem to verify the effectiveness of our proposed method.

Experiment Setup

We evaluated our method in two different scenarios. Firstly, we used the IXI dataset [126] for synthesizing the T2-w image considering the proton density (PD) acquisition from the same subject. We randomly selected 12 subjects from IXI containing both T2-w and corresponding PD-w images. We trained the dictionaries from 5 subjects including both modalities, and the other 7 subjects were used for testing. In the second experiment, we considered the generation of magnetization-prepared rapid gradient-echo (MPRAGE) images based on spoiled gradient recalled (SPGR) acquisitions, allowing us to compare our method with an existing approach [127]. In each experiment, for each co-registered image pair in the training set, we randomly selected 100,000 patch pairs of 5×5×5 voxels size to train our dictionaries. We also took the factor of dictionary size and sparsity into consideration, and fixed the dictionary size as 1024 and $\lambda = 0.15$ based on our experiments trading off cost and synthesis quality. For other parameters, we used the following settings according to our extensive experiments: $q = 5$, $\gamma = 105$, $\delta = 1$, and the searching window for nonlocal reconstruction equals 10. Finally, we adopted Root Mean Square Error (RMSE), Peak Signal to Noise Ratio (PSNR) in decibels (dB), and Structural Similarity Index (SSIM) as evaluation metrics.

Compared Methods

To show the performance of our approach, we compared our results of the following state-of-the-art methods:

1. Joint Dictionary Learning (JDL);
2. MRI example-based contrast synthesis (MIMECS) [127];
3. Geometry Regularized Joint Dictionary Learning (GRiDLE) with only MMD term.
4. The proposed GRiDLE.

Note that JDL is a special case of GRiDLE with $\gamma = \delta = 0$, and GRiDLE with only MMD term is another special case with $\delta = 0$.

Experimental Results

Table 5.1 shows the error measures of the synthesized T2 images using JDL, GRiDLE (δ = 0) and GRiDLE. We did not compare our GRiDLE with MIMECS in this case, because there is no available dictionary within this algorithm to generate arbitrary results. We can see that the proposed method outperforms the other two, obtaining the lowest RMSEs and the highest PSNRs and SSIMs for all 7 subjects. In the second example we compared the performance of the proposed method with that of the state-of-the-art MIMECS. The clear advantage of our approach over the MIMECS and JDL is shown in Figure 5.1, which can be seen in overall

Table 5.1 Performance measures of the synthetic images using JDL, GRiDLE (δ = 0), and GRiDLE.

	RMSE			PSNR (dB)			SSIM		
	JDL	GRiDLE ($\delta = 0$)	GRiDLE	JDL	GRiDLE ($\delta = 0$)	GRiDLE	JDL	GRiDLE ($\delta = 0$)	GRiDLE
Sub.1	9.43	8.53	**8.29**	36.72	39.93	**41.73**	0.9025	0.9069	**0.9075**
Sub.2	9.42	8.53	**8.27**	37.15	39.92	**42.05**	0.9021	0.9054	**0.9062**
Sub.3	10.42	9.73	**9.49**	39.35	38.23	**40.35**	0.8997	0.9018	**0.9029**
Sub.4	10.53	9.26	**9.01**	36.17	37.61	**41.34**	0.8669	0.8999	**0.9016**
Sub.5	12.03	11.07	**10.94**	34.12	36.01	**39.17**	0.8990	0.8962	**0.8970**
Sub.6	10.21	9.30	**9.06**	36.73	38.66	**41.02**	0.9002	0.9049	**0.9062**
Sub.7	10.98	9.87	**9.63**	36.18	38.18	**41.01**	0.8964	0.9028	**0.9034**
Avg.	10.43	9.47	**9.24**	36.63	38.36	**40.95**	0.8953	0.9026	**0.9035**

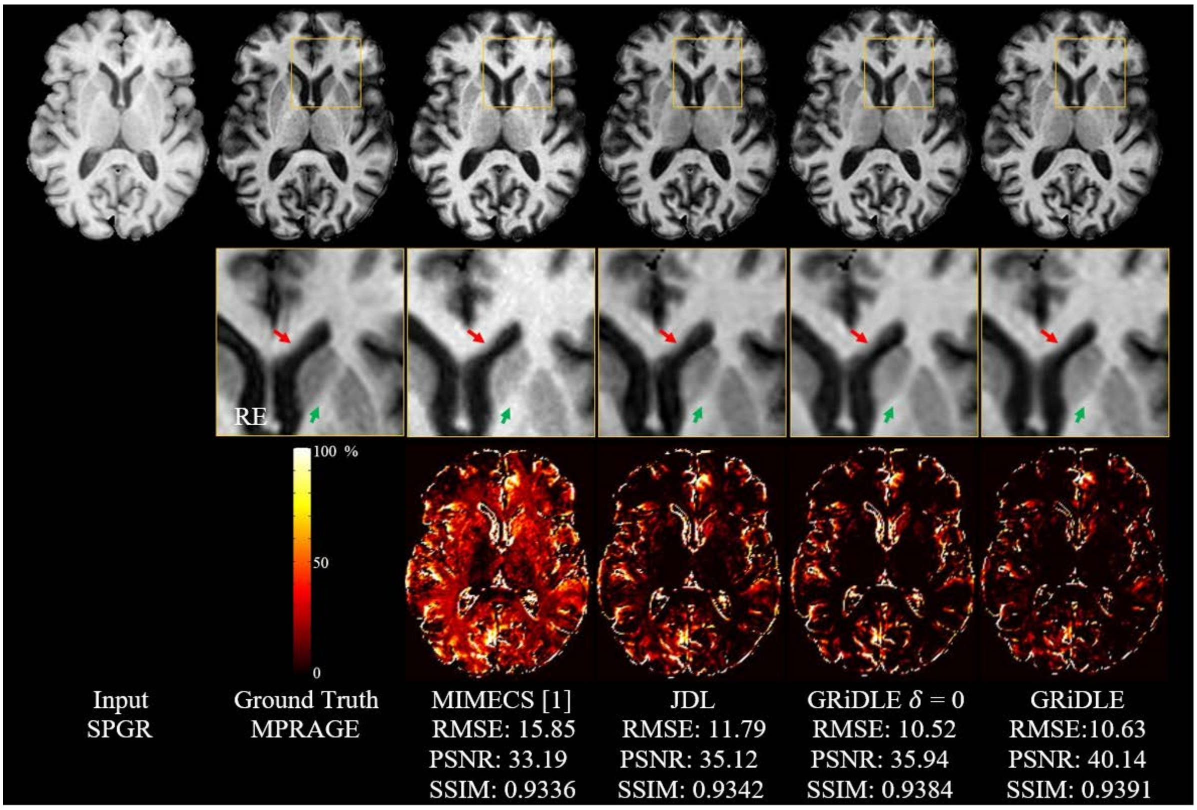


Fig. 5.1 Comparison of the synthesized results with ground truth.

Table 5.2 Comparison of methods used for synthesizing MPRAGE based on SPGR.

	MIMECS [127]	JDL	GRiDLE ($\delta = 0$)	GRiDLE
RMSE	14.55	12.58	11.03	10.89
PSNR (dB)	32.76	34.51	35.52	39.35
SSIM	0.9303	0.9368	0.9403	0.9500

tissue contrast, as well as in the lowest voxelwise RE. Table 5.2 compares the average error measures of all the methods for MPRAGE synthesizing from SPGR images. As shown, the proposed method achieves the best results.

5.1.4 Summary

We proposed a novel geometry regularized joint dictionary learning (GRiDLE) approach for MRI cross-modality synthesis. The distribution divergence is effectively reduced by including the MMD term for both modalities and a mapping function in the sparse domain. The learned dictionary pair can not only minimize the distance between each coupled coefficients but also preserve the geometrical structure in the data while spanning both spaces for stable mapping of image details. Extensive experiments have demonstrated that GRiDLE can achieve superior performance over the state-of-the-art methods. Future work will focus on the simultaneous generation of multimodality images.

5.2 Weakly-Coupled and Geometry Co-Regularized Joint Dictionary Learning

In this work, we propose a Weakly-coupled And Geometry co-regularized (WAG) joint dictionary learning method to address the problem of cross-modality synthesis while considering the fact that collecting large amounts of training data is often impractical. Our learning stage requires only a few registered multi-modality image pairs as training data. To employ both paired images and a large set of unpaired data, a cross-modality image matching criterion is proposed. We then propose a unified model by integrating such a criterion into the joint dictionary learning and the observed common feature space for associating cross-modality data for the purpose of synthesis. Furthermore, two regularization terms are added to construct robust sparse representations. Our experimental results demonstrate superior performance of the proposed model over state-of-the-art methods.

5.2.1 Introduction

Magnetic Resonance Imaging (MRI) is a versatile and noninvasive imaging technique extensively used in neuroimaging studies. MRI comes in several modalities, for example, Proton Density (PD)-weighted images distinguish between fluid and fat, whereas T1-weighted scans have good tissue contrast between gray matter and white matter. Each modality offers diverse and complementary image contrast mechanisms unraveling structural and functional information about brain tissue. Due to variations in the brain images across modalities, multi-modality MRI is preferred in many pharmaceutical clinical trials, in research studies of neurosciences, or in population imaging cohorts targeting to understand neurodegeneration and cognitive decline. However, the acquisitions of a full battery of all these MR images can face constraints associated with their cost, limited availability of scanning time, patient comfort or safety considerations. In large scale studies, it is not uncommon to face incomplete datasets since the presence of imaging artifacts, acquisition errors or corrupted data. While various post-processing solutions such as image imputation [152] and histogram matching [100, 112] have been proposed to compensate for these latter issues, this is usually only at the level of derived imaging biomarkers but not of the data itself [62, 127]. Finally, in longitudinal imaging studies where images are collected over several years, evolution of imaging technology may lead to the appearance of new MRI sequences added to an existing imaging protocol in time, which were not available as part of the imaging battery acquired at earlier time points. In these and other applications, it would be desirable to have a cross-modality image synthesis method that can generate the target modality images from the source modality scans. The ability to synthesize different modalities of the same anatomy can benefit various practical image analysis tasks including multi-modal registration [15, 160], segmentation [68], and atlas construction [25, 115].

In the last few years, cross-modality image synthesis has attracted the attention of the medical image computing community. Most techniques assume such mapping exists between source and target imaging modalities. The problem is then formulated as that of learning the most efficient mapping representation. To synthesize the target from a source modality, some methods have been proposed that construct a dictionary from patches extracted from a single image or from image pairs [59, 72, 95, 127] or that learn the mapping from a large set of training image pairs [14, 150, 153, 172]. Although these approaches have shown great promise, they are supervised and require labeled data sets.

In this work, instead, we propose a single-image cross-modality synthesis method with an application to T1-w, T2-w and PD-w brain MRI that utilizes a few registered multi-modality image pairs[2] while employing a larger set of unpaired data for synthesizing the target image modality from an available source image modality. Our method extracts the common latent features that map different image features of the underlying tissues, preserves global statistical image properties across modalities, and simultaneously, refines extracted features to preserve the local geometrical structure in each modality. In addition, the proposed approach requires only a few registered image pairs to find the mapping between the appearances in different image modalities and employs auxiliary unpaired training images to further exploit the modality-specific geometric structure and obtain a robust sparse representation. To complement the unpaired data with the original training pairs, manifold ranking-based cross-modality image matching is employed as a criterion to pick up features of the target domain from most similar subjects in the source domain. The mapping between multi-modality data can be complex and highly nonlinear. To provide the needed flexibility to map image structures in different modalities, we determine a common feature space by an association function that describes and relates cross-modality data. We call the proposed method <u>W</u>eakly-coupled <u>A</u>nd <u>G</u>eometry (WAG) co-regularized joint dictionary learning, and perform extensive experiments to verify its performance.

The contributions of this work are threefold:

[2] Paired data requires data in both source and target domains from the same subject and registered with each other. Unpaired data means data in the source and target domains are from different subjects without registration.

1. WAG is a unified model, which learns a pair of (not coupled) dictionaries with a common feature space for medical imaging cross-modality synthesis. WAG automatically trains dictionary pairs and computes an association function between source and target modality data based on only a few registered image pairs;
2. To enrich this model, we design a cross-modality image matching criterion that acts collaboratively with a larger set of unpaired images. This allows identifying correspondences across source and target domains that are invariant to pose transformations;
3. We deal with the considerable difference in data distributions from different modalities (i.e. T1-w, T2-w and PD-w MRI) by simultaneously minimizing the distribution discrepancy of similar instances and preserving geometric structures in each domain.

A preliminary version [62] of this work was presented earlier at the SASHIMI Workshop in MICCAI 2016 (www.cistib.org/sashimi2016). This work adds to the seminal version in significant ways. First, we improve the synthesis model by introducing a cross-modality image-matching criterion to connect and integrate all information from both registered and unregistered data in different modalities to describe the diversity of human brain imaging. Second, we extend the single geometry regularization by preserving modality-specific local geometric properties to penalize undesired loss of information. Third, we consider the flexibility of domain-specific information and construct a common feature space by a mapping function that describes and associates cross-modality data. Fourth, we also extend the original experiments from only comparing with one baseline method to several recently published approaches involving both supervised and unsupervised settings. The proposed method demonstrates state-of-the-art synthesis results using two evaluation metrics in all of our experiments.

5.2.2 Method

In this section, we first formulate the problem formally. Then, we propose a general framework for cross-modality image synthesis. Our approach extends the conventional dictionary learning approach by jointly learning a pair of dictionaries from the constructed common feature space that describes and associates cross-modality data. We also consider the minimization of the distribution divergence between both modalities while preserving modality-specific local geometric properties that penalize undesired loss of information. Finally, we utilize unpaired images in both domains as auxiliary training data that enhances the supervised learning process. This additional unsupervised step collaborates with and complements the registered training image pairs. An overview of our proposed method is depicted in Fig. 5.2.

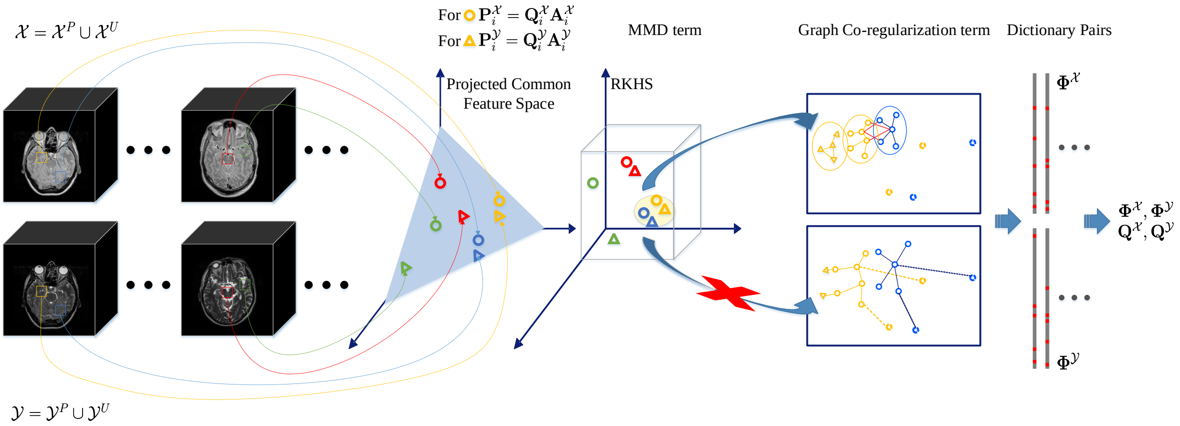

$\mathcal{X} = \mathcal{X}^P \cup \mathcal{X}^U$
For ○ $\mathbf{P}_i^{\mathcal{X}} = \mathbf{Q}_i^{\mathcal{X}}\mathbf{A}_i^{\mathcal{X}}$
For △ $\mathbf{P}_i^{\mathcal{Y}} = \mathbf{Q}_i^{\mathcal{Y}}\mathbf{A}_i^{\mathcal{Y}}$
MMD term
Graph Co-regularization term
Dictionary Pairs
Projected Common Feature Space
RKHS
$\Phi^{\mathcal{X}}$
$\Phi^{\mathcal{X}}, \Phi^{\mathcal{Y}}$
$\mathbf{Q}^{\mathcal{X}}, \mathbf{Q}^{\mathcal{Y}}$
$\Phi^{\mathcal{Y}}$
$\mathcal{Y} = \mathcal{Y}^P \cup \mathcal{Y}^U$

Fig. 5.2 Overview of the proposed cross-modality image synthesis method. Squares within the 3D images represent the extracted 3D patches with size of 5×5×5. The first step is to project the extracted paired/unpaired patches into a common feature space denoted by circles for source modality data and triangles for target modality data respectively. Then, we measure the divergence of the distribution of the maximum mean discrepancy (MMD) over all matched pairs from the first step to seek the intrinsic pairs in the reproducing kernel Hilbert space (RKHS). After that, to better preserve the modality-specific information, we adopt the Laplacian eigenmaps to capture the local geometric structure for each domain and denoted by geometric co-regularization term. Finally, the expected dictionary pairs can be trained based on the processed features.

Problem Definition

Let $\mathscr{X} = \{\mathbf{X}_1, ..., \mathbf{X}_S\}$ be the source modality images of S subjects using modality $\mathscr{M}_1$, and $\mathscr{Y} = \{\mathbf{Y}_1, ..., \mathbf{Y}_T\}$ be the target modality images of T subjects imaged using modality $\mathscr{M}_2$. Therefore, $\mathbf{X}_i$ and $\mathbf{Y}_i$ represent the i-*th* subject-specific images for each modality, and S and T indicate the total numbers of samples in each corresponding training set. Each domain is broken down into a registered/paired domain subset of size R, *i.e.*, $\mathscr{X}^P = \{\mathbf{X}_1, ..., \mathbf{X}_R\}$, $\mathscr{Y}^P = \{\mathbf{Y}_1, ..., \mathbf{Y}_R\}$, and an unregistered/unpaired domain subset of size $T - R$ or $S - R$, respectively, *i.e.* $\mathscr{X}^U = \{\mathbf{X}_{R+1}, ..., \mathbf{X}_S\}$, $\mathscr{Y}^U = \{\mathbf{Y}_{R+1}, ..., \mathbf{Y}_T\}$ so $\mathscr{X} = \mathscr{X}^P \cup \mathscr{X}^U$ and $\mathscr{Y} = \mathscr{Y}^P \cup \mathscr{Y}^U$. The assumption here is that $R \ll S, T$ and we only need access to a few registered pairs and a much larger set of unpaired images. Images in the sets $\mathscr{X}$ and $\mathscr{Y}$ are represented as $m \times n$ matrices whose columns are each of the 3D patches vectorized in lexicographic order. Hence, image data matrices $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_n] \in \mathbb{R}^{m \times n}$ and $\mathbf{Y} = [\mathbf{y}_1, ..., \mathbf{y}_n] \in \mathbb{R}^{m \times n}$, contain n overlapping 3D patches (covering the whole image volume) of dimension m (viz. the cardinality of the 3D patches). The training matrices $\mathbf{X}$ and $\mathbf{Y}$ are comprised of paired training sub-matrices $\mathbf{X}^P$, $\mathbf{Y}^P$ and unpaired training sub-matrices $\mathbf{X}^U$, $\mathbf{Y}^U$. We denote the test image in the same way by a matrix $\mathbf{X}^t$. The test 3D patches in $\mathbf{X}^t$ are acquired with modality $\mathscr{M}_1$, and will be the input to synthesize the corresponding 3D patches in modality $\mathscr{M}_2$.

Problem Statement: We first denote the coding coefficients $\mathbf{A}^{\mathscr{X}}$, $\mathbf{A}^{\mathscr{Y}}$ of $\mathbf{X}$, $\mathbf{Y}$ over the learned dictionaries $\mathbf{\Phi}^{\mathscr{X}}$, $\mathbf{\Phi}^{\mathscr{Y}}$, the projected data $\mathbf{P}^{\mathscr{X}}$, $\mathbf{P}^{\mathscr{Y}}$ of $\mathbf{X}$, $\mathbf{Y}$ in a defined common space, and a mapping function $\mathscr{F}(\cdot)$ to represent the relationship between the sparse codes $\mathbf{A}^{\mathscr{X}}$, $\mathbf{A}^{\mathscr{Y}}$ of two domains, where the detailed descriptions and the corresponding mathematical formulations are provided in the following subsections. Given a pair of training matrices $\mathbf{X}$ and $\mathbf{Y}$ with $\mathbf{X} = \left[\mathbf{X}^P \mathbf{X}^U\right]$ and $\mathbf{Y} = \left[\mathbf{Y}^P \mathbf{Y}^U\right]$, our goal is: 1) to learn a pair of dictionaries $\left\{\mathbf{\Phi}^{\mathscr{X}}, \mathbf{\Phi}^{\mathscr{Y}}\right\}$, their sparse codes $\left\{\mathbf{A}^{\mathscr{X}}, \mathbf{A}^{\mathscr{Y}}\right\}$, and an association function $\mathscr{F}(\cdot) : \mathscr{M}_1 \to \mathscr{M}_2$ using the projected data $\mathbf{P}^{\mathscr{X}}$ and $\mathbf{P}^{\mathscr{Y}}$; and 2) to minimize the inter-modality divergence between $\mathbf{P}^{\mathscr{X}}$ and $\mathbf{P}^{\mathscr{Y}}$, and 3) to preserve the domain-specific local geometric structure.

Dictionary Learning

Assume that $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_n] \in \mathbb{R}^{m \times n}$ is a training dataset, which can be reconstructed by the linear combination of a set of n coefficients that lie on a k-dimensional sparse space, $\mathbf{A}^{\mathscr{X}} = \left[\alpha_1^{\mathscr{X}}, ..., \alpha_n^{\mathscr{X}}\right] \in \mathbb{R}^{k \times n}$ is associated to the dictionary $\mathbf{\Phi}^{\mathscr{X}} = \left[\phi_1^{\mathscr{X}}, ..., \phi_k^{\mathscr{X}}\right] \in \mathbb{R}^{m \times k}$. Here, $k > m$ to make the dictionary over-complete [1]. Considering the reconstruction error for each data point, the problem of learning a dictionary $\mathbf{\Phi}^{\mathscr{X}}$ for sparse representation of $\mathbf{X}$ can be formulated as

$$\min_{\mathbf{\Phi}^{\mathscr{X}}, \mathbf{A}^{\mathscr{X}}} \left\| \mathbf{X} - \mathbf{\Phi}^{\mathscr{X}} \mathbf{A}^{\mathscr{X}} \right\|_F^2 + \lambda \left\| \mathbf{A}^{\mathscr{X}} \right\|_0, \tag{5.9}$$

where $\|\cdot\|_F$ is the Frobenius norm, $\|\cdot\|_0$ is the l_0-norm that penalizes non-zero elements in $\mathbf{A}$, and λ denotes a regularization parameter to trade off sparsity vs. reconstruction error. As shown in [26], the minimization problem in Eq. (5.9) is, in general, NP-hard under the l_0-norm. An alternative solution is to relax the l_0-norm with the l_1-norm and obtain a near-optimal result [20]. The dictionary learning problem in Eq. (5.9) can be reformulated as

$$\min_{\mathbf{\Phi}^{\mathscr{X}}, \mathbf{A}^{\mathscr{X}}} \left\| \mathbf{X} - \mathbf{\Phi}^{\mathscr{X}} \mathbf{A}^{\mathscr{X}} \right\|_F^2 + \lambda \left\| \mathbf{A}^{\mathscr{X}} \right\|_1. \tag{5.10}$$

The above objective function is not simultaneously convex over $\mathbf{\Phi}$ and $\mathbf{A}$. A practical solution is to alternate between optimizing for the dictionary $\mathbf{\Phi}$ and for the sparse codes $\mathbf{A}$ fixing the other degree of freedom. This makes the problem convex and the solution converges to a local minimum [86]. When the dictionary is fixed, the algorithm is known as Lasso/LARS [146] with an l_1 penalty over the coefficients and can be solved by the feature-sign search approach [86]. When sparse codes are fixed, such an optimization problem is reduced to a least squares optimization with quadratic constraints, and can be solved using a Lagrange dual [86].

When dealing with multi-modality data, one can simply construct two independent dictionaries using conventional dictionary learning. Specifically, given two training data sets $\mathbf{X}$ and $\mathbf{Y}$, following the dictionary learning procedure described in [1, 104] and Eq. (5.10), we can learn the dictionaries separately to obtain the two dictionaries, $\Phi^{\mathscr{X}}$ and $\Phi^{\mathscr{Y}}$, and the two corresponding sparse coefficients, $\mathbf{A}^{\mathscr{X}}$ and $\mathbf{A}^{\mathscr{Y}}$, respectively. The data of each modality can be reconstructed using the respective dictionary and associated sparse coefficients.

Cross-Modality Dictionary Learning

Cross-modality image synthesis is based on learning a joint sparse representation [168] with a common set of sparse codes shared between source and target image modalities, *i.e.* $\mathbf{A}^{\mathscr{X}} \equiv \mathbf{A}^{\mathscr{Y}}$. These sparse codes act on independent dictionaries for each modality, *viz.* $\Phi^{\mathscr{X}}$ and $\Phi^{\mathscr{Y}}$, to reconstruct the corresponding source and target images. To this effect, both 3D patches in the source and target modalities must be perfectly co-registered. To map the tissue appearance across modalities, the joint dictionary learning strategy groups two independent reconstruction errors (*viz.* $\|\mathbf{X}-\Phi^{\mathscr{X}}\mathbf{A}^{\mathscr{X}}\|_F^2$ and $\|\mathbf{Y}-\Phi^{\mathscr{Y}}\mathbf{A}^{\mathscr{Y}}\|_F^2$) in a single objective function to be optimized:

$$\begin{aligned} &\min_{\Phi^{\mathscr{X}},\Phi^{\mathscr{Y}},\mathbf{A}} \left\|\mathbf{X}-\Phi^{\mathscr{X}}\mathbf{A}\right\|_F^2 + \left\|\mathbf{Y}-\Phi^{\mathscr{Y}}\mathbf{A}\right\|_F^2 + \lambda \left\|\mathbf{A}\right\|_1 \\ &\text{s.t. } \left\|\phi_i^{\mathscr{X}}\right\|_2^2 \leq 1, \left\|\phi_i^{\mathscr{Y}}\right\|_2^2 \leq 1 \ \forall i = 1,\ldots,k, \end{aligned} \tag{5.11}$$

where $\mathbf{A}$ denotes the same coefficients to be enforced of registered data pairs projected in a common feature space. As in the single dictionary learning optimization problem, the joint optimization function in Eq. (5.11) is convex regarding the learned dictionaries, $\Phi^{\mathscr{X}}$ and $\Phi^{\mathscr{Y}}$, for fixed sparse codes $\mathbf{A}$. Therefore, the computation of $\mathbf{A}$ and of the dictionary pairs can be alternated. Analyzing (5.11), we note that this objective function is suitable to collaboratively learn a pair of dictionaries, so the sparse codes in the source modality space $\mathscr{M}_1$ can directly reconstruct the target modality image $\mathscr{M}_2$ in a transferable feature space.

Although joint dictionary learning achieves very good results, it assumes that source and target images, when represented with jointly learned dictionary pairs, $\Phi^{\mathscr{X}}$ and $\Phi^{\mathscr{Y}}$, must share the same sparse codes. In addition, all previous work requires that the training dataset contains registered image pairs, which imposes additional demands. In this work, we address the above problems by relaxing the need for a common sparse representation and providing more flexibility in reducing the registration requirement to a small training dataset only.

Weak Coupling and Geometry Co-regularization

To make the proposed method effective for generalized cross-modality synthesis, we combine the following ideas:

1. We integrate paired and unpaired training data in both modalities into a unified framework.
2. We relax the need for a shared sparse code in source and target domains.
3. We allow for dissimilar data distributions as required when dealing with very different image modalities.
4. We include a mechanism that preserves the local geometric structure specific to the modalities of the source and target images.

In the following, we introduce each component, and then summarize our overall approach.

Cross-Modality Image Matching

To relate and integrate the information from the paired and unpaired training data subsets of each modality, we introduce a criterion called cross-modality image matching (CMIM) for incorporating the information from the unpaired training data into dictionary learning and cross-modality image synthesis.

In visually matching cross-modality data, it is common to identify the same features across source and target imaging modalities. In this work, we extract High-Frequency (HF) features from source and target images where the resolutions of both modalities are similar. This is based on the assumption that tissues may present different visual appearances one each modality but they share similar high order edge/texture characteristics while modality-specific details affect primarily Low-Frequency (LF) properties [168].

In this chapter, we follow [16, 168] and adopt first- and second- order derivatives involving horizontal and vertical gradients as the HF features for each training data by $\mathbf{X}^h = \mathscr{H} * \mathbf{X}$, $\mathbf{Y}^h = \mathscr{H} * \mathbf{Y}$. Generally, $\mathscr{H}$ is a high-pass filter operator used to extract derivatives, Considering first and second order derivatives, $\mathscr{H}$ is one of the following operators: $\mathbf{H}_1^1, \mathbf{H}_1^2, \mathbf{H}_2^1, \mathbf{H}_2^2$, where $\mathbf{H}_1^1 = \begin{bmatrix} -1 & 0 & 1 \\ -1 & 0 & 1 \\ -1 & 0 & 1 \end{bmatrix}$, $\mathbf{H}_1^2 = \mathbf{H}_1^{1^T}$, and $\mathbf{H}_2^1 = \begin{bmatrix} -2 & -1 & 0 & 1 & 2 \\ -2 & -1 & 0 & 1 & 2 \\ -2 & -1 & 0 & 1 & 2 \\ -2 & -1 & 0 & 1 & 2 \\ -2 & -1 & 0 & 1 & 2 \end{bmatrix}$, $\mathbf{H}_2^2 = \mathbf{H}_2^{1^T}$. Once the features in both domains are computed, we can use them to optimize CMIM and define a mapping

$\mathscr{C}(\cdot) : \mathscr{X} \to \mathscr{Y}$. In particular, CMIM can be thought of as a unilateral matching metric (*i.e.*, the weighted regression) that focuses on a particular goal (*e.g.* matching across resolutions, modalities, or domains. [66, 183, 186]). Given associated HF image feature sets, $\mathbf{X}^h$ and $\mathbf{Y}^h$, corresponding to both paired and unpaired training image data sets, $\mathscr{X}^P$, $\mathscr{X}^U$, $\mathscr{Y}^P$ and $\mathscr{Y}^U$, CMIM represents an ensemble of paired and unpaired cross-modality matching sub-problems. Images in $\mathscr{X}^P$ and $\mathscr{Y}^P$ are endowed with a natural correspondence, $\mathscr{X}^P \rightleftharpoons \mathscr{Y}^P$. In contrast, CMIM finds a mapping for multi-modality unpaired image data for $\mathscr{X}^U$ and $\mathscr{Y}^U$. Since $\mathscr{X}^P$ and $\mathscr{Y}^P$ are already registered/paired, we can assume a perfect matching between them. By integrating the unpaired image data, we can establish a final *affinity matrix* $\mathbb{T} \in \mathbb{R}^{n \times n}$ such that $\mathscr{C}(\mathscr{X}, \mathscr{Y}) = \left\|\mathbf{X}^h - \mathbb{T}\mathbf{Y}^h\right\|_2^2$:

$$\mathbb{T} = \begin{bmatrix} D(\mathbf{x}_1^h, \mathbf{y}_1^h) & \cdots & D(\mathbf{x}_1^h, \mathbf{y}_n^h) \\ \vdots & \ddots & \vdots \\ D(\mathbf{x}_n^h, \mathbf{y}_1^h) & \cdots & D(\mathbf{x}_n^h, \mathbf{y}_n^h) \end{bmatrix}, \tag{5.12}$$

where $D(\mathbf{x}_i^h, \mathbf{y}_j^h)$ is a distance function generally designed to measure the distances between each pair of HF feature vectors in $\mathscr{X}$ and $\mathscr{Y}$ using an g-dimensional Gaussian kernel

$$D(\mathbf{x}_i^h, \mathbf{y}_j^h) = \frac{1}{(\sqrt{2\pi}\sigma)^g} e^{-\frac{\left\|\mathbf{x}_i^h - \mathbf{y}_j^h\right\|^2}{2\sigma^2}}, \tag{5.13}$$

where $\sigma \neq 0$ denotes the *kernel bandwidth*. $\mathbb{T}$ establishes a one-to-one correspondence for each source domain 3D patch. We preserve the most relevant features with the largest D values within $\mathscr{Y}$ while discarding other 3D patches. In this way, from $\mathbb{T}$ we define $\hat{\mathbb{T}}$ as:

$$\hat{\mathbb{T}}(i, j) = \begin{cases} 1, & \text{if } j = j_i, \\ 0, & \text{otherwise.} \end{cases} \tag{5.14}$$

where $j_i = \max_j (\mathbb{T}(i, j))$ is the maximum element of the i-th row in $\mathbb{T}$. Furthermore, we set the maximum element $\hat{\mathbb{T}}(i, j_i)$ to be 1 where all other values are set to 0 resulting in a binary assignment matrix $\hat{\mathbb{T}}$. Given $\hat{\mathbb{T}}$, each source patch is only mapped to one target patch with the most similar tissue texture. Hence, patches across different domains can be treated as the registered pairs after such a processing, *i.e.*, $\mathscr{X} \rightleftharpoons \mathscr{Y}$ for each $\mathbf{x}_i$ paired with $\mathbf{y}_{j_i}$ denoted as $\mathscr{P}_i = \{\mathbf{x}_i, \mathbf{y}_{j_i}\}$ for $i = 1 \ldots n$.

Computing the mapping function

Starting off by Eq. (5.10), by minimizing the reconstruction error, the corresponding sparse codes $\mathbf{A}^{\mathscr{X}}$ and $\mathbf{A}^{\mathscr{Y}}$ for each modality can be computed, respectively. To allow these codes to differ for the paired examples and unpaired data matched via CMIM, we assume there exists a mapping function $\mathscr{F}$: $\mathscr{M}_1 \to \mathscr{M}_2$ with $\mathbf{Y} = \mathscr{F}(\mathbf{X})$. Accordingly, the sparse codes of $\mathbf{X}$ and $\mathbf{Y}$ over the dictionaries will be related by such a mapping function $\mathscr{F}\left(\mathbf{A}^{\mathscr{X}}, \mathbf{A}^{\mathscr{Y}}\right)$. To build a stable mapping between two domains, Wang *et al.* [155] assumed that the sparse codes from the source domain had to be identical to those for the target domain via a linear projection $\mathbf{W}$. As suggested in [60], projecting both source and target domain data into a common feature space can better describe and associate cross-modality data. Inspired by this strategy, we first define the cross-modality relationship in the projected data $\mathbf{P}^{\mathscr{X}}$, $\mathbf{P}^{\mathscr{Y}}$ of $\mathbf{X}$, $\mathbf{Y}$, and replace $\mathscr{F}(\mathbf{A}^{\mathscr{X}}, \mathbf{A}^{\mathscr{Y}})$ by $\mathscr{F}(\mathbf{P}^{\mathscr{X}}, \mathbf{P}^{\mathscr{Y}})$, and then incorporate the projected features into CMIM-driven coupled dictionary learning. The objective function based on the pre-determined CMIM term $\left\|\mathbf{X}^h - \hat{\mathbf{T}}\mathbf{Y}^h\right\|_2^2$ of this learning model is:

$$\begin{aligned} &\min_{\Phi^{\mathscr{X}}, \Phi^{\mathscr{Y}}, \mathbf{A}^{\mathscr{X}}, \mathbf{A}^{\mathscr{Y}}} \left\|\mathbf{X} - \Phi^{\mathscr{X}}\mathbf{A}^{\mathscr{X}}\right\|_F^2 + \left\|\mathbf{Y} - \Phi^{\mathscr{Y}}\mathbf{A}^{\mathscr{Y}}\right\|_F^2 \\ &+ \lambda\left(\left\|\mathbf{A}^{\mathscr{X}}\right\|_1 + \left\|\mathbf{A}^{\mathscr{Y}}\right\|_1\right) \\ &+ \nu\mathscr{F}\left(\mathbf{P}^{\mathscr{X}}, \mathbf{P}^{\mathscr{Y}}\right), \end{aligned} \tag{5.15}$$

where $\mathbf{P}^{\mathscr{X}} = \mathbf{Q}^{\mathscr{X}}\mathbf{A}^{\mathscr{X}} \in \mathbb{R}^{k\times n}$ and $\mathbf{P}^{\mathscr{Y}} = \mathbf{Q}^{\mathscr{Y}}\mathbf{A}^{\mathscr{Y}} \in \mathbb{R}^{k\times n}$ denote the projected data of $\mathbf{X}$ and $\mathbf{Y}$, respectively, in the common feature space. Here, λ and ν are regularization parameters. The projection matrices, $\mathbf{Q}^{\mathscr{X}} \in \mathbb{R}^{k\times k}$ and $\mathbf{Q}^{\mathscr{Y}} \in \mathbb{R}^{k\times k}$ are the projection matrices for $\mathbf{A}^{\mathscr{X}}$ and $\mathbf{A}^{\mathscr{Y}}$, respectively. Generally, $\mathscr{F}(\mathbf{P}^{\mathscr{X}}, \mathbf{P}^{\mathscr{Y}})$ can be applied to any joint dictionary learning scheme with $\mathscr{F}(\mathbf{P}^{\mathscr{X}}, \mathbf{P}^{\mathscr{Y}}) = \|\mathbf{P}^{\mathscr{X}} - \mathbf{P}^{\mathscr{Y}}\|_F^2 = \|\mathbf{Q}^{\mathscr{X}}\mathbf{A}^{\mathscr{X}} - \mathbf{Q}^{\mathscr{Y}}\mathbf{A}^{\mathscr{Y}}\|_F^2$. For example, in Eq. (5.11) of [168], $\mathscr{F}$ is defined with an infinitely large ν having $\mathbf{Q}^{\mathscr{X}} = \mathbf{Q}^{\mathscr{Y}} = \mathbf{I}$, while in [155] $\mathscr{F}$ is defined so $\mathbf{Q}^{\mathscr{X}} = \mathbf{I}$ and $\mathbf{Q}^{\mathscr{Y}} = \mathbf{W}$, where $\mathbf{I}$ is the identity matrix. The solutions of $\mathbf{Q}^{\mathscr{X}}$ and $\mathbf{Q}^{\mathscr{Y}}$ are not unique. Following [60], an additional regularization constraint should be added to ensure the uniqueness of these solutions. Moreover, to guarantee the projected data lands in a common space and we can synthesize data of the target modality from projected data of the source modality, an additional regularization constraint is provided to make the function separately convex with respect to each variable. Given $\mathbf{P}^{\mathscr{X}}$ and $\mathbf{P}^{\mathscr{Y}}$, we minimize their distance in the projected common space considering the projections separately, *viz.* $\nu\left(\|\mathbf{A}^{\mathscr{X}}\mathbf{Q}^{\mathscr{X}} - \mathbf{P}^{\mathscr{Y}}\|_F^2 + \|\mathbf{A}^{\mathscr{Y}}\mathbf{Q}^{\mathscr{Y}} - \mathbf{P}^{\mathscr{X}}\|_F^2\right)$. Solving this objective function, we obtain $\mathbf{A}^{\mathscr{X}} = \mathbf{Q}^{\mathscr{X}^{-1}}\mathbf{P}^{\mathscr{Y}}$ and $\mathbf{A}^{\mathscr{Y}} = \mathbf{Q}^{\mathscr{Y}^{-1}}\mathbf{P}^{\mathscr{X}}$, where $\mathbf{P}^{\mathscr{X}} = \mathbf{Q}^{\mathscr{X}}\mathbf{A}^{\mathscr{X}}$ and $\mathbf{P}^{\mathscr{Y}} = \mathbf{Q}^{\mathscr{Y}}\mathbf{A}^{\mathscr{Y}}$ denote the projected data of $\mathbf{X}$ and $\mathbf{Y}$, respectively, in the constructed common feature space.

Maximum Mean Discrepancy Regularization

When the source and target image modalities have very different tissue appearances, corresponding patches may be associated to very different features and, hence, the mapping derived from CMIM may not be optimal. Matching by HF features can be insufficient here. We therefore add an extra term to CMIM to better constrain the optimal match between image pairs. We measure the divergence of the distribution of the empirical maximum mean discrepancy (MMD) [48, 8] over all matched image pairs. MMD is a nonparametric statistic utilized to assess whether two samples are drawn from the same distribution. In this chapter, we seek that the probability distributions of the projected data $\mathbf{P}^{\mathscr{X}}$ and $\mathbf{P}^{\mathscr{Y}}$ are identical in the common HF feature space. To this effect, we follow [48, 98, 139] and estimate the largest difference of $\mathbf{P}^{\mathscr{X}}$ and $\mathbf{P}^{\mathscr{Y}}$ in expectations over functions in the unit ball of a reproducing kernel Hilbert space:

$$\begin{aligned}\frac{1}{n^2}\left\|\sum_{i=1}^{n}\mathbf{p}_i^{\mathscr{X}}-\sum_{j=1}^{n}\mathbf{p}_j^{\mathscr{Y}}\right\|_H^2 &= \sum_{i,j=1}^{n}\mathbf{p}_i^{\mathscr{X}^{\mathrm{T}}}\mathbf{m}_{i,j}\mathbf{p}_j^{\mathscr{Y}} \\ &= \mathrm{Tr}\left(\mathbf{P}^{\mathscr{X}^{\mathrm{T}}}\mathbf{M}\mathbf{P}^{\mathscr{Y}}\right),\end{aligned} \tag{5.16}$$

where $(\cdot)^{\mathrm{T}}$ is the transpose operator, and $\mathbf{M}\in\mathbb{R}^{n\times n}$ denotes the matrix defined as:

$$\mathbf{m}_{i,j}=\begin{cases}1/n^2, & \text{if } j=j_i \text{, hence, } \{\mathbf{p}_i,\mathbf{p}_{j_i}\}\in\mathscr{P}_i,\\ -1/n^2, & \text{otherwise.}\end{cases} \tag{5.17}$$

The objective function is then rewritten by incorporating the MMD regularization term into Eq. (5.15).

Geometry Co-Regularization

During dictionary learning, features of $\mathbf{X}$ and $\mathbf{Y}$ are jointly captured in the dictionary atoms. However, this process focuses on the common space learning and fails to preserve modality/domain-specific information within the training image dataset. In this chapter, we attempt to represent specific modality properties by introducing the domain-specific graph Laplacian (a.k.a. geometry co-regularization term). To realize this idea, Lu *et al.* [99] and Zheng *et al.* [184] proposed the use of Laplacian eigenmaps to respect the intrinsic geometrical structure (manifold assumption) but their work focused on single-domain problems. Inspired by such a strategy, we capture and preserve the local geometric structure of each modality using the projected feature space. To be specific, given $\mathbf{P}^{\mathscr{X}}$ and $\mathbf{P}^{\mathscr{Y}}$ of $\mathbf{X}$ and $\mathbf{Y}$, respectively, one can construct two q-nearest neighbor graphs, $\mathscr{G}^{\mathscr{X}}$ and $\mathscr{G}^{\mathscr{Y}}$, with n vertices each based on prior work by [184]. The weight matrices $\mathbf{W}^{\mathscr{X}}$ and $\mathbf{W}^{\mathscr{Y}}$ of $\mathbf{G}^{\mathscr{X}}$ and $\mathbf{G}^{\mathscr{Y}}$ are then defined as the matrices with elements $w_{i,j}^{\mathscr{X}} = 1$ and $w_{i,j}^{\mathscr{Y}} = 1$ if and only if for any two features $\mathbf{p}_i^{\mathscr{X}}$, $\mathbf{p}_j^{\mathscr{X}}$ or $\mathbf{p}_i^{\mathscr{Y}}$, $\mathbf{p}_j^{\mathscr{Y}}$ satisfying: $\mathbf{p}_i^{\mathscr{X}}$ or $\mathbf{p}_i^{\mathscr{Y}}$ is among the q-nearest neighbors of $\mathbf{p}_j^{\mathscr{X}}$ or $\mathbf{p}_j^{\mathscr{Y}}$, otherwise $w_{i,j}^{\mathscr{X}} = 0$ or $w_{i,j}^{\mathscr{Y}} = 0$. Let $\mathbf{D}^{\mathscr{X}} = \mathrm{diag}\left(d_1^{\mathscr{X}}, \cdots, d_n^{\mathscr{X}}\right)$ and $\mathbf{D}^{\mathscr{Y}} = \mathrm{diag}\left(d_1^{\mathscr{Y}}, \cdots, d_n^{\mathscr{Y}}\right)$ be the degree matrices of $\mathbf{P}^{\mathscr{X}}$ and $\mathbf{P}^{\mathscr{Y}}$, with elements $d_j^{\mathscr{X}} = \sum_{i=1}^{n} w_{i,j}^{\mathscr{X}}$ and $d_j^{\mathscr{Y}} = \sum_{i=1}^{n} w_{i,j}^{\mathscr{Y}}$. Based on the graph Laplacian [5], we can define $\mathbf{G}^{\mathscr{X}} = \mathbf{D}^{\mathscr{X}} - \mathbf{W}^{\mathscr{X}}$ and $\mathbf{G}^{\mathscr{Y}} = \mathbf{D}^{\mathscr{Y}} - \mathbf{W}^{\mathscr{Y}}$, respectively. Considering the case of mapping the graphs $\mathbf{G}^{\mathscr{X}}$ and $\mathbf{G}^{\mathscr{Y}}$ to the projected features $\mathbf{P}^{\mathscr{X}}$ and $\mathbf{P}^{\mathscr{Y}}$, a reasonable criterion [96] for preserving the domain-specific geometrical strictures is designed by minimizing the following objective function:

$$\begin{aligned}
&\frac{1}{2}\sum_{i,j=1}^{n}\left(w_{i,j}^{\mathscr{X}}\left\|\mathbf{p}_i^{\mathscr{X}} - \mathbf{p}_j^{\mathscr{X}}\right\|_2^2 + w_{i,j}^{\mathscr{Y}}\left\|\mathbf{p}_i^{\mathscr{Y}} - \mathbf{p}_j^{\mathscr{Y}}\right\|_2^2\right) \\
=&\frac{1}{2}\sum_{i,j=1}^{n}\left(\mathbf{p}_i^{\mathscr{X}}\mathbf{p}_i^{\mathscr{X}^T} d_{i,i} - \mathbf{p}_i^{\mathscr{X}}\mathbf{p}_j^{\mathscr{X}^T} w_{i,j}^{\mathscr{Y}}\right. \\
&\left. + \mathbf{p}_i^{\mathscr{Y}}\mathbf{p}_i^{\mathscr{Y}^T} d_{i,i} - \mathbf{p}_i^{\mathscr{Y}}\mathbf{p}_j^{\mathscr{Y}^T} w_{i,j}^{\mathscr{Y}}\right) \\
=&\frac{1}{2}\mathrm{Tr}\left(\mathbf{P}^{\mathscr{X}}\mathbf{G}^{\mathscr{X}}\mathbf{P}^{\mathscr{X}^T} + \mathbf{P}^{\mathscr{Y}}\mathbf{G}^{\mathscr{Y}}\mathbf{P}^{\mathscr{Y}^T}\right).
\end{aligned} \tag{5.18}$$

The regularization criterion in Eq. (5.18) guarantees that the projected data varies smoothly along the geodesics of the manifold defined by the corresponding graph.

Objective Function

To summarize: we start-off with few registered cross-modal image-pairs and complement them with extensive unpaired images which are projected onto a common feature space. We then minimize the statistical divergence of the distributions of the projected data pairs. Finally, we preserve domain-specific properties by integrating the MMD and geometry co-regularization terms into Eq. (5.15) leading to the pre-determined CMIM $\left\|\mathbf{X}^h - \hat{\mathbb{T}}\mathbf{Y}^h\right\|_2^2$ driven objective function:

$$\begin{aligned}
\min_{\Phi,\mathbf{A},\mathbf{Q}} &\left\|\mathbf{X} - \Phi^{\mathscr{X}}\mathbf{A}^{\mathscr{X}}\right\|_F^2 + \left\|\mathbf{Y} - \Phi^{\mathscr{Y}}\mathbf{A}^{\mathscr{Y}}\right\|_F^2 \\
&+ \nu\left(\left\|\mathbf{A}^{\mathscr{X}}\mathbf{Q}^{\mathscr{X}} - \mathbf{P}^{\mathscr{Y}}\right\|_F^2 + \left\|\mathbf{A}^{\mathscr{Y}}\mathbf{Q}^{\mathscr{Y}} - \mathbf{P}^{\mathscr{X}}\right\|_F^2\right) \\
&+ \lambda\left(\left\|\mathbf{A}^{\mathscr{X}}\right\|_1 + \left\|\mathbf{A}^{\mathscr{Y}}\right\|_1\right) + \gamma\mathrm{Tr}\left(\mathbf{P}^{\mathscr{X}^T}\mathbf{M}\mathbf{P}^{\mathscr{Y}}\right) \\
&+ \frac{\mu}{2}\mathrm{Tr}\left(\mathbf{P}^{\mathscr{X}}\mathbf{G}^{\mathscr{X}}\mathbf{P}^{\mathscr{X}^T} + \mathbf{P}^{\mathscr{Y}}\mathbf{G}^{\mathscr{Y}}\mathbf{P}^{\mathscr{Y}^T}\right),
\end{aligned} \tag{5.19}$$

where γ and μ are the regularization parameters for trading off the effects of the MMD and geometry co-regularization terms, respectively.

Optimization

Similarly to existing joint dictionary learning methods [60, 99, 155], the optimization problem of Eq. (5.19) is not simultaneously convex regarding the dictionaries, sparse codes, and projection matrices. Instead, we divide the proposed method into three sub-problems: learning sparse coefficients, identifying a dictionary pair, and updating the projection matrices.

Computing Sparse Codes

We initialize the dictionary pair $\Phi^{\mathscr{X}}$, $\Phi^{\mathscr{Y}}$ and the projection matrices $\mathbf{Q}^{\mathscr{X}}$, $\mathbf{Q}^{\mathscr{Y}}$, fix them, and solve for $\mathbf{A}^{\mathscr{X}}$ and $\mathbf{A}^{\mathscr{Y}}$. Particularly, $\Phi^{\mathscr{X}}$ and $\Phi^{\mathscr{Y}}$ can be simply initialized as two random matrices (or use PCA or DCT bases), and $\mathbf{Q}^{\mathscr{X}}$, $\mathbf{Q}^{\mathscr{Y}}$ can be initialized to two identity matrices. Unlike conventional sparse coding, two additional terms are related to the projected feature space. Given $\Phi^{\mathscr{X}}$, $\Phi^{\mathscr{Y}}$ and $\mathbf{Q}^{\mathscr{X}}$, $\mathbf{Q}^{\mathscr{Y}}$, pre-determined by $\left\|\mathbf{X}^h - \hat{\mathbf{T}}\mathbf{Y}^h\right\|_2^2$ we can rewrite Eq. (5.19) as follows:

$$\begin{aligned}
&\min_{\mathbf{A}^{\mathscr{X}}} \left\|\mathbf{X} - \Phi^{\mathscr{X}}\mathbf{A}^{\mathscr{X}}\right\|_F^2 + \nu\left\|\mathbf{A}^{\mathscr{X}}\mathbf{Q}^{\mathscr{X}} - \mathbf{P}^{\mathscr{Y}}\right\|_F^2 \\
&+ \lambda\left\|\mathbf{A}^{\mathscr{X}}\right\|_1 + \mathrm{Tr}\left(\gamma\mathbf{P}^{\mathscr{X}^T}\mathbf{M}\mathbf{P}^{\mathscr{Y}} + \frac{\mu}{2}\mathbf{P}^{\mathscr{X}}\mathbf{G}^{\mathscr{X}}\mathbf{P}^{\mathscr{X}^T}\right), \\
&\min_{\mathbf{A}^{\mathscr{Y}}} \left\|\mathbf{Y} - \Phi^{\mathscr{Y}}\mathbf{A}^{\mathscr{Y}}\right\|_F^2 + \nu\left\|\mathbf{A}^{\mathscr{Y}}\mathbf{Q}^{\mathscr{Y}} - \mathbf{P}^{\mathscr{X}}\right\|_F^2 \\
&+ \lambda\left\|\mathbf{A}^{\mathscr{Y}}\right\|_1 + \mathrm{Tr}\left(\gamma\mathbf{P}^{\mathscr{X}^T}\mathbf{M}\mathbf{P}^{\mathscr{Y}} + \frac{\mu}{2}\mathbf{P}^{\mathscr{Y}}\mathbf{G}^{\mathscr{Y}}\mathbf{P}^{\mathscr{Y}^T}\right).
\end{aligned} \tag{5.20}$$

However, the problem in Eq. (5.20) is non-differentiable when the sparse codes take zero values. Coordinate Descent is usually adopted [1, 99, 184] to solve this l_1-regularized least squares problem. This is done by updating each vector $\alpha_i^{\mathscr{X}}$ or $\alpha_i^{\mathscr{Y}}$ individually while considering constant all other vectors $\alpha_j^{\mathscr{X}}$ or $\alpha_j^{\mathscr{Y}}$ where $j \neq i$. To optimize over each $\alpha_i^{\mathscr{X}}$ or $\alpha_i^{\mathscr{Y}}$, Eq. (5.20) can be expanded using vector-wise manipulations. Sparse representations in vector form can be solved by the feature-sign search algorithm [86].

Identifying Dictionary Pairs

Fixing the sparse codes $\mathbf{A}^{\mathscr{X}}$ and $\mathbf{A}^{\mathscr{Y}}$, learning dictionary pairs $\Phi^{\mathscr{X}}$ and $\Phi^{\mathscr{Y}}$ can be simplified and casted into quadratically constrained quadratic programing (QCQP):

$$\begin{aligned}
&\min_{\Phi^{\mathscr{X}},\Phi^{\mathscr{Y}}} \left\|\mathbf{X} - \Phi^{\mathscr{X}}\mathbf{A}^{\mathscr{X}}\right\|_F^2 + \left\|\mathbf{Y} - \Phi^{\mathscr{Y}}\mathbf{A}^{\mathscr{Y}}\right\|_F^2 \\
&\text{s.t.} \left\|\phi_i^{\mathscr{X}}\right\|_2^2 \leq 1, \left\|\phi_i^{\mathscr{Y}}\right\|_2^2 \leq 1 \;\; \forall i = \{1, \ldots, k\}.
\end{aligned} \tag{5.21}$$

Algorithm 6: WAG Algorithm

```
   Input: Training data $\mathbf{X}$ and $\mathbf{Y}$, parameters $\lambda$, $\mu$, $\sigma$, $\gamma$.
   Initialize $\Phi_0^{\mathscr{X}}$, $\Phi_0^{\mathscr{Y}}$, $\mathbf{A}_0^{\mathscr{X}}$, $\mathbf{A}_0^{\mathscr{Y}}$, $\mathbf{Q}_0^{\mathscr{X}}$, $\mathbf{Q}_0^{\mathscr{Y}}$.
   Let $\mathbf{Q}_0^{\mathscr{X}} = \mathbf{I}$, $\mathbf{Q}_0^{\mathscr{Y}} = \mathbf{I}$, $\mathbf{P}_0^{\mathscr{X}} \leftarrow \mathbf{A}_0^{\mathscr{X}}\mathbf{Q}_0^{\mathscr{X}}$, $\mathbf{P}_0^{\mathscr{Y}} \leftarrow \mathbf{A}_0^{\mathscr{Y}}\mathbf{Q}_0^{\mathscr{Y}}$.
   while not converged do
       Fix other variables, update $\mathbf{A}_{i+1}^{\mathscr{X}}$ and $\mathbf{A}_{i+1}^{\mathscr{Y}}$ by sparse coding according to Eq. (5.20).
       Fix other variables, update $\Phi_{i+1}^{\mathscr{X}}$ and $\Phi_{i+1}^{\mathscr{Y}}$ by dictionary learning according to Eq. (5.21).
       Fix other variables, update $\mathbf{Q}_{i+1}^{\mathscr{X}}$ and $\mathbf{Q}_{i+1}^{\mathscr{Y}}$ according to Eq. (5.24) based on $\mathbf{A}_{i+1}^{\mathscr{X}}$, $\mathbf{A}_{i+1}^{\mathscr{Y}}$ and $\Phi_{i+1}^{\mathscr{X}}$, $\Phi_{i+1}^{\mathscr{Y}}$.
       Update $\mathbf{P}_{i+1}^{\mathscr{X}} \leftarrow \mathbf{A}_{i+1}^{\mathscr{X}}\mathbf{Q}_{i+1}^{\mathscr{X}}$, $\mathbf{P}_{i+1}^{\mathscr{Y}} \leftarrow \mathbf{A}_{i+1}^{\mathscr{Y}}\mathbf{Q}_{i+1}^{\mathscr{Y}}$.
   end
   Output: $\Phi^{\mathscr{X}}$, $\Phi^{\mathscr{Y}}$ and $\mathbf{Q}^{\mathscr{X}}$, $\mathbf{Q}^{\mathscr{Y}}$.
```

The optimization in Eq. (5.21) can be solved by the Lagrange dual method [9].

Updating Projection Matrices

Considering constant the dictionary pairs and the corresponding sparse codes, we can then update the projection matrices by only considering $\mathbf{Q}^{\mathscr{X}}$ and $\mathbf{Q}^{\mathscr{Y}}$:

$$\min_{\mathbf{Q}^{\mathscr{X}},\mathbf{Q}^{\mathscr{Y}}} \nu(\left\|\mathbf{A}^{\mathscr{X}}\mathbf{Q}^{\mathscr{X}} - \mathbf{P}^{\mathscr{Y}}\right\|_F^2 + \left\|\mathbf{A}^{\mathscr{Y}}\mathbf{Q}^{\mathscr{Y}} - \mathbf{P}^{\mathscr{X}}\right\|_F^2). \tag{5.22}$$

Eq. (5.22) can be solved using simple ridge regression. Following [60], additional constraints, *viz.* $\delta\left(\|\mathbf{Q}^{\mathscr{X}}\|_F^2 + \|\mathbf{Q}^{\mathscr{Y}}\|_F^2\right)$ regarding the projection matrices $\mathbf{Q}^{\mathscr{X}}$ and $\mathbf{Q}^{\mathscr{Y}}$, are imposed to avoid over-fitting. We can rewrite Eq. (5.22) by combining the constraints as:

$$\begin{aligned} &\min_{\mathbf{Q}^{\mathscr{X}}} \nu\left\|\mathbf{A}^{\mathscr{X}}\mathbf{Q}^{\mathscr{X}} - \mathbf{P}^{\mathscr{Y}}\right\|_F^2 + \delta\left\|\mathbf{Q}^{\mathscr{X}}\right\|_F^2, \\ &\min_{\mathbf{Q}^{\mathscr{Y}}} \nu\left\|\mathbf{A}^{\mathscr{Y}}\mathbf{Q}^{\mathscr{Y}} - \mathbf{P}^{\mathscr{X}}\right\|_F^2 + \delta\left\|\mathbf{Q}^{\mathscr{Y}}\right\|_F^2. \end{aligned} \tag{5.23}$$

The solution of Eq. (5.23) can be analytically derived as

$$\begin{aligned} \mathbf{Q}^{\mathscr{X}} &= \mathbf{P}^{\mathscr{Y}}\mathbf{A}^{\mathscr{X}T}\left(\mathbf{A}^{\mathscr{X}}\mathbf{A}^{\mathscr{X}T} + (\delta/\nu)\mathbf{I}\right)^{-1}, \\ \mathbf{Q}^{\mathscr{Y}} &= \mathbf{P}^{\mathscr{X}}\mathbf{A}^{\mathscr{Y}T}\left(\mathbf{A}^{\mathscr{Y}}\mathbf{A}^{\mathscr{Y}T} + (\delta/\nu)\mathbf{I}\right)^{-1}, \end{aligned} \tag{5.24}$$

where $\mathbf{I}$ indicates an identity matrix. Algorithm 6 summarizes the proposed method.

Algorithm 7: Cross-Modality Image Synthesis

```
Input: Test image $\mathbf{X}^t$, dictionary pairs $\Phi^{\mathscr{X}}$ and $\Phi^{\mathscr{Y}}$, projection matrices $\mathbf{Q}^{\mathscr{X}}$ and $\mathbf{Q}^{\mathscr{Y}}$.
Initialize $\mathbf{A}_0^{t\mathscr{X}}$, $\mathbf{A}_0^{t\mathscr{Y}}$ by Eq. (5.25).
Let $\mathbf{A}_0^{t\mathscr{Y}} \leftarrow \mathbf{Q}^{\mathscr{Y}^{-1}}\mathbf{Q}^{\mathscr{X}}\mathbf{A}_0^{t\mathscr{X}}$, $\mathbf{Y}_0^t \leftarrow \mathbf{A}_0^{t\mathscr{Y}}\Phi_0^{\mathscr{Y}}$.
while not converged do
    Solve $\mathbf{A}_{i+1}^{t\mathscr{X}}$, $\mathbf{A}_{i+1}^{t\mathscr{Y}}$ using Eq. (5.20) with $\mathbf{Q}^{\mathscr{X}}$, $\mathbf{Q}^{\mathscr{Y}}$ and $\mathbf{Y}_i^t$.
    Update $\mathbf{Y}_{i+1}^t \leftarrow \mathbf{Q}^{\mathscr{Y}^{-1}}\mathbf{Q}^{\mathscr{X}}\mathbf{A}_{i+1}^{t\mathscr{X}}\Phi^{\mathscr{Y}} = \mathbf{A}_{i+1}^{t\mathscr{Y}}\Phi^{\mathscr{Y}}$.
end
Output: Synthesized image $\mathbf{Y}^t$.
```

Cross-Modality Image Synthesis

Once the optimization is completed, we can obtain the trained dictionary pairs, sparse coefficients and their projection matrices, and then apply the learned model to synthesize images across modalities. Given a test image $\mathbf{X}^t$, we first compute the coefficients $\mathbf{A}^{t\mathscr{X}}$ of $\mathbf{X}^t$ related to $\Phi^{\mathscr{X}}$ by solving a single sparse coding problem in Eq. (5.10). After that, we associate $\mathbf{A}^{t\mathscr{X}}$ to the expected sparse codes $\mathbf{A}^{t\mathscr{Y}}$ via $\mathbf{Q}^{\mathscr{X}}$ and $\mathbf{Q}^{\mathscr{Y}}$ leading to

$$\mathbf{A}^{t\mathscr{Y}} \approx \mathbf{Q}^{\mathscr{Y}^{-1}}\mathbf{P}^{t\mathscr{X}} = \mathbf{Q}^{\mathscr{Y}^{-1}}\mathbf{Q}^{\mathscr{X}}\mathbf{A}^{t\mathscr{X}}, \tag{5.25}$$

where $\mathbf{P}^{t\mathscr{X}}$ is the projected data of $\mathbf{X}^t$. Finally, the data in the target $\mathscr{M}_2$ modality, $\mathbf{Y}^t$, can be synthesized by $\mathbf{Y}^t = \mathbf{A}^{t\mathscr{Y}}\Phi^{\mathscr{Y}}$. Algorithm 7 summarizes the process for cross-modality image synthesis.

5.2.3 Experiments

Herewith, we describe an extensive experimental evaluation of the proposed method. We first introduce the datasets used for the evaluation, the experimental settings, and the methods we benchmark against. Finally, we show the statistical significance test to assess the importance of our improvements.

Databases and Pre-processing

We validate our method on two public multi-modality brain datasets, *viz.* IXI[3] and NAMIC[4] databases, respectively. The IXI database involves 578 healthy subjects each imaged using a matrix of 256×256×v (v = 112∼136) scanned with a Magnetic Resonance Imaging (MRI)

[3] http://brain-development.org/ixi-dataset/

[4] http://hdl.handle.net/1926/1687

Table 5.3 The number of selected paired/unpaired images.

	IXI		NAMIC		RATIO
	PAIRED SETS	UNPAIRED SETS	PAIRED SETS	UNPAIRED SETS	PAIRED/FULL SET
Scenario #1	289	–	10	–	100%
Scenario #2	145	72	6	2	50.2%
Scenario #3	73	108	4	3	25.3%
Scenario #4	37	126	2	4	12.8%

system. The NAMIC database, instead, contains 20 subjects (ten are normal controls and the other ten are schizophrenic) each imaged using a matrix of $128\times128\times z$ ($z = 88$) scanned with a 3T MRI system. For our experiments, we adopt PD-w, T2-w MRI scans from the IXI dataset, and T1-w, T2-w acquisitions form the NAMIC dataset. Following [127, 153, 172], all the experimental images are skull stripped, linearly registered and/or inhomogeneity corrected. In the experiments, we perform a more challenging division by applying half of the dataset for training while the remaining for testing. Particularly, by fixing the number of test data (*i.e.*, 289 subjects for IXI and 10 subjects for NAMIC, respectively), we divide our training set into two subsets with registered image pairs and unpaired image sets (in each domain). We evaluate these four cases listed in Table 5.3 for two datasets separately. Specifically, Table 5.3 shows the number of selected paired/unpaired images with respect to different modalities for each scenario we explored. The ratio of paired images over the full training set are 100%, 50%, 25% and 13% for Scenarios #1 to #4, respectively. Correspondingly, WAG has 289, 145, 73 and 37 original registered pairs for training for each scenario. To create a set of unpaired images valid for a fair comparison, we remove the other half of available paired to generated a similar amount of paired image sets for each scenario. For instance, at the Scenario #2, 72 out of 144 sets (for 145 registered image pairs) are used for training as the unpaired data, and so on. The logical presentation of Scenario #2 can be expressed as:

- Paired sets: A = 145 subjects with both PD-w and T2-w images.
- Unpaired sets: B = 72 subjects with PD-w images.
- Unpaired sets: C = 72 subjects with T2-w images.
- $\mathrm{A} \cap \mathrm{B} \cap \mathrm{C} = \varnothing$

Experimental Setup

We evaluate our method in two scenarios. First, we use the IXI dataset for synthesizing the T2-w images from the PD-w acquisitions and *vice versa*. Second, we adopt the NAMIC dataset for generating the T1-w scans from the T2-w inputs and *vice versa*. In our experiments, we

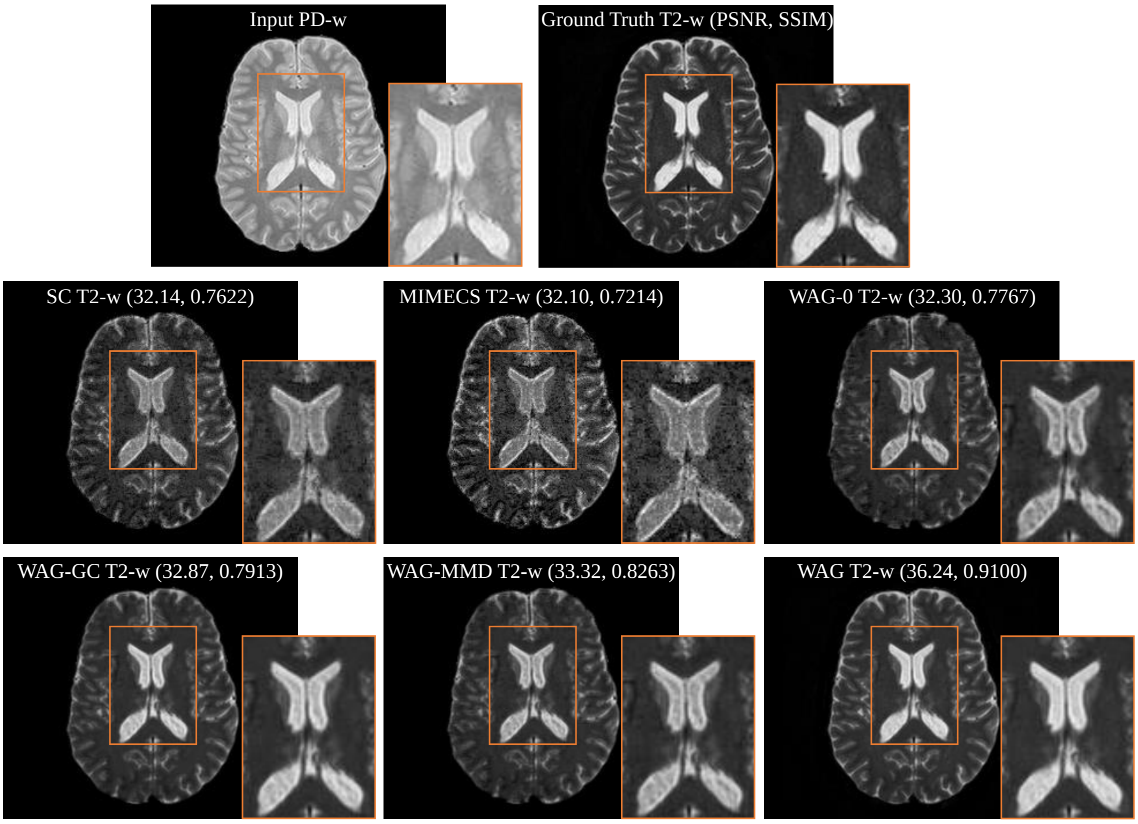

Input PD-w
Ground Truth T2-w (PSNR, SSIM)
SC T2-w (32.14, 0.7622)
MIMECS T2-w (32.10, 0.7214)
WAG-0 T2-w (32.30, 0.7767)
WAG-GC T2-w (32.87, 0.7913)
WAG-MMD T2-w (33.32, 0.8263)
WAG T2-w (36.24, 0.9100)

Fig. 5.3 Synthesized results generated using SC, MIMECS, WAG-0, WAG-GC, WAG-MMD and WAG (zoom in for details).

randomly select 100 thousand training patch pairs from both datasets respectively, which have no relation with the test images used in our experiments. We consider patches of dimension $5 \times 5 \times 5$ voxels. Following [60, 98], the regularization parameters γ, λ, μ, and ν are empirically set to be 10^5, 0.15, 1, 0.01, respectively. The number of atoms in the learned dictionary is set as 1024 according to [168]. Correspondingly, matrix $\mathbf{P}$ has n items in the k dimensional space, $\mathbf{Q}$ has k elements in the k dimensional space, $\mathbf{G}$ and $\mathbb{T}$ have n items in the n dimensional space, where n is the size of the training set and k is the size of the trained dictionary. Unless otherwise explicitly stated, we always use scenario #4 in all our experiments, which is a more challenging case between paired training data and unpaired training data (we will examine the effects of all scenarios in Section 5.2.3). For the evaluation metrics, we adopt the widely used Peak Signal to Noise Ratio (PSNR) and Structural Similarity Index (SSIM) [159] to objectively assess the quality of the synthesized images.

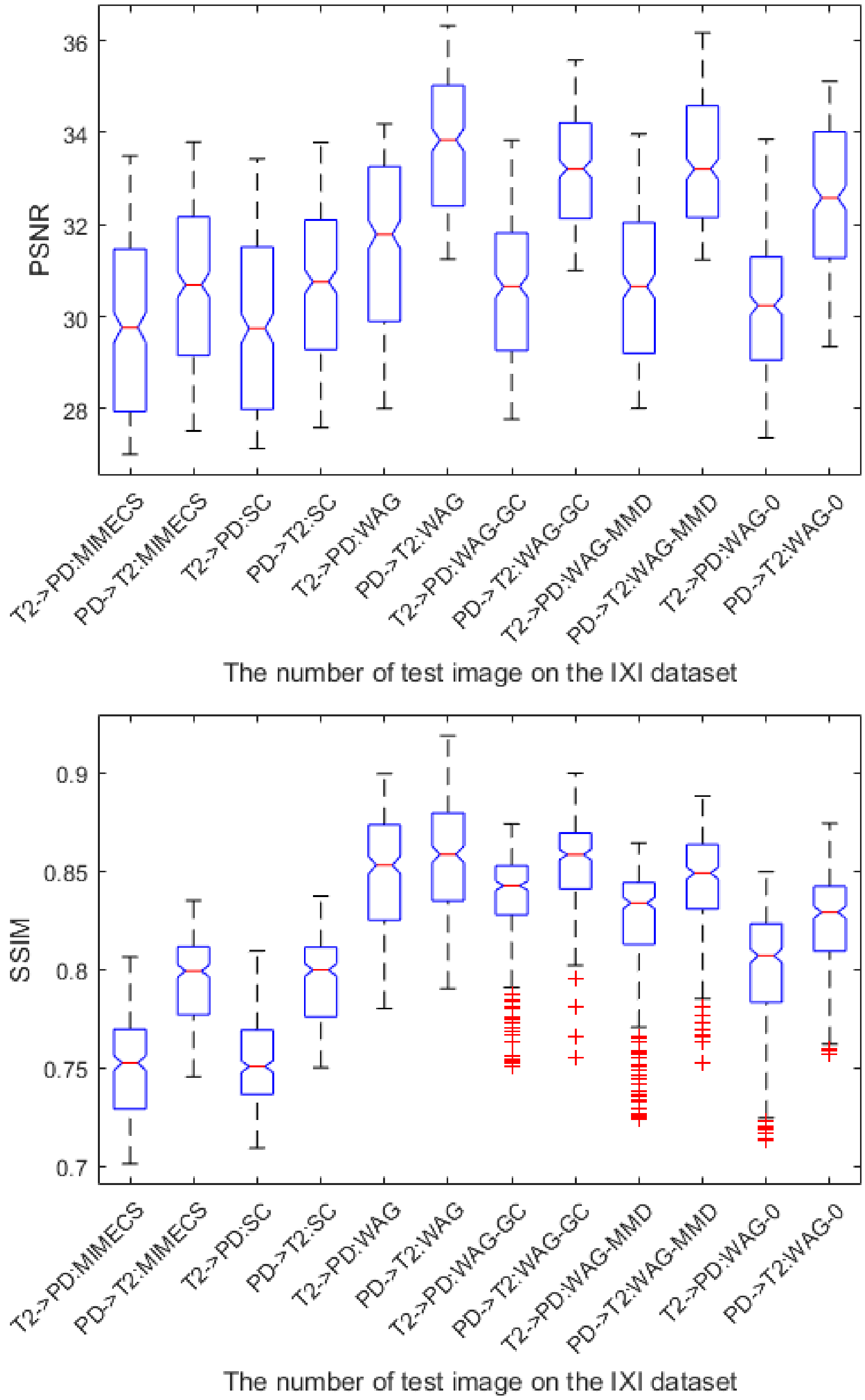


Fig. 5.4 Cross-modality synthesis results: MIMECS, SC, WAG, WAG-MMD, WAG-GC and WAG-0 on the IXI dataset.

Compared Methods

To fully evaluate the effectiveness of the proposed method in different patient groups (*e.g.* health or pathology), we conduct comprehensive evaluation on two public datasets and compare WAG with four state-of-the-art (related) approaches for cross-modality image synthesis:

Table 5.4 PSNRs and SSIMs of the WAG-synthesized images resulting from different paired/full set ratios during dictionary training.

IXI Dataset								
Metric (mean)	Scenario #1		Scenario #2		Scenario #3		Scenario #4	
	T2-w ↦ PD-w	PD-w ↦ T2-w	T2-w ↦ PD-w	PD-w ↦ T2-w	T2-w ↦ PD-w	PD-w ↦ T2-w	T2-w ↦ PD-w	PD-w ↦ T2-w
PSNR (dB)	**32.11**	**34.46**	31.97	34.27	31.68	34.02	31.54	33.73
SSIM	**0.8551**	**0.8602**	0.8539	0.8589	0.8527	0.8578	0.8506	0.8549

Table 5.5 Performance measures of the WAG-synthesized images resulting from different paired/full set ratios during dictionary training.

Metric (mean)	Fixing the number of paired data as 145							
	no unpaired data		36 unpaired data		48 unpaired data		72 unpaired data	
	T2-w ↦ PD-w	PD-w ↦ T2-w	T2-w ↦ PD-w	PD-w ↦ T2-w	T2-w ↦ PD-w	PD-w ↦ T2-w	T2-w ↦ PD-w	PD-w ↦ T2-w
PSNR (dB)	31.58	33.88	31.60	33.97	31.71	34.04	**31.97**	**34.27**
SSIM	0.8514	0.8563	0.8519	0.8570	0.8528	0.8580	**0.8539**	**0.8589**

- SC: Sparse Coding-based method [168]
- MIMECS: MRI example-based contrast synthesis [127]
- Ve-S: Vemulapalli's supervised [153]
- Ve-US: Vemulapalli's unsupervised [153]
- WAG-0: WAG without any regularization terms
- WAG-MMD: WAG using MMD regularization only
- WAG-GC: WAG using Geometric Co-regularization only
- WAG: Fully fledged WAG method

In particular, SC can be cast as a fundamental baseline only considering the joint dictionary learning. MIMECS, Ve-S and Ve-US are the most relevant and state-of-the-art cross-modality image synthesis approaches. We consider three special cases of the proposed method by excluding all regularization terms (WAG-0) or including only either MMD term (WAG-MMD) or geometric co-regularization term (WAG-GC) for proving that each of the added term is useful for more accurate synthesis. The mathematical models of WAG-MMD and WAG-GC are provided in Section 5.2.2.

Experimental Results

As we mentioned in Section 5.2.3, we first address cross-modality synthesis on the IXI dataset. In this scenario, we investigate both PD-w and T2-w images for evaluating and comparing the proposed WAG method with SC and MIMECS. To validate that our regularization terms are beneficial, we compare WAG with WAG-0, WAG-MMD and WAG-GC and show a set of visual

Table 5.6 Performance measures of the WAG-synthesized images resulting from different paired/full set ratios during dictionary training.

Metric (mean)	Fixing the number of unpaired data as 72					
	37 paired data		73 paired data		145 paired data	
	T2-w ↦ PD-w	PD-w ↦ T2-w	T2-w ↦ PD-w	PD-w ↦ T2-w	T2-w ↦ PD-w	PD-w ↦ T2-w
PSNR (dB)	31.35	33.54	31.57	33.86	**31.97**	**34.27**
SSIM	0.8487	0.8532	0.8514	0.8560	**0.8539**	**0.8589**

Table 5.7 Averaged PSNRs and SSIMs of the synthesized images using different methods on the NAMIC dataset.

NAMIC Dataset										
Metric (mean)	T1-w ↦ T2-w					T2w ↦ T1-w				
	MIMECS	SC	Ve-US	Ve-S	WAG	MIMECS	SC	Ve-US	Ve-S	WAG
PSNR (dB)	23.88	24.58	26.70	27.76	**27.96**	27.05	26.90	27.66	29.40	**30.40**
SSIM	0.8779	0.8778	0.8832	0.8874	**0.8991**	0.9165	0.9177	0.9168	0.9182	**0.9259**

results in Fig. 5.3, while reporting all quantitative results in Fig. 5.4. From Fig. 5.4, we see that WAG is always better than MIMECS especially with the standard WAG (with two regularization terms). We also explore the effectiveness of different numbers of paired and unpaired subjects listed in Table 5.3. The averaged PSNRs and SSIMs are shown in Table 5.4. Generally, a larger number of paired subjects leads to better synthesis results. The proposed method under the weakly coupled settings (*i.e.* small number of paired images in scenario #4) can match the performance of fully coupled method (in scenario #1) for cross-modality synthesis. To see the impact of the number of registered image pairs or unpaired data in WAG, in Tables 5.5 and 5.6, we show the mean performance of our proposed method based on different ratios of paired and unpaired data. In those results, we first fix the number of registered image pairs to be 145 (referring to scenario #2) to observe the performance variation by increasing the number of unpaired data from 36 to 72. Generally, more unpaired data yield better results. We evaluate how the number of paired data influences the synthesized results given the fixed number of unpaired images as 72. The number of paired images is set to 37, 73 and 145 (the same sets in scenario #2-#4). The more existing paired data, the better the synthesized results.

In the second scenario, we evaluate WAG and other relevant methods on the NAMIC dataset involving two sets of major experiments. The representative and stat-of-the-art synthesis methods, including SC, MIMECS, Ve-S and Ve-US are employed to compare with our WAG model. We demonstrate visual and quantitative results in Fig. 5.5, Fig. 5.6 and summarize the averaged values in Table 5.7, respectively. It can be seen that our method yields the best results against the compared approaches proving our claim of being able to synthesize better results through the added two regularization terms under weakly-supervised setting.

All of our experiments were performed on an Intel Xeon E5-1620 CPU (3.5 GHz, 8 cores) machine running Windows 10 with 32 GB of RAM. Training of WAG took, on average,

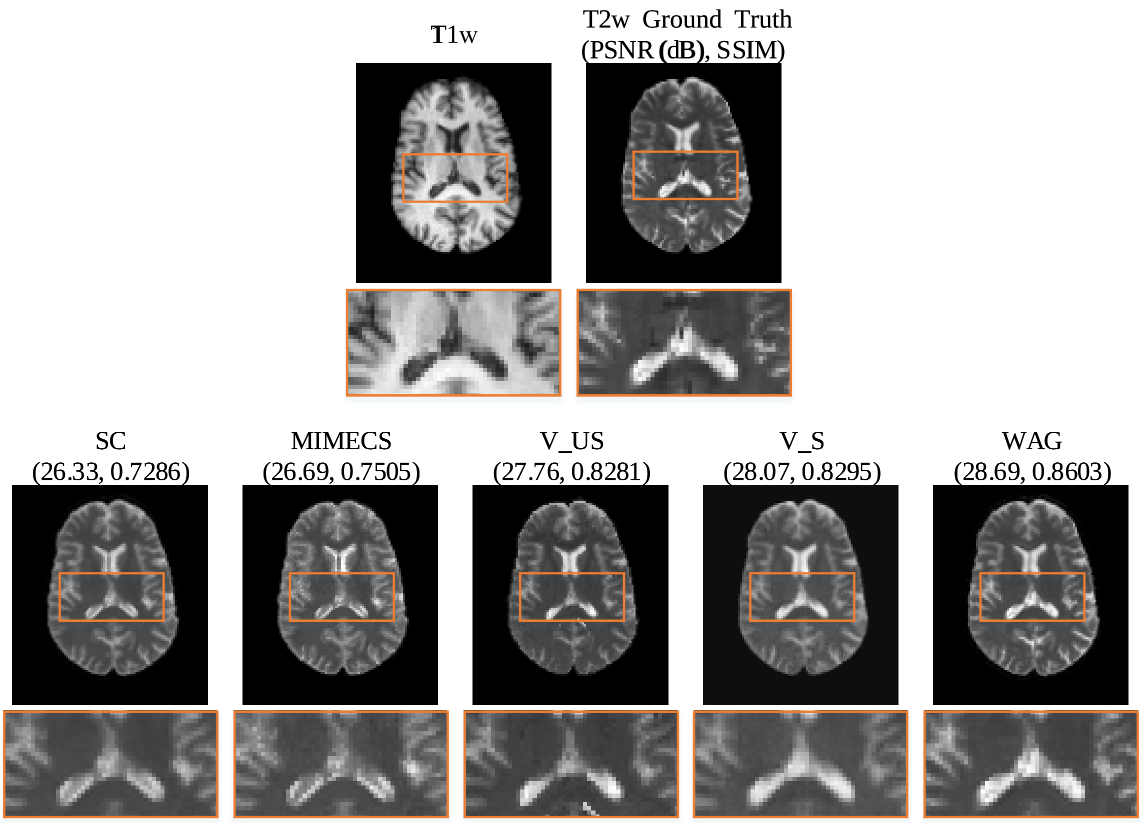


Fig. 5.5 Example cross-modality synthesis results generated by MIMECS, SC, Ve-S, Ve-US and WAG on the NAMIC dataset.

about 30 minutes using a Matlab R2013a code. Execution time for the synthesis of one 3D representative image with size 256×256×100 pixels took about 7 minutes.

Statistical Test

We conduct two statistical tests illustrating the significance of the improvements introduced by (1) the various regularization terms within WAG, and (2) our method compared with other state-of-the-art approaches. Regarding the characteristics of the comparison, we employ a paired-sample t-test for group (1) and independent (two-samples) t-test for group (2) at 5% significance level. Table 5.8 lists the results of paired t-test for case (1), which shows our improvements are all statistically significant. Tables 5.9 and 5.10 show the results of independent t-test for case (2), which demonstrates that the performance benefits of our method against others are statistically significant in all but one case, i.e., synthesizing T1-w images from T2-w data on the NAMIC dataset using Ve-S method.

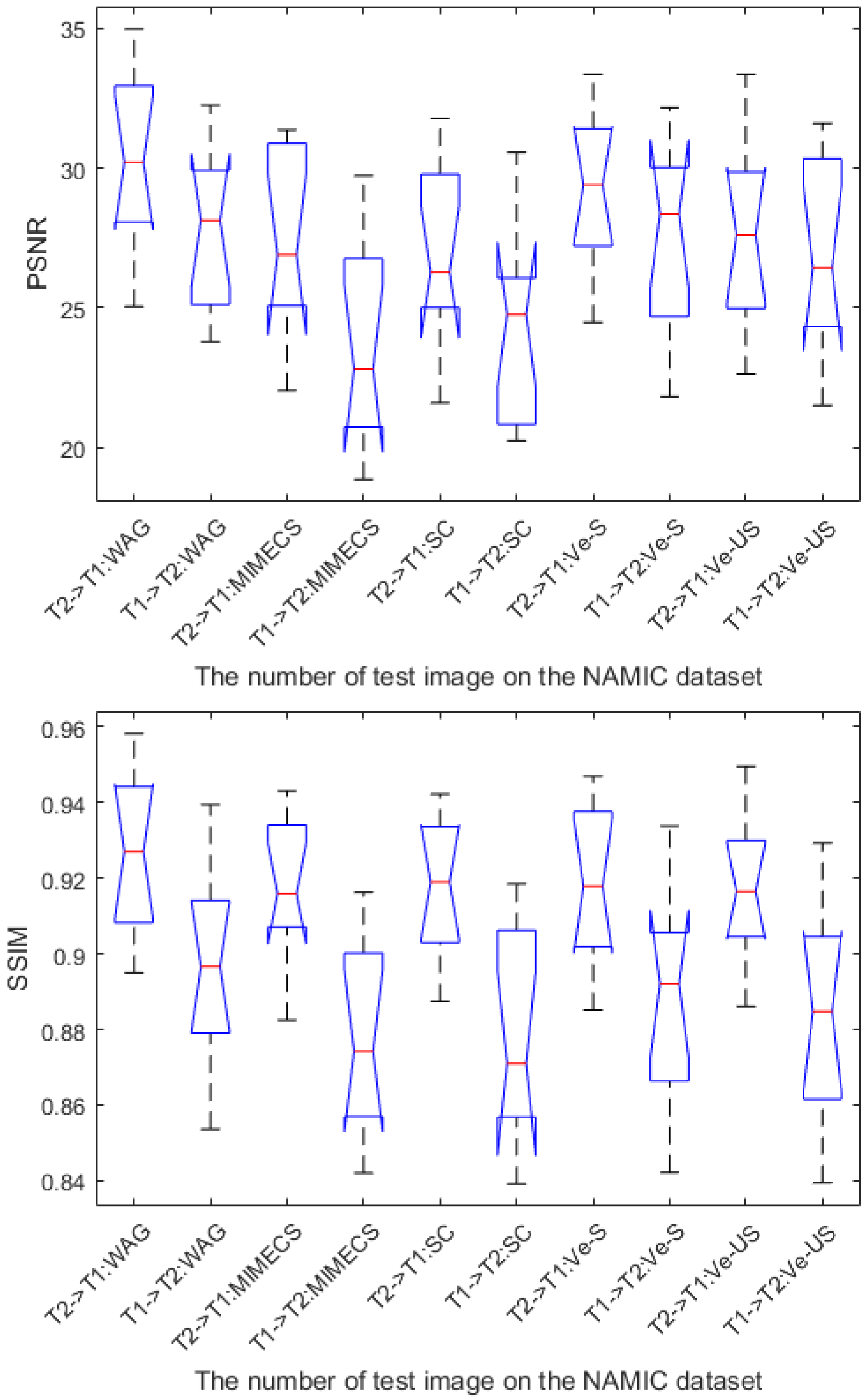


Fig. 5.6 Cross-modality synthesis results: MIMECS, SC, WAG, Ve-S and Ve-US on the NAMIC dataset.

Table 5.8 Paired t-test on the WAG improvements using the IXI dataset.

Paired t-test	WAG vs. WAG-0	WAG vs. WAG-MMD	WAG vs. WAG-GC
IXI: T2-w ↦ PD-w			
p-value (PSNR)	< 0.001	< 0.001	< 0.001
p-value (SSIM)	< 0.001	< 0.001	< 0.001
IXI: PD-w ↦ T2-w			
p-value (PSNR)	< 0.001	< 0.001	< 0.001
p-value (SSIM)	< 0.001	< 0.001	< 0.001

Table 5.9 Independent t-test on the performance benefits using the IXI dataset.

Independent t-test	WAG vs. MIMECS	WAG vs. SC
IXI: T2-w ↦ PD-w		
p-value (PSNR)	< 0.001	< 0.001
p-value (SSIM)	< 0.001	< 0.001
IXI: PD-w ↦ T2-w		
p-value (PSNR)	< 0.001	< 0.001
p-value (SSIM)	< 0.001	< 0.001

Table 5.10 Independent t-test on the performance benefits using the NAMIC dataset.

Independent t-test	WAG vs. MIMECS	WAG vs. SC	WAG vs. Ve-US	WAG vs. Ve-S
NAMIC: T1-w ↦ T2-w				
p-value (PSNR)	0.0319	0.0308	0.0450	0.0363
p-value (SSIM)	0.0347	0.0396	0.0468	0.0392
NAMIC: T2-w ↦ T1-w				
p-value (PSNR)	0.0168	0.0361	**0.0809**	0.041
p-value (SSIM)	0.0143	0.0138	0.0464	0.0345

5.2.4 Discussions

To investigate the performance of the proposed method, in this paper, we extensively validated WAG on two public datasets, i.e., IXI and NAMIC. We compared our results with other state-of-the-art methods for cross-modality image synthesis. We illustrated our method on different synthesis scenarios of structural brain MRI and synthesized images of both healthy and schizophrenic subjects. A few registered multi-modality image pairs were employed and then enriched with a larger set of unpaired data showing improved synthesis quality. Broadly, unlike most of state-of-the-art methods heavily relying on supervised learning, the proposed method allows using weakly-supervised data for generating competitive synthesis results. Surprisingly, from Table 5.4 and Table 5.7, we can see that with only 12.8% originally paired data, WAG achieves comparable results as WAG using 100% paired data for synthesis of either T2-w or PD-w images from the opposite modality. For the synthesis from T2-w data to PD-w data, WAG (#1) with 100% registered image pairs outperforms WAG (#4) with 12.8% registered image pairs by 0.57 dB for PSNR and 0.0045 for SSIM, on average. For the synthesis from PD-w data to T2-w data, WAG (#1) outperforms WAG (#4) by 0.73 dB for PSNR and 0.0053 for SSIM, on average. By using 12.8% paired data, WAG offers an excellent performance in two evaluation metrics in all experiments compared with MIMECS, SC, Ve-US and Ve-S while the compared methods required 100% registered image pairs for training. Specifically, as shown in Fig. 5.5, 5.6 and Table 5.7, for generating T2-w from T1-w data, the average gains on PSNR and SSIM achieved by WAG are 4.08 dB and 0.0212 higher than the worst performing approach on the NAMIC dataset. Also, for synthesizing T1-w images from T2-w data, the average gains on PSNR and SSIM achieved by WAG are 3.35 dB and 0.0094 higher than the worst performing approach on the NAMIC dataset as well. WAG achieves the best performance among all supervised state-of-the-art methods under a weakly-supervised setting (with only 12.8% registered data) in both healthy and pathological scenarios. This reveals its capability in effectively leveraging data to boost the learning system. Therefore, the proposed method is usable in clinical practice considering the fact that collecting parallel image pairs is costly and usually limited in many situations.

WAG achieves compelling synthesis results in this paper for the specific MRI modalities investigated here. However, our method could be potentially applied to other imaging modalities having the assumption that images with similar high order edge/texture characteristics and resolutions. It remains to be demonstrated the synthesis quality in more complex settings like, for instance, for the synthesis of PET images from MRI data, for the synthesis of MRI data from CT images, and for the more challenging cases such as the synthesis of a tumor case. In addition, to address multi-modality image synthesis involving more than two modalities, the natural extension of the proposed method would currently required that all source modalities would be available at once at the input. We are aware of very recent work by other researchers that handle multi-modality image synthesis even in the absence of one of some source modalities [17]. In our future work, we plan to explore extensions to our framework based on multi-modality image fusion of the source modalities before the synthesis. Fused features can better express multiple source modalities and thus synthesize the target image modality even with only partial input sources.

5.2.5 Conclusions

We proposed a weakly-coupled and geometry co-regularized joint dictionary learning (WAG) method for cross-modality synthesis of MRI images. Most conventional joint dictionary learning methods with sparse representations assume a fully supervised setting. Instead, our method only requires a small subset of registered image pairs and automatically finds correspondences for a much larger set of unpaired images. This process assists and enriches the supervised learning on the smaller subset while booting synthesis performance. With the proposed cross-modality image matching criterion, the derived common feature space associates cross-modality data effectively by updating a pair of dictionaries in both domains. We integrated our model with both MMD and modality-specific geometric co-regularization terms to further improve image synthesis quality. The proposed WAG approach was applied to cross-modality image synthesis of brain MRI and experimental results demonstrated that WAG significantly outperforms competing state-of-the-art methods on two public databases with healthy and schizophrenic subjects.

Chapter 6

Task-Driven Bidirectional Fault-Aware Adversarial Networks

Different imaging modalities present valuable and complementary anatomical information about the underlying organ for medical image analysis. However, a full battery of multi-modality data is often scarce, costly and even impossible to obtain. The ability to automatically generate meaningful synthetic information is highly desirable for many computer-aided medical applications. With recent progress in generative adversarial networks (GANs) particularly on natural images, instead of collecting acquisitions relying on the scanner, an appealing alternative is to generate highly compelling subject-specific images, such as structural magnetic resonance imaging. Inspired by the framework of adversarial learning, we propose the task-driven bidirectional fault-aware adversarial networks (T-GAN) for volumetric neuroimage processing in an unsupervised manner. T-GAN provides a generic way of learning a dual mapping between source and target domains while considering both visually high-fidelity synthesis and task-practicability. Through combining the bidirectional adversarial loss, the cycle-consistency loss and domain adapted loss in volumetric space, T-GAN is robust for medical images under multiple conditions. In addition to generating the desirable modality data, we complement both discriminators conditioned on the performance of any optional post-processing which balances synthesis fidelity and task performance. Experiments evaluated on several cross-modality synthesis tasks show that T-GAN not only produces visually impressive results but also can substitute real acquisitions for the clinical post-processing, and also outperforms some related state-of-the-art methods.

6.1 Introduction

Medical imaging enjoys a multitude of imaging modalities such as Computed Tomography (CT), Magnetic Resonance Imaging (MRI) and Positron-Emission Tomography, and opens up the opportunity to gain insights into diverse tissues' characteristics via different physical acquisition principles or parameters. The simultaneous availability of multi-modal imaging has benefited a wide range of brain image analysis tasks, for example, providing complementary information to discriminate specific tissues, anatomies and pathologies in image segmentation [54], or improved cross-modal image registration under the great variability of tissue or organ appearance [136].

Despite that neuroimaging such as MRI can produce highly detailed images with great soft tissue contrast, one practical problem remains largely unsolved: imaging has notoriously long acquisition times, which hampers multi-protocol MRI acquisitions at high-resolution. Additional complications arise when imaging patients with some specific medical conditions, for instance, Alzheimer's Disease and Parkinson's Disease, require that the acquisitions are collected in even short times resulting in very low-quality images.

Automatic generation of synthesized images for recovering corrupt/missing modality data or extending a dataset to additional modalities has been shown to be an effective way in several applications of the medical community [65, 150, 153]. Especially in recent years, for auxiliary diagnosis and scientific research, propagating and synthesizing non-acquired or higher-resolution imaging modalities has seen rapid progress, fueled increasingly by the development of the machine learning and computer vision techniques. This problem can be more broadly casted as cross-modality image synthesis [64, 75], which is closely related to *texture transfer* [3, 32, 35, 59], also known as image-to-image translation or image style transfer in the context of computer vision. That is, we seek to render a source modality image into a target modality, for instance, to synthesize a CT image as an MRI image, or a PD-w MRI image as a T2-w MRI image, or a structural MRI image as the corresponding diffusion tensor MRI image.

Among techniques for addressing such cross-modality synthesis tasks, sparse representation [42, 127] and deep learning-based approaches [44, 69] particularly favor the scenarios in which fully paired images in different modalities are available and have shown their promising results. However, collecting a large amount of paired data is impractical in the medical imagining community. To cope with cases where paired data are scarce or even unavailable, semi-supervised [21, 155], weakly-supervised [65] and unsupervised methods [93, 119, 173, 189] are developed respectively, aiming at learning a joint representation or distribution of images in different modalities of source and target domains. Due to lack of paired information, the unsupervised cross-modality synthesis problem is much harder but more applicable in clinical routine and post-processing (e.g., segmentation and registration).

While the existing synthesis approaches have shown visually impressive results, it is unclear if the synthesized data are mere pastiches that imitates the modality of the desirable one for pure visualization or they can substitute real acquisitions in the respective tasks for accurate quantitative image analysis and diagnosis. When the underlying relationship between different modalities is known a priori, the design of cross-modality image synthesis can ideally be task-driven, such that the synthesized results are tailored to the information to determine them within the respective reliability standards. Moreover, the generated images present additional information to improve the performance of the task interactively.

In this work, we propose to tackle the problem of unsupervised cross-modality synthesis by exploring the task-driven deep architecture for volumetric neuroimage processing. The concept of task is broad, given that the performance is dependent on the information acquired from the image. This chapter mainly focuses on the case of segmentation which is central to medical image analysis in research and primordial to quantify the severity of the pathology in clinical practice. We propose a novel task-driven bidirectional fault-aware adversarial network (which we term T-GAN) that is able to learn the cross-modality transformation in an unsupervised manner for 3D medical images. In the absence of any paired data, we define a jointly-adapted bidirectional loss consisting of three sub-components: conditional dual mapping, volumetric cycle-consistency and domain specific-distance. The presented loss function utilizes the deep features of hidden representation to penalize domain discrepancy while ensuring the bidirectional mappings modeled in a closed loop while using the criterion of cycle-consistency to improve data variations within multiple conditions. To achieve not only visually-realistic synthesis but also task-practicability (i.e., the synthesized results can substitute the real acquisitions for the optional post-processing clinically and technically), we complement the bidirectional discriminators with a constraint that balances both synthesis fidelity and task performance. Rather than conditioned on the task-specific descriptions, the proposed fault-aware discriminator is suitable to improve the performance of the chosen task while preserving strong visual effects.

The contribution of the proposed method (T-GAN) is fourfold:

1. We propose to solve the cross-modality synthesis problem driven by any possible task in volumetric neuroimaging under an unsupervised setting.

2. Through combining both adversarial losses of dual mappings, the cycle-consistency loss in volumetric space and domain adapted loss, we propose the jointly-adapted bidirectional loss as an entire loss function for solving the problem of acquired medical images under multiple conditions.

3. We present a fault-aware discriminator to improve the performance of any post-processing from the synthesized results using a combination of bidirectional adversarial losses conditioned on the task performance.

4. We evaluate the proposed approach on several cross-modality synthesis tasks from different datasets involving both healthy and pathological cases and show the generative network can synthesize both visually desirable modality data and practically beneficial to medical image analysis.

6.2 Related Work

6.2.1 Cross-Modality Synthesis

Cross-modality synthesis, otherwise known as image style transfer (also called image-to-image translation), can be defined as synthesizing the target modality data from the representations of the source images. A common strategy is to learn the underlying relationship between both domains and predict the desirable data in one modality conditioned on data in another modality. Examples of pioneering works include the classic texture synthesis [32, 33, 83] relying on the non-parametric settings to resample the feature statistics of a given source texture. More advanced image synthesis methods such as neighbor embedding-based image super-resolution, eigentransformation-based face sketch synthesis [143] and image hallucination [157] assumed that images in source and target spaces have similar local geometrics or representations, such that the target image can be expressed as a linear combination of its neighbors or training samples. While achieving remarkable results at that time, the assumption is hardly held for images which are highly dimensional, statistically non-Gaussian, texturally rich and varied, especially for the nonlinear relationship between both modalities. To find alternative solutions, the strategy of projecting images to the feature spaces was explored [60, 70, 91], which has achieved great results for many image synthesis applications. They mapped cross-modality data into either coupled or common subspace to associate images between both domains, also learned a mapping function for the parametric image transformation.

Deep learning [84] has enabled substantial progress in the problem of generating missing/target modality data due to the very impressive results obtained. Gatys *et al.* [44] first exploited a pre-trained Convolutional Neural Network (CNN) to extract features from both superficial modalities and substantial contents of images for texture synthesis. This method investigated images based on their high-level representations derived from CNNs, and generated new images by seeking the matched feature representations close to the input one while providing a correlation-map for emphasizing the domain information. Other approaches, for example, learning the capability of deep convolutional decoder networks for synthesizing 3D chair renderings [31], inverting the image representations of certain layers to learn the degrees of the geometric and photometric invariance about the image [101], and training a compact feed-forward generative network for image stylization [149], extended the diversity of image synthesis considered in this filed.

More recently, Generative Adversarial Networks (GANs) [46] have shown very promising performance in the cross-modality image generation. The key contribution of GANs is that two neural networks are trained simultaneously, i.e., a generative network captures the distributions of real data for generating visually similar images, and a discriminative network evolves to distinguish the generated images from real ones. Through learning a minimax objective between both networks, GANs are optimized to fool the adversarial discriminator into predicting whether the synthesized images are real. Built upon the original GANs' model, follow-up works, e.g. [85, 94, 122, 135, 154] have been proposed for a wide variety of cross-modality synthesis problems. Conditional GANs (cGANs) typically employed GANs to synthesize samples conditioned on either image attributions, textures or class labels. For instance, Wang *et al.* [156] proposed the style and structure GANs by factorizing the image generation process to achieve more realistic synthesis. Zhu *et al.* [188] tackled the visual manipulation problem using the image manifold-guided GAN to reformulate the traditional editing operations. Isola *et al.* [69] used a cGAN for the generic image-to-image translation problem, in which the loss function penalized arbitrary dissimilar structures between the synthesized images and real ones. In SRGAN [85], the authors exploited a well-trained VGG network [137] to form a perceptual loss function for feature-guided photo-realistic image super-resolution. Liu *et al.* proposed a CoGAN framework [94] by employing a weight-shared strategy between networks to learn the cross-modality representations.

However, the algorithms described above suffer from a limitation associated with the supervised learning. Nevertheless, lacking paired data reveals an opposite reality. Fortunately, dual learning [55] provides a new learning paradigm that forms a closed loop between the source and target domains to generate the informative feedbacks for the unpaired machine translation. DualGAN [173] benefits from dual learning and GANs, loosening the limitation on requiring paired images by constraining a dual mapping for the general image style transfer. In parallel, the cycle-consistency property was leveraged in CycleGAN [189], UNIT [93] and DiscoGAN [79] to achieve the unsupervised joint distribution matching without paired data.

Inspired by their success in computer vision, cross-modality synthesis of medical images [13, 72, 114] has been receiving significant attention in recent years. Methods that put particular focus on the dictionary learning-based common representation have been proposed [124, 127]. More approaches aim to establish a stable mapping between the source modality data and the target modality images. To effectively leverage data while avoiding explicit modeling acting on some specific cases, Ye *et al.* [172] proposed the modality propagation method for coherent synthesis of population data. Iglesias *et al.* [67] used a patch matching algorithm to synthesize the target tissue contrasts and investigated whether the synthesized results can be useful for inter-modality analysis. Van *et al.* [150] built a location-sensitive deep network by combining the intensity features from image voxels and spatial information into a deep network for the cross-modality synthesis of brain images. Unlike the supervised approaches with the benefit of paired training data, Vemulapalli *et al.* [153] proposed a general unsupervised method to capture similarities of the cross-modality image patches, followed by a joint maximization of both global mutual information and local spatial consistency. Huang *et al.* [64, 65] further improved the quality of the synthesized images across different modalities using the joint convolutional sparse coding scheme. A more recent work [17] explored the fully convolutional neural network to formulate a modality-invariant latent representation which has shown the increased performance for multi-modal MRI synthesis.

6.2.2 Medical Image Analysis

The analysis of medical images [52, 54, 182] is always required for both diagnostic and therapeutic medicine in many clinical tasks. Different imaging modalities present valuable and complementary anatomical information about the underlying tissues. Although multi-modality data can be very informative in medical imaging, acquiring them is a very time-consuming process and generally impractical. Clinically, to gather more information about anatomical structures of organs, synthesizing the desirable modality images is sometimes necessary and expected to infer an image from one modality to another for better visualization, prediction and detection purposes during post-processing. Among applications for investigating the properties of tissue organizations, image segmentation [109] is a major task. It faces a problem when acquisitions are having a variability of tissue appearance obtained from different physical principles. Therefore, efforts to tackle it can focus on the cross-modality synthesis. While an automated generation of different structures in cross-modality 3D volumes is highly desirable, it remains a problem whether the synthesized results can improve the performance of post-processing.

6.3 Task-Driven Bidirectional Fault-Aware Adversarial Networks

6.3.1 Preliminaries

Generative Adversarial Networks (GANs) [46] are composed of two models: generator G and discriminator D. G is trained to imitate the real image by mapping a latent random vector $\mathbf{z}$ sampled from the uniform noise distribution p_z to the real data distribution p_{data}. D is optimized to distinguish whether an image is the generated counterpart $G(\mathbf{z}) \sim p_g$ or the real one $\mathbf{x} \sim p_{data}$. Concretely, given the vectorized image $\mathbf{x}$, G and D are defined to solve the following adversarial minimax objective on $V(D,G)$:

$$\begin{aligned}\min_G \max_D V(D,G) = &\mathbb{E}_{\mathbf{x}\sim p_{data}}[\log D(\mathbf{x})] + \\ &\mathbb{E}_{\mathbf{z}\sim p_z}[\log(1 - D(G(\mathbf{z})))]\,.\end{aligned} \tag{6.1}$$

Eq. (6.1) can be solved in an alternative manner over generator G and discriminator D. That is, by fixing the parameters of G, we can optimize D and *vice versa*. There exists a global optimality when $p_g = p_{data}$ with a mild condition that G and D have enough capacity to make p_g converges to p_{data}.

CycleGAN [189], DualGAN [173], DiscoGAN [79] and UNIT [93] are the extensions of GANs where the cycle-consistency property $\mathscr{L}_{cyc}(G,F)$ is combined with the dual mapping functions $G : \mathbf{X} \rightarrow \mathbf{Y}$ and $F : \mathbf{Y} \rightarrow \mathbf{X}$ into the GANs' objective for unpaired cross-modality synthesis. $\mathbf{X}$ and $\mathbf{Y}$ are two sets of training samples as $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, ...\mathbf{x}_s)$ and $\mathbf{Y} = (\mathbf{y}_1, \mathbf{y}_2, ...\mathbf{y}_t)$.

6.3.2 Problem Formulation

Instead of working with 2D image stacks of original volumetric neuroimaging, we input 3D volumes (i.e., $\mathbf{X}$ and $\mathbf{Y}$) directly to ensure the intrinsic sequential information between consecutive slices. Given training samples $\mathscr{X} = \{\mathbf{X}_i\}_{i=1}^{S} \in \mathbb{R}^{m\times n\times t\times S}$ in the source domain and $\mathscr{Y} = \{\mathbf{Y}_i\}_{i=1}^{T} \in \mathbb{R}^{m\times n\times t\times T}$ in the target domain, our goal is to form a closed loop between the dual tasks, i.e. $\mathscr{X} \leftrightarrow \mathscr{Y}$ without supervision of paired examples. Here, m and n are the dimensions of the axial view of the volumetric image, t denotes the size of an image along the z-axis, while S and T are the numbers of elements in the source and target training sets respectively. As with the existing dual GANs' learning, we construct two mappings: $G : \mathscr{X} \rightarrow \mathscr{Y}$ and $F : \mathscr{Y} \rightarrow \mathscr{X}$ in the volumetric space, therefore the generation of G and F can be represented as $\hat{\mathbf{Y}} = G(\mathbf{X})$ and $\hat{\mathbf{X}} = F(\mathbf{Y})$ respectively. Two adversarial discriminators D_G and D_F are modeled to distinguish the fake products corresponding to G and F. The difficulties of our synthesis work

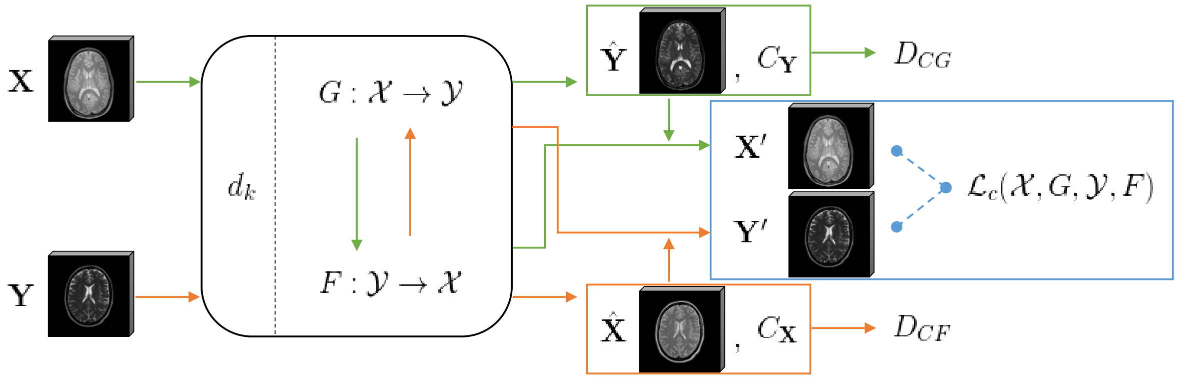


Fig. 6.1 Flowchart of the proposed method (T-GAN) for cross-modality synthesis. G and F are the dual mapping functions which are used to establish domain exchange among the source domain $\mathscr{X}$ and the target domain $\mathscr{Y}$, $\mathbf{X}$ and $\mathbf{Y}$ are the 3D volumes belongs to $\mathscr{X}$ and $\mathscr{Y}$ respectively, $\hat{\mathbf{X}}$ and $\hat{\mathbf{Y}}$ represent the first generated results while $\mathbf{X}$ and $\mathbf{Y}$ are their dual generations, D_{CG} and D_{CF} are the discriminators corresponding to G and F, $\mathscr{L}_c$ denotes the cycle-consistent GAN, d_k is the MK-MMD-based jointly-adapted regularizer, C_X and C_Y are the task-driven results.

vary with multiple conditions, e.g., the volumetric representations, an unsupervised setting, different imaging modalities, imaging angles and even systems (such as Philips, Siemens and GE). To encourage the synthesized results actually leading to the improvement of any possible post-processing, we generate visually-realistic images conditioned on performing application specific task. Therefore, we address the task-driven cross-modality synthesis by first building two GANs in a dual manner with the bidirectional loss, then minimizing the misaligned representations, while conditioning the whole model on performing the selected post-processing task. In this chapter, we attempt to synthesize MRI of human brains with segmentation. The proposed method is diagrammed in Fig. 6.1.

6.3.3 Jointly-Adapted Bidirectional Loss

To translate an image $\mathbf{X}_i$ in $\mathscr{X}$ to an image in $\mathscr{Y}$ through applying GANs' model, we can learn a function $G : \mathscr{X} \to \mathscr{Y}$ with the expected output $\hat{\mathbf{Y}}_i = G(\mathbf{X}_i)$. The generator G is then argued by a discriminator D_G giving the likelihood that the input image $\mathbf{X}_i$ has been sampled from the target domain. Similarly, to map an image $\mathbf{Y}_i$ in $\mathscr{Y}$ to an image in $\mathscr{X}$ which can be the dual task of G by training an inverse generator $F : \mathscr{Y} \to \mathscr{X}$ having $\hat{\mathbf{X}}_i = F(\mathbf{Y}_i)$ with the corresponding discriminator D_F. Following the standard GAN [46], the adversarial losses of both mapping functions are jointly expressed in the volumetric space:

$$\begin{aligned}\mathscr{L}_d(D_G, D_F, G, F) = &\mathbb{E}_{\mathbf{Y}\sim p_{data}(\mathbf{Y})}\left[\log D_G(\mathbf{Y})\right] + \\ &\mathbb{E}_{\mathbf{X}\sim p_{data}(\mathbf{X})}\left[\log(1 - D_G(G(\mathbf{X})))\right] + \\ &\mathbb{E}_{\mathbf{X}\sim p_{data}(\mathbf{X})}\left[\log D_F(\mathbf{X})\right] + \\ &\mathbb{E}_{\mathbf{Y}\sim p_{data}(\mathbf{Y})}\left[\log(1 - D_F(F(\mathbf{Y})))\right],\end{aligned} \tag{6.2}$$

where $\mathscr{L}_d$ is the dual loss. The function in Eq. (6.2) forms a simple closed loop between two losses which extends the volumetric GANs into a dual learning manner and joint representations into a unified framework. In the unsupervised dual learning problem, one typical property is to force both learnings from each other to produce the pseudo-input. This is done by generating $\mathbf{X}'$ for task $\mathscr{X} \rightarrow \mathscr{Y}$ and $\mathbf{Y}'$ for task $\mathscr{Y} \rightarrow \mathscr{X}$ respectively, where $\mathbf{X}' = F(\hat{\mathbf{Y}}) = F(G(\mathbf{X}))$ and $\mathbf{Y}' = G(\hat{\mathbf{X}}) = G(F(\mathbf{Y}))$. We note that some contemporary works [79, 93, 173, 189] infer the pseudo-input by involving the cycle-consistency constraint to regularize both mappings. Inspired by such a strategy, we enforce a volumetric cycle-consistency with GANs' model using

$$\begin{aligned}\mathscr{L}_c(\mathscr{X}, G, \mathscr{Y}, F) = &\mathbb{E}_{\mathbf{X}\sim p_{data}(\mathbf{X})}\left\|\mathbf{X} - F(G(\mathbf{X}))\right\|_1 + \\ &\mathbb{E}_{\mathbf{Y}\sim p_{data}(\mathbf{Y})}\left\|\mathbf{Y} - G(F(\mathbf{Y}))\right\|_1,\end{aligned} \tag{6.3}$$

where $\|\cdot\|_1$ means we adopt l_1 distance[1] to quantitatively compare the input data and reconstructed pseudo.

Although the modeled image distributions over the latent feature space are adopted in the unsupervised cross-modality synthesis problem, a potential assumption is implied that the representations of both modalities are almost domain invariant. Some recent approaches [97, 174] reveal a fact that deep features can disentangle explanatory factors of variations in the data distributions [6], but the cross-modality distribution discrepancy is still remnant. Motivated by the requirement of boosting unpaired cross-modality data (underlying the same distributions) to be close to each other, especially in the medical images under multiple conditions, we define a jointly-adapted regularization term that intrinsically manifests invariant structures across modalities. To achieve this effect, recent work [49, 65, 183] employed either the traditional Maximum Mean Discrepancy (MMD) or the extended Multi-Kernel MMD (MK-MMD) criterion to explore the data statistics of different domains. The jointly-adapted regularizer is proposed to release the assumption of domain invariance which was explored in [174]. We follow such an idea and make use of the MK-MMD in our method. MK-MMD is employed for the two-sample matching, along with other components of T-GAN, to align the 'real' paired data. Specifically, we use the unbiased estimate of MK-MMD to reduce the domain

[1] l_1-based loss has been explored in [173, 189] to avoid blurriness and show promising results.

discrepancy, hence our model is independent of the assumption of 'same latent variables'. This is done by adding the MK-MMD-based jointly-adapted regularizer to our bidirectional GAN model:

$$d_k(\mathscr{A}^{\mathscr{X}}, \mathscr{A}^{\mathscr{Y}}) = \left\| \mathbb{E}_{p_{data(\mathbf{X})}} \psi(\mathbf{A}^{\mathscr{X}}) - \mathbb{E}_{p_{data(\mathbf{Y})}} \psi(\mathbf{A}^{\mathscr{Y}}) \right\|^2_{\mathscr{H}_k}, \tag{6.4}$$

where d_k is interpreted as matching all orders of statistics which can be performed by stochastic gradient descent with the gradient calculated by back-propagation through the generative network, and k is the characteristic kernel defined on the vectorized element $\mathbf{A}$ combining a set of positive definite kernels $\{k_u\}_{u=1}^d$:

$$k \in \mathscr{K} := \left\{ \sum_{u=1}^{d} \beta_u k_u : \sum_{u=1}^{d} \beta_u = 1, \beta_u \geq 0, \forall u \in \{1, ..., d\} \right\}, \tag{6.5}$$

where $\{\beta_u\}_{u=1}^d$ is the coefficient for constraining the characteristic of each k_u. $\psi(\cdot)$ denotes the nonlinear mapping with $k(\mathbf{A}^{\mathscr{X}}, \mathbf{A}^{\mathscr{Y}}) = \langle \psi(\mathbf{A}^{\mathscr{X}}), \psi(\mathbf{A}^{\mathscr{Y}}) \rangle$, $\mathscr{A}^{\mathscr{X}} = \{\mathbf{A}_i^{\mathscr{X}}\}_{i=1}^S$ and $\mathscr{A}^{\mathscr{Y}} = \{\mathbf{A}_i^{\mathscr{Y}}\}_{i=1}^T$ are the deep features of $\mathbf{X}_i$ and $\mathbf{Y}_i$ for source and target domains respectively, and $\mathscr{H}_k$ indicates the reproducing kernel Hilbert space (RKHS) induced by k and ψ.

We then integrate the volumetric cycle-consistency and joint adaptation into Eq. (6.1), yielding the proposed jointly-adapted bidirectional loss:

$$\begin{aligned} \mathscr{L}_b(\mathscr{X}, \mathscr{Y}) = & \mathscr{L}_d(D_G, D_F, G, F) + \\ & \mathscr{L}_c(\mathscr{X}, G, \mathscr{Y}, F) + d_k(\mathscr{A}^{\mathscr{X}}, \mathscr{A}^{\mathscr{Y}}). \end{aligned} \tag{6.6}$$

6.3.4 Fault-Aware Discriminator

Rather than only encouraging the visually-realistic synthesis to approximate the ground truth in an unsupervised bidirectional learning manner, we instead encourage them to have better performance in an optional post-processing as used by the real image. As we introduced in Section 6.3.3, jointly-adapted bidirectional loss penalizes both generators and domain discrepancy in the latent space, while the corresponding discriminators simply distinguish whether the generation is real or fake. To ensure the synthesized results can satisfy the requirement of later processing, e.g., image segmentation or registration, and substitute the real acquisitions for medical image analysis, we need to bridge the gap of task performance between the distributions of generated and real ones. The ideal discriminator has an explicit notion that whether the image is real or synthesized and also match the task performance. To achieve this, we propose the fault-aware discriminator to meet the additional condition on the selected task.

Given the solution C[2] of an arbitrary task, and the input data $\mathbf{X}$, the result of $\mathbf{X}$ for the chosen task can be denoted as $C_{\mathbf{X}}$. More precisely, the responsibility of the proposed discriminator is to judge two major tasks (i.e., the performance of synthesis and possible task) derived an obvious discriminant: real data & right performance. Similar concept has been explored in some recent works [122, 177] conditioned on the embedded text for text-to-image translation. Instead of relying on the task-specific (e.g. text, label and even image) descriptions, we build on the ideas but in theory more general and suitable for an application. Conditioned on the result of C, the mapping functions $G : \mathscr{X} \to \mathscr{Y}$ and $F : \mathscr{Y} \to \mathscr{X}$ are updated to $G : \{\mathscr{X}, C\} \to \mathscr{Y}$ and $F : \{\mathscr{Y}, C\} \to \mathscr{X}$ with the corresponding outputs $\hat{\mathbf{Y}} = G(\mathbf{X}, C)$ and $\hat{\mathbf{X}} = F(\mathbf{Y}, C)$, respectively. The fault-aware discriminators $\mathscr{L}_{D_{CG}}$ and $\mathscr{L}_{D_{CF}}$ for the mappings G and F are then expressed as

$$\begin{aligned}
\mathscr{L}_{D_{CG}} =& \mathbb{E}_{(\mathbf{Y},C)\sim p_{data}(\mathbf{Y})}\left[\log D_{CG}(\mathbf{Y},C)\right] + \\
& \mathbb{E}_{(\mathbf{X},C)\sim p_{data}(\mathbf{X})}\left[\log(1 - D_{CG}(G(\mathbf{X}),C))\right], \\
\mathscr{L}_{D_{CF}} =& \mathbb{E}_{(\mathbf{X},C)\sim p_{data}(\mathbf{X})}\left[\log D_{CF}(\mathbf{X},C)\right] + \\
& \mathbb{E}_{(\mathbf{Y},C)\sim p_{data}(\mathbf{Y})}\left[\log(1 - D_{CF}(F(\mathbf{Y}),C))\right].
\end{aligned} \tag{6.7}$$

We can then update the parameters of the discriminative network by minimizing the above losses in Eq. (6.7).

6.3.5 Objective Function

Essentially, the properties of dual mapping, visual similarity and task performance form the whole adversarial loss that enforces associations between similar contents conditioned on an extra fault (i.e. the quantification of an optional application) of both domains. We thus update the adversarial loss in Eq. (6.2) based on the fault-aware discriminator, which is given below:

$$\begin{aligned}
&\mathscr{L}_{cd}(D_{CG}, D_{CG}, G, F) = \\
&\mathbb{E}_{(\mathbf{Y},C)\sim p_{data}(\mathbf{Y})}\left[\log D_{CG}(\mathbf{Y},C)\right] + \\
&\mathbb{E}_{(\mathbf{X},C)\sim p_{data}(\mathbf{X})}\left[\log(1 - D_{CG}(G(\mathbf{X}),C))\right] + \\
&\mathbb{E}_{(\mathbf{X},C)\sim p_{data}(\mathbf{X})}\left[\log D_{CF}(\mathbf{X},C)\right] + \\
&\mathbb{E}_{(\mathbf{Y},C)\sim p_{data}(\mathbf{Y})}\left[\log(1 - D_{CF}(F(\mathbf{Y}),C))\right].
\end{aligned} \tag{6.8}$$

where $\mathscr{L}_{cd}$ is the conditional dual mapping loss. Correspondingly, the overall loss function for the proposed model can be further updated as $\mathscr{L}_b(\mathscr{X}, \mathscr{Y}) = \mathscr{L}_{cd}(D_{CG}, D_{CF}, G, F) +$

[2] C can be treated as performing any task measured by an appropriate quantification such as dice score.

$\mathscr{L}_c(\mathscr{X},G,\mathscr{Y},F)+d_k(\mathscr{A}^{\mathscr{X}},\mathscr{A}^{\mathscr{Y}})$. Our optimization objective then becomes:

$$\begin{aligned}\min_{G,F}\max_{D_{CG},D_{CF}}\ & \mathscr{L}_{cd}(D_{CG},D_{CF},G,F)+ \\ & \delta\mathscr{L}_c(\mathscr{X},G,\mathscr{Y},F)+\gamma d_k(\mathscr{A}^{\mathscr{X}},\mathscr{A}^{\mathscr{Y}}).\end{aligned} \tag{6.9}$$

where δ denotes a balance coefficient γ for the cycle-consistency loss $\mathscr{L}_c$ and γ is the trade-off parameter for the jointly-adapted penalty. In Eq. (6.9), as with the conventional two-player minimax problem, we propose to train the entire model by alternatively maximizing discriminators D_{CG}, D_{CF} and minimizing a combination of conditional dual mapping loss $\mathscr{L}_{cd}$, cycle-consistency loss $\mathscr{L}_c$ and domain RKHS-distance d_k in the volumetric space.

6.4 Implementation

6.4.1 Network Structures

The proposed method (T-GAN) roughly follows the network architectures set forth by [74, 119] for the generative networks. We construct T-GAN inspired from [173, 189] including both identical configurations for the generators. In addition to the bidirectional framework, similar to [163], 3D-GAN is adopted for generating the volumetric neuroimage to solve the discontinuity problem against stacked 2D slices.

The generator in T-GAN consists of 3 convolutional layers with strides 1, 2 and 2 as the front-end, 6 residual blocks [56], 2 fractionally-strided convolutions with same stride 1/2, and 1 convolutional layer as the back-end with stride 1. Following the settings in [74], the general convolutional layers with spatial batch normalizations and ReLU nonlinearity are added in between (i.e., formed as convolution-BatchNorm-ReLU), while the output layer applies the tanh activation at the end. The 6 residual blocks each includes 2 convolutional layers with the fixed 128 filters on both layers. Particularly, we use the 7×7×7 volumetric kernels for the first and last layers while using the 3×3×3 volumetric kernels for the remaining layers.

For the discriminative network, we adopt the Markovian PatchGAN previously explored in [69, 88, 173] but mirrors the generator in the volumetric space in this work. Instead of modeling the full image-sized discriminator, Markovian PatchGAN effectively models an image as a Markov random field in local image patches for distinguishing whether the selected size of patch in an image is real or fake. Such a configuration is effective especially in the very large-sized images and 3D volumes since it contains parameters fewer than others. We fix the patch size as 70×70×70 in the overlapped manner and use the stack of convolution-BatchNorm-Leaky ReLU layers (i.e., instead of using ReLU, Leaky ReLU activation is applied here) to train the discriminative network. Similar to [55, 69, 173], our discriminator is run convolutionally across the volumes and finally averaging all responses to give the ultimate results.

6.4.2 Training Details

To train the proposed method (T-GAN), a common procedure is to take alternating steps of updating the generator and discriminator in every batch. We set the learning rate of 0.0002, employ the Stochastic Gradient Descent with mini-batch of size 1 and apply the Adam solver [80] for optimization. Empirically, to control the influence between G and F, we follow [97, 189] and set the balance coefficient δ =10, the trade-off parameter γ =0.3. However, with 3D generation, a potential problem can be exposed that synthesizing 3D voxels is harder than differentiating between the synthesized result and ground truth, easily leading to faster learning progress of the discriminator than the generator. The proposed discriminator is conditioned on the extra criteria that the performance of both synthesis and its application need to be balanced and satisfied. As the suggestion in [163], to make the training of both generator and discriminator in pace meanwhile taking account of the task performance, we update the discriminator when its accuracy and dice scores [3] in the last batch are less than 0.8.

6.5 Experiments

6.5.1 Experimental Setup

Datasets

Our T-GAN is evaluated on two brain datasets: IXI [4] and NAMIC Multimodality [5] datasets. The IXI dataset involves 578 healthy subjects, each was imaged using a matrix of $256\times256\times p$ scanned with either Philips 3T system, Philips 1.5T system, or GE 3T system having $0.94\times0.94\times1.2$ mm^3 voxel dimensions. The NAMIC dataset includes 20 subjects (10 normal controls and 10 schizophrenics), each was imaged using a matrix of $128\times128\times q$ scanned with a 3T GE system having 1x1x1 mm^3 voxel dimensions.

[3] The Dice overlap is commonly used as a proxy for evaluating the quality of segmentation and registration for medical image analysis.

[4] http://brain-development.org/ixi-dataset/

[5] http://hdl.handle.net/1926/1687

Experiments

We evaluate our method in two scenarios: (1) synthesizing the T2-w images from the PD-w acquisitions and *vice versa* in the IXI dataset, (2) generating the T1-w images from the T2-w inputs and *vice versa* in the NAMIC dataset. However, we note that only a few groups of unpaired data are available in the NAMIC dataset. To deal with this problem, the images in the IXI dataset are scaled to $128 \times 128 \times p$ voxels for extending the size of training data in this scenario. For quantitative evaluation, we perform two-fold cross-validation to test our method. By selecting 230 unpaired Proton Density-weighted (PD-w) and T2-weighted (T2-w) MRI scans from the IXI dataset and 7 unpaired T1-w, T2-w acquisitions form the NAMIC dataset for training, while the remaining data i.e., 118 (IXI) and 6 (NAMIC) for testing. In the segmentation experiments, we feed the real scans and the synthesized results to the fully-automatic FreeSurfer (i.e., a structural MRI analysis software with default parameters) [38] pipeline. It should work independently of the primary training procedure. Once the regular training (i.e., GAN) has done, we input the first generated result into the task, e.g. segmentation, and collect the segmented results to 'reboot' our T-GAN for judging two major tasks (synthesis and segmentation) derived two discriminants. The segmentation (carried out by FreeSurfer) produces several major brain tissue classes, i.e., Cerebrospinal fluid (CSF), (GM) and (WM), yielding the averaged quantification of a whole brain volume. The tissue prior probability templates used in FreeSurfer are based on averaging multiple automatically segmented images in standard space of images from either the IXI or the NAMIC dataset in this chapter, so there is no guarantee that CSF, GM and WM classes will exactly follow other methods. In addition, this work provides a reliable synthesis mechanism for generating both visually-realistic and task-effective products, and we focus on our synthesized results in these experiments.

Baselines

We compare our results against several state-of-the-art cross-modality synthesis algorithms including MR image exampled-based contrast synthesis (MIMECS) [127], Vemulapalli's supervised cross-modal synthesis of subject-specific scans (V-s) [153], Vemulapalli's unsupervised synthesis method (V-us) [153], dual convolutional filter learning (DOTE) [64], GAN [46] and CycleGAN [189]. Specifically, MIMECS, V-s, DOTE and GAN have supervision in form of paired data, while V-us, CycleGAN and our method are unsupervised. The implementations are mostly from the available codes provided by the authors, except Vemulapalli's methods that the results are provided by the authors and only available on the NAMIC dataset. Moreover, following [64, 127, 153], the brain MRI scans are skull-stripped, bias-field corrected and linearly registered. To demonstrate the robustness of our method, we skip the skull stripping

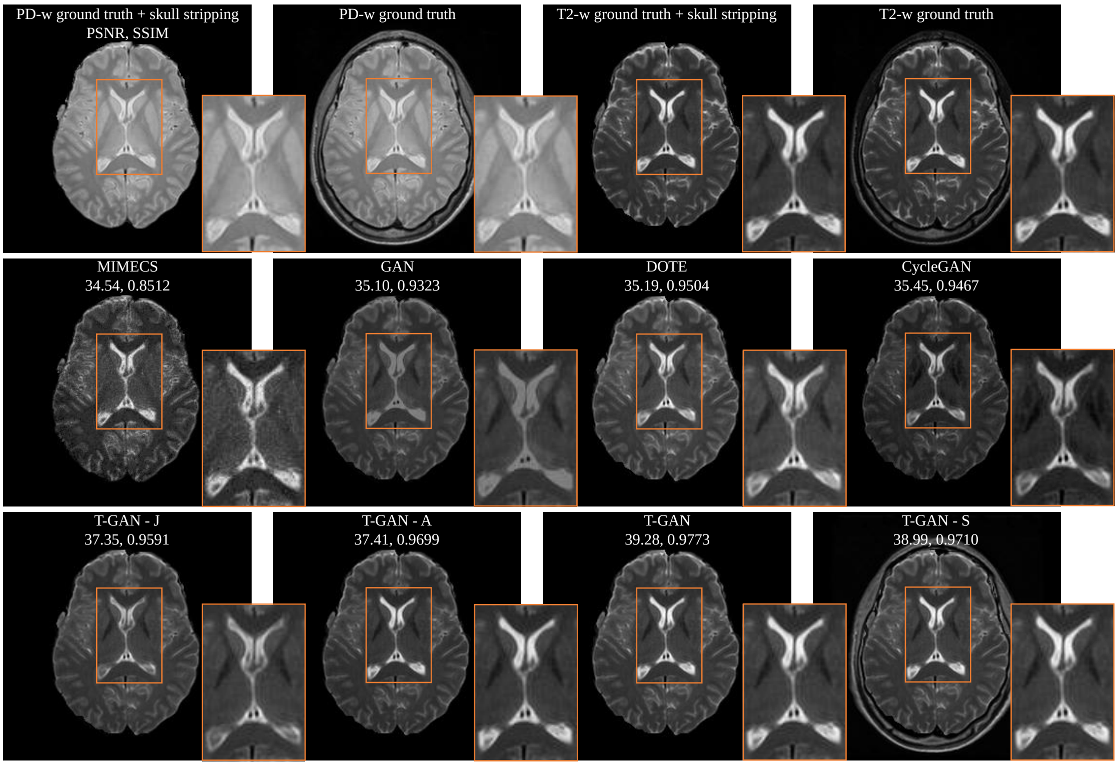


Fig. 6.2 Visual comparison for PD-w → T2-w with/without skull stripping MRI brain image cross-modality synthesis.

processing and show the quantitative results along with others. Without skull stripping, we denote our method here as T-GAN−S (without skull stripping). We also perform an ablation study measuring the impact of the jointly-adapted penalty and task-driven condition by separately removing the relative term, in which our method can be written as T-GAN−J (without jointly-adapted penalty), T-GAN−T (without task-driven condition).

Quantitative Evaluation

For the evaluation criteria, we adopt PSNR, SSIM indices and Dice overlap to objectively assess the quality of the synthesized results and the use of generations on segmentation. Besides the widely used PSNR and SSIM, the Dice overlap is also a well-known volume metric for comparing the quality of two binary label masks. Mathematically, given the mask of ground truth as the reference plane $\mathscr{M}_g$ and the mask generated by the synthesized product $\mathscr{M}_s$, the Dice overlap can be defined as $2\frac{\mathscr{M}_g \cap \mathscr{M}_s}{|\mathscr{M}_g|+|\mathscr{M}_s|}$.

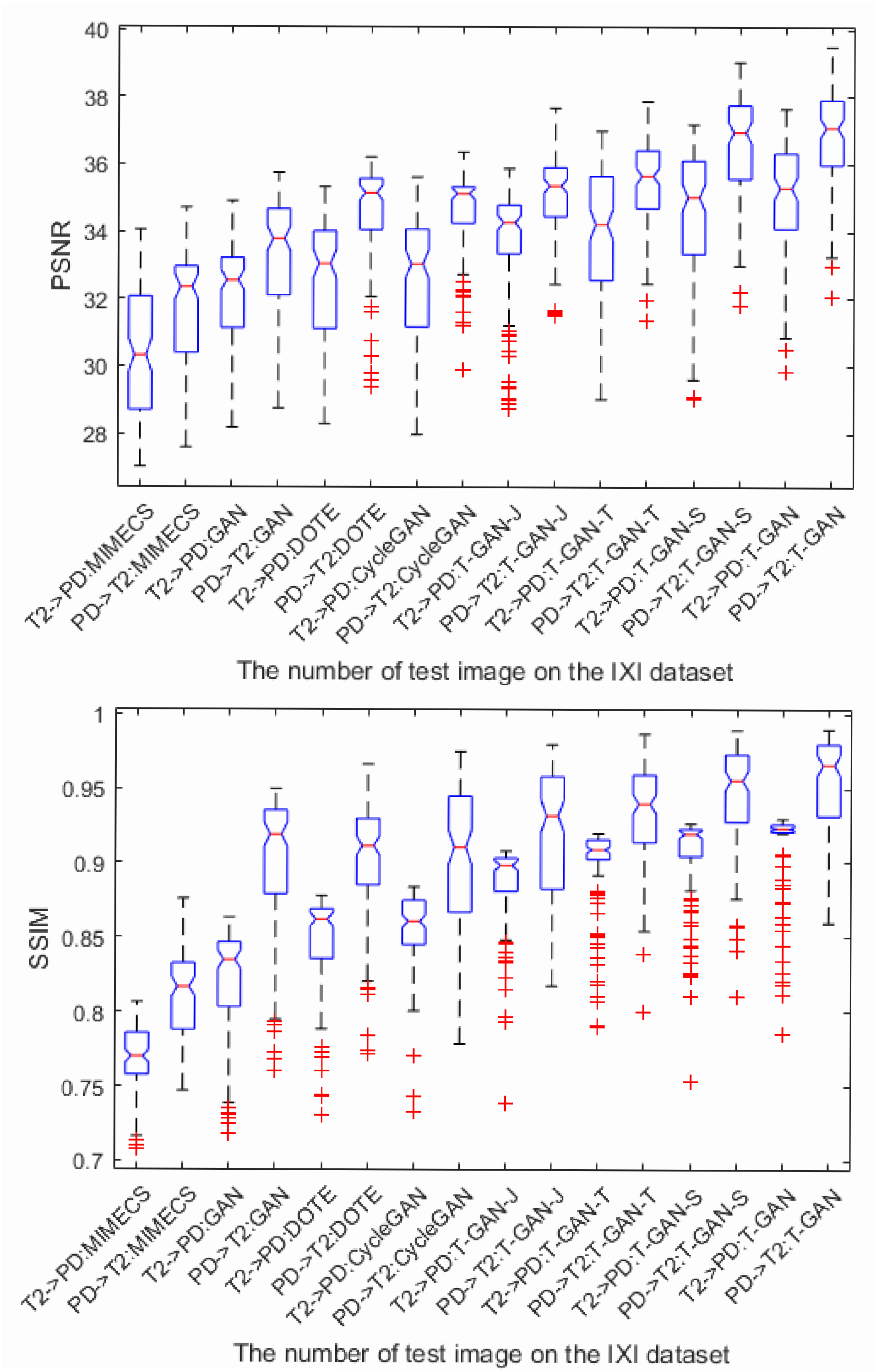


Fig. 6.3 Quantitative performance of PD-w → T2-w with/without skull stripping MRI brain image cross-modality synthesis.

Metric(avg.)	MIMECS [127]	GAN [46]	DOTE [64]	CycleGAN [189]	T-GAN−J	T-GAN−T	T-GAN−S	T-GAN
IXI: T2-w → PD-w								
PSNR (dB)	30.24	32.07	32.49	32.59	33.72	33.97	34.56	**34.99**
SSIM	0.7675	0.8219	0.8493	0.8563	0.8861	0.8994	0.9059	**0.9131**
IXI: PD-w → T2-w								
PSNR (dB)	31.69	33.32	34.60	34.66	35.06	35.55	36.64	**36.93**
SSIM	0.8126	0.9001	0.9016	0.9026	0.9206	0.9352	0.9468	**0.9517**
Dice Score	79.97%	69.53%	82.11%	81.57%	85.21%	83.06%	86.35%	**89.81%**

Table 6.1 Quantitative evaluation (PSNR (dB) and SSIM): T-GAN vs. other synthesis methods on the IXI dataset.

Metric(avg.)	MIMECS [127]	GAN [46]	DOTE [64]	CycleGAN [189]	V-us [153]	V-s [153]	T-GAN
NAMIC: T2-w → T1-w							
PSNR (dB)	24.98	25.94	31.99	29.69	27.30	29.44	**33.31**
SSIM	0.8733	0.9274	0.9405	0.9379	0.9081	0.9276	**0.9510**
NAMIC: T1-w → T2-w							
PSNR (dB)	23.52	24.93	29.99	28.73	26.60	27.81	**30.37**
SSIM	0.8788	0.9004	0.9060	0.9020	0.8857	0.8923	**0.9141**

Table 6.2 Quantitative evaluation (PSNR (dB) and SSIM): T-GAN vs. other synthesis methods on the NAMIC dataset.

6.5.2 Segmentation-Driven Synthesis

To quantitatively evaluate both visual quality of the synthesized results and the segmentation performance comparing with ground truths, also explore the generality of the proposed T-GAN, we test on many tasks using two independent datasets. For brevity, we refer to different synthesis tasks as (1) PD-w → T2-w, (2) T2-w → PD-w, (3) T1-w → T2-w, (4) T2-w → T1-w, in which (1-2) are conducted on the IXI dataset corresponded to the first scenario and (3-4) are explored on the NAMIC dataset corresponded to the second scenario. We demonstrate both visual and quantitative results in Figs. 6.2, 6.3 and Table 6.1 for tasks (1-2), and in Figs. 6.4, 6.5 and Table 6.2 for tasks (3-4), respectively. Our algorithm consistently yields the best results against the state-of-the-art supervised and unsupervised cross-modality synthesis methods for two datasets. Specifically, Table 6.1 shows the quantitative results on the averaged PSNRs, SSIMs and Dice scores. Numbers in boldface indicates the best performance which validated our claim of synthesizing visual better results while these results can effectively improve the segmentation performance through the proposed method.

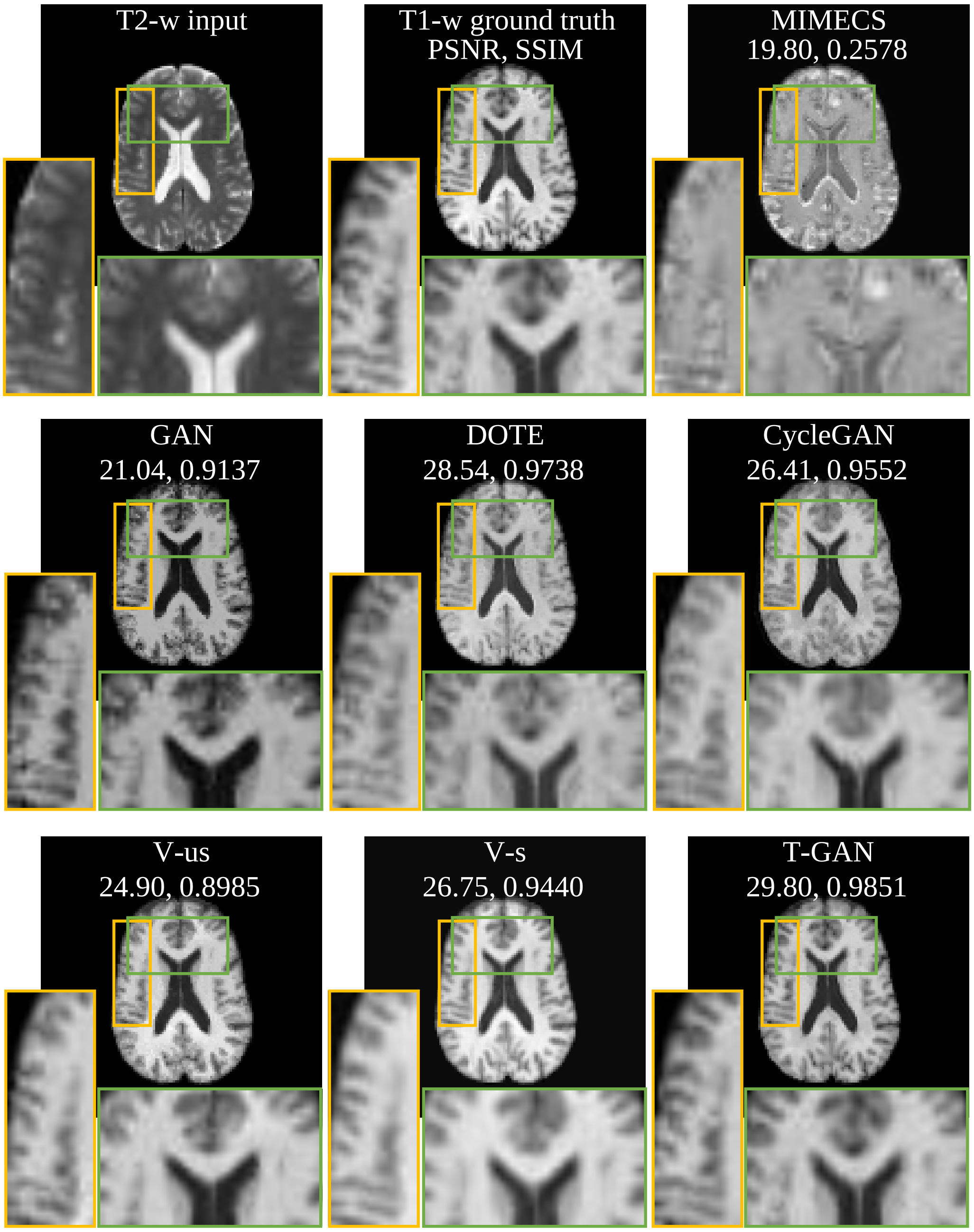


Fig. 6.4 Visual comparison for PD-w → T2-w with/without skull stripping MRI brain image cross-modality synthesis.

6.6 Conclusions

In this chapter, we proposed the task-driven bidirectional fault-aware adversarial networks (T-GAN) for the cross-modality synthesis of 3D brain images. Our framework is based on

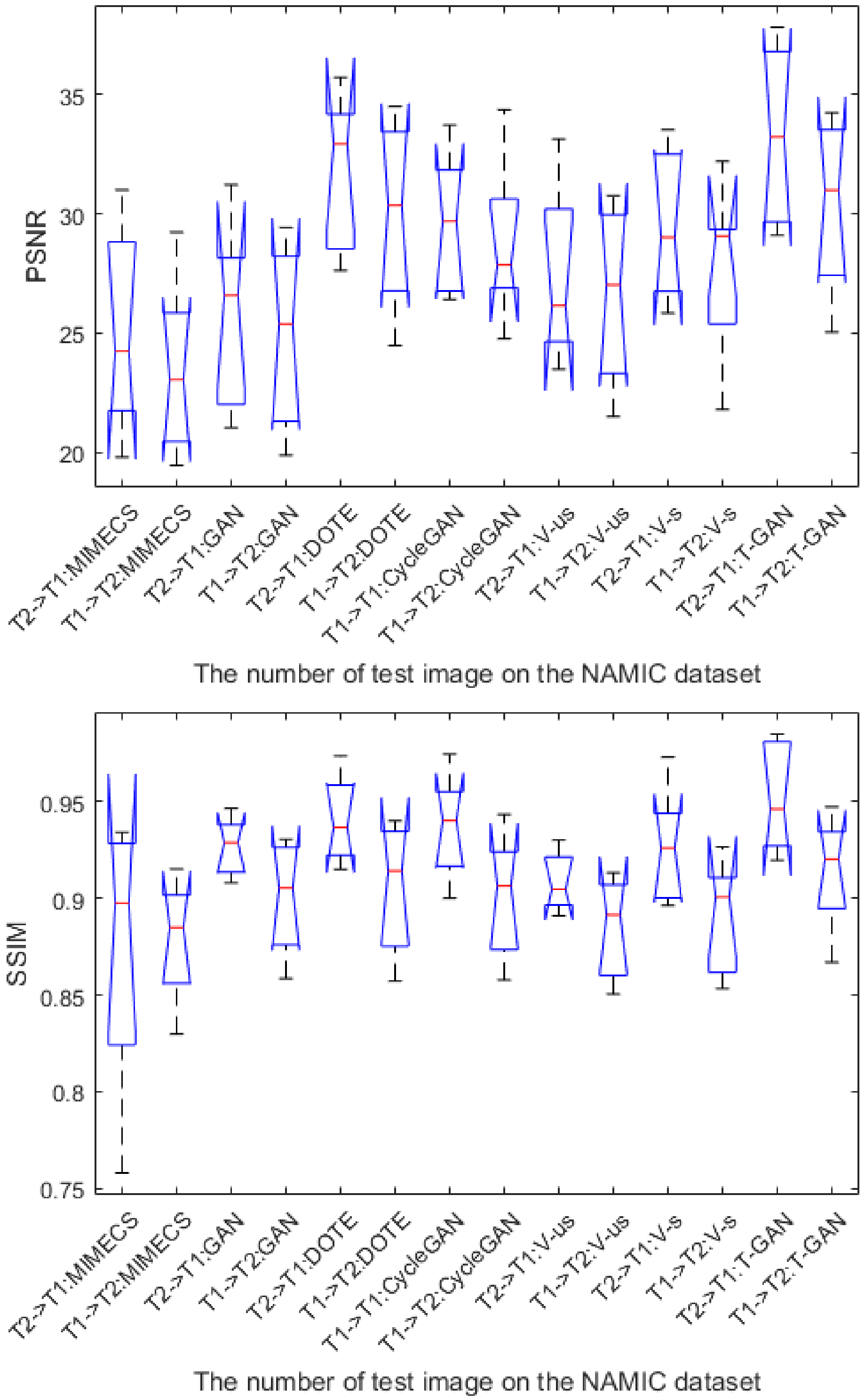


Fig. 6.5 Quantitative analysis of T2-w → T1-w MRI brain image cross-modality synthesis on the NAMIC dataset.

GANs' model using high-level representations over both generative and discriminative networks across image spaces. To learn the mapping relations between both modalities, we exploited the jointly-adapted bidirectional loss for simultaneously minimizing the domain divergence and providing volumetric consistency mappings in a dual manner. Imposing the task-effectiveness constraint in the synthesis provides a segmentation that achieves better results with visual interaction compared to a separate synthesis and segmentation process. In future work, we hope to explore the use of T-GAN for other medical image synthesis tasks such as the more challenging 4D cardiac images.

Chapter 7

Conclusions and Future Directions

7.1 Conclusions

This dissertation has focused on the problem of learning cross-modality features for three-dimensional brain image synthesis. Specifically, we have addressed two major problems including image super-resolution and cross-modality synthesis. We first explored the supervised models which rely on a large number well-processed (*i.e.*, skull-stripped and strictly registered) multi-modality brain images. In contrast to the fully supervised methods, we also learn from a few registered multi-modality image pairs which are weakly supervised. We attempt to resolve the modality divergence associated with weak pairs through constructing a common space processed by the rigorous mathematical constraints on the features. We demonstrated that the resulting approaches for the general cross-modality synthesis task could be trained in a weakly-supervised manner. Further, we addressed the problem of unsupervised cross-modality synthesis by exploring the task-driven deep architecture for volumetric neuroimage processing.

Concretely, in Chapter 2 we developed a region-character clustering model that can match cross-modality images in the sparse feature space. This allowed us to synthesize the desired modality data with normalized domain specific property. We represented a general approach for bringing transitional independent sparse representations of two domains for cross-modality image synthesis into joint learning, leveraging freely available pairwise data for training our model. This method projected the sparse codes of both domains into a common feature space, which was used to describe the correspondence between two modalities and generate the normalized target modality data via inputting the source modality image.

In Chapter 3, we addressed the problem of cross-modality synthesis for MR brain images. We introduced an approach which constructs a closed loop joint filter learning strategy to generate informative feedback for model self-optimization. The proposed method can efficiently leverage data and therefore reducing the size of the required training set. We presented dual learning and interaction based features which are informative about modality characteristics. The features were used in a convolutional sparse coding space where the role assignments and modality-specific feature weights were jointly inferred through a dual convolutional filter learning algorithm. The proposed method was evaluated against multiple baselines in two challenging tasks and demonstrated the superior performance over other state-of-the-art approaches.

In Chapters 4 and 5 we presented two models for automatically learning different modality features for image super-resolution and cross-modality synthesis. Both works had the benefits of improved synthesis performance via hetero-domain alignment without requiring fully paired training data, making synthesis more flexible. The proposed models can take a source modality image and generate the corresponding target modality image without being constrained to merely training from a finite collection of strictly pairwise data pre-processed by an extra tool (*e.g.*, FSL, AFNI, and ANTs for brain image registration). Our models generate both visually and quantitatively impressive results, and we showed in experimental evaluations that the proposed models outperform previous cross-modality synthesis methods based on the ranking results.

Finally, in Chapter 6, we presented an image synthesis model even more scalable by removing the requirement of either fully or weakly-supervised setting at training stage. This model made use of freely and publicly available brain dataset from the web, realizing that data is fundamentally a limiting factor in learning-based methods for cross-modality synthesis. As a result, we introduced the task-driven bidirectional fault-aware adversarial networks for volumetric neuroimage processing in an unsupervised manner. This method is related to the well-studied problem of the adversarial model in computer vision and machine learning. In particular, we defined a jointly-adapted bidirectional loss which utilizes the deep features of hidden representation to penalize domain discrepancy. The bidirectional mappings were modeled in a closed loop with the criterion of cycle-consistency to improve data variations within multiple conditions. We also introduced the fault-aware discriminator with a constraint on the task-specific descriptions to balance both synthesis fidelity and task performance. On the algorithmic side, our approach is suitable to improve the performance of the chosen task while preserving strong visual effects.

7.2 Future Works

Due to the large amount of progress on image synthesis made in recent years, caused only partially by the contributions included in this dissertation, it is challenging to say that some future work can be directed towards achieving better performance in brain image synthesis.

Nonetheless, based on the potential extensions of our works, there are some open questions lead for future research.

7.2.1 Discrimination Capability

Cross-modality synthesis has thus far focused on the missing/target data reconstruction problem. What advances need to be made to push only cross-modality synthesis into other problems, *e.g.* segmentation, detection of lesion area, diagnosis, or prognosis. The key limiting factor for these problems is the discrimination capability of the learned model. Most existing approaches treat the learning procedure and classifier training as two separate processes since it is easier to consider a one-way operation rather than mixing reconstructive and discriminative formulation into a unified whole. Some previous methods which are patch-based mechanism learn a dictionary and a classifier simultaneously, however, they might not scale well to a large number of data/classes. Learning a discriminative model while guaranteeing the modality-specific features tend to be required. Thus, one interesting research direction along these lines is to combine cross-modality synthesis with image classification, and doing so with transfer learning or domain adaption strategy.

7.2.2 Conditional GAN

GAN learns a generative network jointly with a corresponding discriminative adversarial network in a mini-max objective, which enables a wide variety of applications, particularly in image synthesis task. Unfortunately in practice, GAN does not provide a stable training process, and thus people suggested that adversarial training might be unstable and prone to failure even for single-modality image reconstruction problem. As GAN continues to improve, some works explore GAN in the conditional setting making the yielded conditional GAN (cGAN) suitable for both image reconstruction and image-to-image translation tasks. It makes sense to start reasoning about even more complex cases since cGAN can be extended in a principled way. While this has been partially addressed in Chapter 6, more work remains, such as leveraging the perceptual similarity between the source and target modality images with GAN to produce high-quality results. On the other hand, although the proposed method introduced in Chapter 6 can achieve compelling results in cross-modality synthesis, the results are far from uniformly positive. A more general and promising demonstration can be explored to perform cross-modality synthesis between an inexpensive modality and one that is expensive either monetarily or concerning exposure to ionizing radiation such as CT or PET. Therefore, handling more varied and extreme synthesis problems, especially a great changes in geometry, is an interesting direction for our future work.

7.2.3 4D Cardiac Data

A limitation of the proposed synthesis method is its flexibility to handle different subjects (*e.g.* brain, bone, and heart), and complicated situations (*e.g.* abnormalities affecting the morphology). As no close match exists in the currently explored datasets, the synthesis of various subjects is limited and cannot be compensated by the utilized datasets. In this situation, finding several appropriate datasets is preferential. Fortunately, UK Biobank[1] as a national and international health resource provides unparalleled research opportunities for improving the prevention, diagnosis, and treatment of a wide range of serious and life-threatening illnesses. A simple solution can then be to apply the proposed synthesis method to these subjects. However, it is worth noting that none of the sequence data (like 4D cardiac images) is used in this thesis as an input for the synthesis purpose. We can potentially design a method which can act on the four-dimensional sequences to generate good quality data while providing the complementary/required modality information.

[1] http://www.ukbiobank.ac.uk/

References

[1] Aharon, M., Elad, M., and Bruckstein, A. (2006). K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on Signal Processing*, 54(11):4311–4322.

[2] Akhtar, N., Shafait, F., and Mian, A. (2014). Bayesian sparse representation for hyperspectral image super resolution. *IEEE Transactions on Geoscience and Remote Sensing*, 52(3):1693–1704.

[3] Ashikhmin, N. (2003). Fast texture transfer. *IEEE Computer Graphics and Applications*, 23(4):38–43.

[4] Bahrami, K., Shi, F., Zong, X., Shin, H. W., An, H., and Shen, D. (2015). Hierarchical reconstruction of 7t-like images from 3t mri using multi-level cca and group sparsity. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 659–666. Springer.

[5] Belkin, M. and Niyogi, P. (2001). Laplacian eigenmaps and spectral techniques for embedding and clustering. In *Neural Information Processing Systems*, volume 14, pages 585–591.

[6] Bengio, Y., Courville, A., and Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828.

[7] Bogunović, H., Pozo, J. M., Villa-Uriol, M. C., Majoie, C. B., van den Berg, R., van Andel, H. A. G., Macho, J. M., Blasco, J., San Román, L., and Frangi, A. F. (2011). Automated segmentation of cerebral vasculature with aneurysms in 3dra and tof-mra using geodesic active regions: an evaluation study. *Medical Physics*, 38(1):210–222.

[8] Borgwardt, K. M., Gretton, A., Rasch, M. J., Kriegel, H.-P., Schölkopf, B., and Smola, A. J. (2006). Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics*, 22(14):e49–e57.

[9] Boyd, S. and Vandenberghe, L. (2004). *Convex optimization*. Cambridge university press.

[10] Bristow, H., Eriksson, A., and Lucey, S. (2013). Fast convolutional sparse coding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 391–398.

[11] Bruckstein, A. M., Donoho, D. L., and Elad, M. (2009). From sparse solutions of systems of equations to sparse modeling of signals and images. *SIAM Review*, 51(1):34–81.

[12] Buades, A., Coll, B., and Morel, J.-M. (2010). Image denoising methods. a new nonlocal principle. *SIAM Review*, 52(1):113–147.

[13] Burgos, N., Cardoso, M. J., Modat, M., Pedemonte, S., Dickson, J., Barnes, A., Duncan, J. S., Atkinson, D., Arridge, S. R., Hutton, B. F., et al. (2013). Attenuation correction synthesis for hybrid PET-MR scanners. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 147–154. Springer.

[14] Burgos, N., Cardoso, M. J., Thielemans, K., Modat, M., Pedemonte, S., Dickson, J., Barnes, A., Ahmed, R., Mahoney, C. J., Schott, J. M., et al. (2014). Attenuation correction synthesis for hybrid pet-mr scanners: application to brain studies. *IEEE Transactions on Medical Imaging*, 33(12):2332–2341.

[15] Cao, T., Zach, C., Modla, S., Powell, D., Czymmek, K., and Niethammer, M. (2014). Multi-modal registration for correlative microscopy using image analogies. *Medical Image Analysis*, 18(6):914–926.

[16] Chang, H., Yeung, D.-Y., and 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*, volume 1, pages I–I. IEEE.

[17] Chartsias, A., Joyce, T., Giuffrida, M. V., and Tsaftaris, S. A. (2017). Multimodal mr synthesis via modality-invariant latent representation. *IEEE Transactions on Medical Imaging*.

[18] Chen, B., Polatkan, G., Sapiro, G., Blei, D., Dunson, D., and Carin, L. (2013). Deep learning with hierarchical convolutional factor analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1887–1901.

[19] Chen, L., Li, W., and Xu, D. (2014). Recognizing rgb images by learning from rgb-d data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1418–1425.

[20] Chen, S. S., Donoho, D. L., and Saunders, M. A. (2001). Atomic decomposition by basis pursuit. *SIAM Review*, 43(1):129–159.

[21] Cheng, L., Vishwanathan, S. N., and Zhang, X. (2008). Consistent image analogies using semi-supervised learning. In *Computer Vision and Pattern Recognition (CVPR), 2008 IEEE Conference on*, pages 1–8. IEEE.

[22] Cho, D., Lee, M., Kim, S., and Tai, Y.-W. (2013). Modeling the calibration pipeline of the lytro camera for high quality light-field image reconstruction. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 3280–3287. IEEE.

[23] Choudhury, B., Swanson, R., Heide, F., Wetzstein, G., and Heidrich, W. (2017). Consensus convolutional sparse coding. *IEEE International Conference on Computer Vision*.

[24] Christensen, J. D. (2003). Normalization of brain magnetic resonance images using histogram even-order derivative analysis. *Magnetic Resonance Imaging*, 21(7):817–820.

[25] Cordier, N., Delingette, H., Lê, M., and Ayache, N. (2016). Extended modality propagation: Image synthesis of pathological cases. *IEEE Transactions on Medical Imaging*, 35(12):2598–2608.

[26] Davis, G., Mallat, S., and Avellaneda, M. (1997). Adaptive greedy approximations. *Constructive Approximation*, 13(1):57–98.

[27] Dong, C., Loy, C. C., He, K., and Tang, X. (2016). Image super-resolution using deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(2):295–307.

[28] Dong, W., Li, X., Zhang, L., and Shi, G. (2011). Sparsity-based image denoising via dictionary learning and structural clustering. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 457–464. IEEE.

[29] Dong, W., Zhang, L., Shi, G., and Li, X. (2013). Nonlocally centralized sparse representation for image restoration. *IEEE Transactions on Image Processing*, 22(4):1620–1630.

[30] Donoho, D. L. (2006). For most large underdetermined systems of linear equations the minimal l1-norm solution is also the sparsest solution. *Communications on Pure and Applied Mathematics*, 59(6):797–829.

[31] Dosovitskiy, A., Tobias Springenberg, J., and Brox, T. (2015). Learning to generate chairs with convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1538–1546.

[32] Efros, A. A. and Freeman, W. T. (2001). Image quilting for texture synthesis and transfer. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 341–346. ACM.

[33] Efros, A. A. and Leung, T. K. (1999). Texture synthesis by non-parametric sampling. In *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, volume 2, pages 1033–1038. IEEE.

[34] Elad, M. and Aharon, M. (2006). Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Transactions on Image Processing*, 15(12):3736–3745.

[35] Elad, M. and Milanfar, P. (2017). Style transfer via texture synthesis. *IEEE Transactions on Image Processing*, 26(5):2338–2351.

[36] Elhamifar, E. and Vidal, R. (2013). Sparse subspace clustering: Algorithm, theory, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(11):2765–2781.

[37] Engan, K., Aase, S. O., and Husoy, J. H. (1999). Method of optimal directions for frame design. In *Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on*, volume 5, pages 2443–2446. IEEE.

[38] Fischl, B., Salat, D. H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., Van Der Kouwe, A., Killiany, R., Kennedy, D., Klaveness, S., et al. (2002). Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. *Neuron*, 33(3):341–355.

[39] Frangi, A. F., Niessen, W. J., Vincken, K. L., and Viergever, M. A. (1998). Multiscale vessel enhancement filtering. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 130–137. Springer.

[40] Freeman, W. T., Jones, T. R., and Pasztor, E. C. (2002). Example-based super-resolution. *IEEE Computer Graphics and Applications*, 22(2):56–65.

[41] Freeman, W. T., Pasztor, E. C., and Carmichael, O. T. (2000). Learning low-level vision. *International Journal of Computer Vision*, 40(1):25–47.

[42] Gao, X., Wang, N., Tao, D., and Li, X. (2012a). Face sketch–photo synthesis and retrieval using sparse representation. *IEEE Transactions on Circuits and Systems for Video Technology*, 22(8):1213–1226.

[43] Gao, X., Zhang, K., Tao, D., and Li, X. (2012b). Image super-resolution with sparse neighbor embedding. *IEEE Transactions on Image Processing*, 21(7):3194–3205.

[44] Gatys, L. A., Ecker, A. S., and Bethge, M. (2016). Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2414–2423.

[45] Gersho, A. and Gray, R. M. (2012). *Vector quantization and signal compression*, volume 159. Springer Science & Business Media.

[46] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680.

[47] Greenspan, H., Oz, G., Kiryati, N., and Peled, S. (2002). Mri inter-slice reconstruction using super-resolution. *Magnetic Resonance Imaging*, 20(5):437–446.

[48] Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012a). A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773.

[49] Gretton, A., Sejdinovic, D., Strathmann, H., Balakrishnan, S., Pontil, M., Fukumizu, K., and Sriperumbudur, B. K. (2012b). Optimal kernel choice for large-scale two-sample tests. In *Advances in Neural Information Processing Systems*, pages 1205–1213.

[50] Grevera, G. J. and Udupa, J. K. (1998). An objective comparison of 3d image interpolation methods. *IEEE Transactions on Medical Imaging*, 17(4):642–652.

[51] Gu, S., Zuo, W., Xie, Q., Meng, D., Feng, X., and Zhang, L. (2015). Convolutional sparse coding for image super-resolution. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1823–1831.

[52] Han, J., Chen, C., Shao, L., Hu, X., Han, J., and Liu, T. (2015). Learning computational models of video memorability from fmri brain imaging. *IEEE Transactions on Cybernetics*, 45(8):1692–1703.

[53] Hartigan, J. A. and Wong, M. A. (1979). Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108.

[54] Havaei, M., Guizard, N., Chapados, N., and Bengio, Y. (2016). Hemis: Hetero-modal image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 469–477. Springer.

[55] He, D., Xia, Y., Qin, T., Wang, L., Yu, N., Liu, T., and Ma, W.-Y. (2016a). Dual learning for machine translation. In *Advances in Neural Information Processing Systems*, pages 820–828.

[56] He, K., Zhang, X., Ren, S., and Sun, J. (2016b). Deep residual learning for image recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[57] Heide, F., Heidrich, W., and Wetzstein, G. (2015). Fast and flexible convolutional sparse coding. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5135–5143. IEEE.

[58] Herman, G. T., Rowland, S. W., and Yau, M.-m. (1979). A comparative study of the use of linear and modified cubic spline interpolation for image reconstruction. *IEEE Transactions on Nuclear Science*, 26(2):2879–2894.

[59] Hertzmann, A., Jacobs, C. E., Oliver, N., Curless, B., and Salesin, D. H. (2001). Image analogies. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 327–340. ACM.

[60] Huang, D.-A. and Frank Wang, Y.-C. (2013). Coupled dictionary and feature space learning with applications to cross-domain image synthesis and recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 2496–2503.

[61] Huang, J.-B., Singh, A., and Ahuja, N. (2015). Single image super-resolution from transformed self-exemplars. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5197–5206. IEEE.

[62] Huang, Y., Beltrachini, L., Shao, L., and Frangi, A. F. (2016a). Geometry regularized joint dictionary learning for cross-modality image synthesis in magnetic resonance imaging. In *International Workshop on Simulation and Synthesis in Medical Imaging*, pages 118–126. Springer.

[63] Huang, Y., Shao, L., and Frangi, A. F. (2017a). Cross-modality image synthesis via weakly-coupled and geometry co-regularized joint dictionary learning. *IEEE Transactions on Medical Imaging*.

[64] Huang, Y., Shao, L., and Frangi, A. F. (2017b). Dote: Dual convolutional filter learning for super-resolution and cross-modality synthesis in MRI. pages 89–98.

[65] Huang, Y., Shao, L., and Frangi, A. F. (2017c). Simultaneous super-resolution and cross-modality synthesis of 3D medical images using weakly-supervised joint convolutional sparse coding. In *IEEE Computer Vision and Pattern Recognition*.

[66] Huang, Y., Zhu, F., Shao, L., and Frangi, A. F. (2016b). Color object recognition via cross-domain learning on rgb-d images. In *Robotics and Automation (ICRA), 2016 IEEE International Conference on*, pages 1672–1677. IEEE.

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

[68] Iglesias, J. E. and Sabuncu, M. R. (2015). Multi-atlas segmentation of biomedical images: a survey. *Medical Image Analysis*, 24(1):205–219.

[69] Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. (2016). Image-to-image translation with conditional adversarial networks. *arXiv preprint arXiv:1611.07004*.

[70] Jia, K., Wang, X., and Tang, X. (2013). Image transformation based on learning dictionaries across image spaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(2):367–380.

[71] Jiang, Z., Lin, Z., and Davis, L. S. (2013). Label consistent k-svd: Learning a discriminative dictionary for recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(11):2651–2664.

[72] Jog, A., Carass, A., Roy, S., Pham, D. L., and Prince, J. L. (2017). Random forest regression for magnetic resonance image synthesis. *Medical Image Analysis*, 35:475–488.

[73] Jog, A., Roy, S., Carass, A., and Prince, J. L. (2013). Magnetic resonance image synthesis through patch regression. In *2013 IEEE 10th International Symposium on Biomedical Imaging*, pages 350–353. IEEE.

[74] Johnson, J., Alahi, A., and Fei-Fei, L. (2016). Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, pages 694–711. Springer.

[75] Joyce, T., Chartsias, A., and Tsaftaris, S. A. (2017). Robust multi-modal MR image synthesis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 347–355. Springer.

[76] 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.

[77] Kavukcuoglu, K., Sermanet, P., Boureau, Y.-L., Gregor, K., Mathieu, M., and Cun, Y. L. (2010). Learning convolutional feature hierarchies for visual recognition. In *Advances in Neural Information Processing Systems*, pages 1090–1098.

[78] Keys, R. (1981). Cubic convolution interpolation for digital image processing. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 29(6):1153–1160.

[79] Kim, T., Cha, M., Kim, H., Lee, J., and Kim, J. (2017). Learning to discover cross-domain relations with generative adversarial networks. *arXiv preprint arXiv:1703.05192*.

[80] Kinga, D. and Adam, J. B. (2015). A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*.

[81] Kong, S. and Wang, D. (2012). A dictionary learning approach for classification: separating the particularity and the commonality. In *European Conference on Computer Vision*, pages 186–199. Springer.

[82] Kroon, D.-J. and Slump, C. H. (2009). MRI modalitiy transformation in demon registration. In *IEEE International Symposium on Biomedical Imaging*, pages 963–966. IEEE.

[83] Kwatra, V., Schödl, A., Essa, I., Turk, G., and Bobick, A. (2003). Graphcut textures: image and video synthesis using graph cuts. In *ACM Transactions on Graphics (ToG)*, volume 22, pages 277–286. ACM.

[84] LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.

[85] Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al. (2016). Photo-realistic single image super-resolution using a generative adversarial network. *arXiv preprint arXiv:1609.04802*.

[86] Lee, H., Battle, A., Raina, R., and Ng, A. Y. (2007). Efficient sparse coding algorithms. *Neural Information Processing Systems*, 19:801.

[87] Lehmann, T. M., Gonner, C., and Spitzer, K. (1999). Survey: Interpolation methods in medical image processing. *IEEE Transactions on Medical Imaging*, 18(11):1049–1075.

[88] Li, C. and Wand, M. (2016). Precomputed real-time texture synthesis with markovian generative adversarial networks. In *European Conference on Computer Vision*, pages 702–716. Springer.

[89] Li, S., Yin, H., and Fang, L. (2012). Group-sparse representation with dictionary learning for medical image denoising and fusion. *IEEE Transactions on Biomedical Engineering*, 59(12):3450–3459.

[90] Li, X. and Orchard, M. T. (2001). New edge-directed interpolation. *IEEE Transactions on Image Processing*, 10(10):1521–1527.

[91] Lin, D. and Tang, X. (2005). Coupled space learning of image style transformation. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 2, pages 1699–1706. IEEE.

[92] Lin, X., Liu, Y., Wu, J., and Dai, Q. (2014). Spatial-spectral encoded compressive hyperspectral imaging. *ACM Transactions on Graphics (TOG)*, 33(6):233.

[93] Liu, M.-Y., Breuel, T., and Kautz, J. (2017). Unsupervised image-to-image translation networks. *arXiv preprint arXiv:1703.00848*.

[94] Liu, M.-Y. and Tuzel, O. (2016). Coupled generative adversarial networks. In *Advances in Neural Information Processing Systems*, pages 469–477.

[95] Liu, Z., Zhang, Z., and Shan, Y. (2004). Image-based surface detail transfer. *IEEE Computer Graphics and Applications*, 24(3):30–35.

[96] Long, M., Ding, G., Wang, J., Sun, J., Guo, Y., and Yu, P. S. (2013). Transfer sparse coding for robust image representation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 407–414.

[97] Long, M., Wang, J., Cao, Y., Sun, J., and Philip, S. Y. (2016). Deep learning of transferable representation for scalable domain adaptation. *IEEE Transactions on Knowledge and Data Engineering*, 28(8):2027–2040.

[98] Long, M., Wang, J., Ding, G., Shen, D., and Yang, Q. (2014). Transfer learning with graph co-regularization. *IEEE Transactions on Knowledge and Data Engineering*, 26(7):1805–1818.

[99] Lu, X., Yuan, H., Yan, P., Yuan, Y., and Li, X. (2012). Geometry constrained sparse coding for single image super-resolution. In *IEEE Computer Vision and Pattern Recognition*, pages 1648–1655.

[100] Madabhushi, A. and Udupa, J. K. (2006). New methods of mr image intensity standardization via generalized scale. *Medical Physics*, 33(9):3426–3434.

[101] Mahendran, A. and Vedaldi, A. (2015). Understanding deep image representations by inverting them. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5188–5196.

[102] Mairal, J., Bach, F., Ponce, J., and Sapiro, G. (2009a). Online dictionary learning for sparse coding. In *Proceedings of the 26th annual international conference on machine learning*, pages 689–696. ACM.

[103] Mairal, J., Bach, F., Ponce, J., Sapiro, G., and Zisserman, A. (2008a). Discriminative learned dictionaries for local image analysis. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE.

[104] Mairal, J., Elad, M., and Sapiro, G. (2008b). Sparse representation for color image restoration. *IEEE Transactions on Image Processing*, 17(1):53–69.

[105] Mairal, J., Ponce, J., Sapiro, G., Zisserman, A., and Bach, F. R. (2009b). Supervised dictionary learning. In *Advances in Neural Information Processing Systems*, pages 1033–1040.

[106] Manjón, J. V., Carbonell-Caballero, J., Lull, J. J., García-Martí, G., Martí-Bonmatí, L., and Robles, M. (2008). Mri denoising using non-local means. *Medical Image Analysis*, 12(4):514–523.

[107] Manjón, J. V., Coupé, P., Buades, A., Fonov, V., Collins, D. L., and Robles, M. (2010). Non-local mri upsampling. *Medical Image Analysis*, 14(6):784–792.

[108] Marwah, K., Wetzstein, G., Bando, Y., and Raskar, R. (2013). Compressive light field photography using overcomplete dictionaries and optimized projections. *ACM Transactions on Graphics (TOG)*, 32(4):46.

[109] Menze, B. H., Van Leemput, K., Lashkari, D., Riklin-Raviv, T., Geremia, E., Alberts, E., Gruber, P., Wegener, S., Weber, M.-A., Szekely, G., et al. (2016). A generative probabilistic model and discriminative extensions for brain lesion segmentation—with application to tumor and stroke. *IEEE Transactions on Medical Imaging*, 35(4):933–946.

[110] Monay, F. and Gatica-Perez, D. (2007). Modeling semantic aspects for cross-media image indexing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(10):1802–1817.

[111] Mueller, S. G., Weiner, M. W., Thal, L. J., Petersen, R. C., Jack, C., Jagust, W., Trojanowski, J. Q., Toga, A. W., and Beckett, L. (2005). The alzheimer's disease neuroimaging initiative. *Neuroimaging Clinics of North America*, 15(4):869–877.

[112] Nyúl, L. G., Udupa, J. K., et al. (1999). On standardizing the mr image intensity scale. *Image*, 1081.

[113] Nyúl, L. G., Udupa, J. K., and Zhang, X. (2000). New variants of a method of mri scale standardization. *IEEE Transactions on Medical Imaging*, 19(2):143–150.

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

[115] Olivier Commowick, S. K. W. and Malandain, G. (2009). Using Frankenstein's creature paradigm to build a patient specific atlas. In *International Conference On Medical Image Computing and Computer Assisted Intervention*, pages 993–1000. Springer.

[116] Olshausen, B. A. and Field, D. J. (1997). Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision Research*, 37(23):3311–3325.

[117] Ongie, G. and Jacob, M. (2015). Recovery of discontinuous signals using group sparse higher degree total variation. *IEEE Signal Processing Letters*, 22:1414–1418.

[118] Pan, S. J. and Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359.

[119] Radford, A., Metz, L., and Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*.

[120] Rao, S. R., Tron, R., Vidal, R., and Ma, Y. (2008). Motion segmentation via robust subspace separation in the presence of outlying, incomplete, or corrupted trajectories. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE.

[121] Ravishankar, S. and Bresler, Y. (2011). Mr image reconstruction from highly undersampled k-space data by dictionary learning. *IEEE Transactions on Medical Imaging*, 30(5):1028–1041.

[122] Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., and Lee, H. (2016). Generative adversarial text to image synthesis. *arXiv preprint arXiv:1605.05396*.

[123] Robitaille, N., Mouiha, A., Crépeault, B., Valdivia, F., and Duchesne, S. (2012). Tissue-based mri intensity standardization: application to multicentric datasets. *International Journal of Biomedical Imaging*, 2012:4.

[124] Rousseau, F. (2008). Brain hallucination. *European Conference on Computer Vision*, pages 497–508.

[125] Rousseau, F., Initiative, A. D. N., et al. (2010). A non-local approach for image super-resolution using intermodality priors. *Medical Image Analysis*, 14(4):594–605.

[126] Rowland, A., Burns, M., Hartkens, T., Hajnal, J., Rueckert, D., and Hill, D. (2004). Information extraction from images (ixi): Image processing workflows using a grid enabled image database. *Proceedings of DiDaMIC*, 4:55–64.

[127] Roy, S., Carass, A., and Prince, J. L. (2013). Magnetic resonance image example-based contrast synthesis. *IEEE Transactions on Medical Imaging*, 32(12):2348–2363.

[128] Rueda, A., Malpica, N., and Romero, E. (2013). Single-image super-resolution of brain mr images using overcomplete dictionaries. *Medical Image Analysis*, 17(1):113–132.

[129] Shao, L., Gao, X., and Li, H. (2014a). Image restoration and enhancement: Recent advances and applications. *Signal Processing*, 103:1–5.

[130] Shao, L., Yan, R., Li, X., and Liu, Y. (2014b). From heuristic optimization to dictionary learning: A review and comprehensive comparison of image denoising algorithms. *IEEE Transactions on Cybernetics*, 44(7):1001–1013.

[131] Shao, L. and Zhao, M. (2007). Order statistic filters for image interpolation. In *2007 IEEE International Conference on Multimedia and Expo*, pages 452–455. IEEE.

[132] Shao, L., Zhu, F., and Li, X. (2015). Transfer learning for visual categorization: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 26(5):1019–1034.

[133] Shi, F., Cheng, J., Wang, L., Yap, P.-T., and Shen, D. (2015). Lrtv: Mr image super-resolution with low-rank and total variation regularizations. *IEEE Transactions on Medical Imaging*, 34(12):2459–2466.

[134] Shinohara, R. T., Sweeney, E. M., Goldsmith, J., Shiee, N., Mateen, F. J., Calabresi, P. A., Jarso, S., Pham, D. L., Reich, D. S., Crainiceanu, C. M., et al. (2014). Statistical normalization techniques for magnetic resonance imaging. *NeuroImage: Clinical*, 6:9–19.

[135] Shrivastava, A., Pfister, T., Tuzel, O., Susskind, J., Wang, W., and Webb, R. (2016). Learning from simulated and unsupervised images through adversarial training. *arXiv preprint arXiv:1612.07828*.

[136] Simonovsky, M., Gutiérrez-Becker, B., Mateus, D., Navab, N., and Komodakis, N. (2016). A deep metric for multimodal registration. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 10–18. Springer.

[137] Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556".

[138] Smeulders, A. W., Worring, M., Santini, S., Gupta, A., and Jain, R. (2000). Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1349–1380.

[139] Steinwart, I. (2001). On the influence of the kernel on the consistency of support vector machines. *Journal of Machine Learning Research*, 2(Nov):67–93.

[140] Stytz, M. R. and Parrott, R. W. (1993). Using kriging for 3d medical imaging. *Computerized Medical Imaging and Graphics*, 17(6):421–442.

[141] Sun, J., Zheng, N.-N., Tao, H., and Shum, H.-Y. (2003). Image hallucination with primal sketch priors. In *IEEE Computer Vision and Pattern Recognition*, volume 2, pages II–729. IEEE.

[142] Szlam, A., Kavukcuoglu, K., and LeCun, Y. (2010). Convolutional matching pursuit and dictionary training. *arXiv preprint arXiv:1010.0422*.

[143] Tang, X. and Wang, X. (2003). Face sketch synthesis and recognition. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pages 687–694. IEEE.

[144] Tang, Y. and Shao, L. (2017). Pairwise operator learning for patch-based single-image super-resolution. *IEEE Transactions on Image Processing*, 26(2):994–1003.

[145] Thaipanich, T., Oh, B. T., Wu, P.-H., Xu, D., and Kuo, C.-C. J. (2010). Improved image denoising with adaptive nonlocal means (anl-means) algorithm. *IEEE Transactions on Consumer Electronics*, 56:2623–2630.

[146] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, pages 267–288.

[147] Timofte, R., De Smet, V., and Van Gool, L. (2013). Anchored neighborhood regression for fast example-based super-resolution. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1920–1927.

[148] Timofte, R., De Smet, V., and Van Gool, L. (2014). A+: Adjusted anchored neighborhood regression for fast super-resolution. In *Asian Conference on Computer Vision*, pages 111–126. Springer.

[149] Ulyanov, D., Lebedev, V., Vedaldi, A., and Lempitsky, V. S. (2016). Texture networks: Feed-forward synthesis of textures and stylized images. In *International Conference on Machine Learning*, pages 1349–1357.

[150] Van Nguyen, H., Zhou, K., and 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*, pages 677–684. Springer.

[151] Van Ouwerkerk, J. (2006). Image super-resolution survey. *Image and Vision Computing*, (10):1039–1052.

[152] van Tulder, G. and de Bruijne, M. (2015). Why does synthesized data improve multi-sequence classification? In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 531–538. Springer.

[153] Vemulapalli, R., Van Nguyen, H., and Kevin Zhou, S. (2015). Unsupervised cross-modal synthesis of subject-specific scans. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 630–638.

[154] Wang, C., Xu, C., Wang, C., and Tao, D. (2017). Perceptual adversarial networks for image-to-image transformation. *arXiv preprint arXiv:1706.09138*.

[155] Wang, S., Zhang, L., Liang, Y., and Pan, Q. (2012). Semi-coupled dictionary learning with applications to image super-resolution and photo-sketch synthesis. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2216–2223. IEEE.

[156] Wang, X. and Gupta, A. (2016). Generative image modeling using style and structure adversarial networks. In *European Conference on Computer Vision*, pages 318–335. Springer.

[157] Wang, X. and Tang, X. (2005). Hallucinating face by eigentransformation. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 35(3):425–434.

[158] Wang, X. and Tang, X. (2009). Face photo-sketch synthesis and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(11):1955–1967.

[159] Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612.

[160] Woo, J., Stone, M., and Prince, J. L. (2015). Multimodal registration via mutual information incorporating geometric and spatial context. *IEEE Transactions on Image Processing*, 24(2):757–769.

[161] Wright, J., Ma, Y., Mairal, J., Sapiro, G., Huang, T. S., and Yan, S. (2010). Sparse representation for computer vision and pattern recognition. *Proceedings of the IEEE*, 98(6):1031–1044.

[162] Wright, J., Yang, A. Y., Ganesh, A., Sastry, S. S., and Ma, Y. (2009). Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(2):210–227.

[163] Wu, J., Zhang, C., Xue, T., Freeman, B., and Tenenbaum, J. (2016). Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. In *Advances in Neural Information Processing Systems*, pages 82–90.

[164] Xu, Q., Yu, H., Mou, X., Zhang, L., Hsieh, J., and Wang, G. (2012). Low-dose x-ray CT reconstruction via dictionary learning. *IEEE Transactions on Medical Imaging*, 31(9):1682–1697.

[165] Yan, R., Shao, L., Cvetkovic, S. D., and Klijn, J. (2012). Improved nonlocal means based on pre-classification and invariant block matching. *Journal of Display Technology*, 8(4):212–218.

[166] Yan, R., Shao, L., and Liu, Y. (2013). Nonlocal hierarchical dictionary learning using wavelets for image denoising. *IEEE Transactions on Image Processing*, 22(12):4689–4698.

[167] Yang, J., Wang, Z., Lin, Z., Cohen, S., and Huang, T. (2012a). Coupled dictionary training for image super-resolution. *IEEE Transactions on Image Processing*, 21(8):3467–3478.

[168] Yang, J., Wright, J., Huang, T. S., and Ma, Y. (2010). Image super-resolution via sparse representation. *IEEE Transactions on Image Processing*, 19(11):2861–2873.

[169] Yang, J., Yu, K., Gong, Y., and Huang, T. (2009). Linear spatial pyramid matching using sparse coding for image classification. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1794–1801. IEEE.

[170] Yang, M., Zhang, L., Feng, X., and Zhang, D. (2014). Sparse representation based fisher discrimination dictionary learning for image classification. *International Journal of Computer Vision*, 109(3):209–232.

[171] Yang, S., Wang, M., Chen, Y., and Sun, Y. (2012b). Single-image super-resolution reconstruction via learned geometric dictionaries and clustered sparse coding. *IEEE Transactions on Image Processing*, 21(9):4016–4028.

[172] Ye, D. H., Zikic, D., Glocker, B., Criminisi, A., and 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*, pages 606–613. Springer.

[173] Yi, Z., Zhang, H., Gong, P. T., et al. (2017). Dualgan: Unsupervised dual learning for image-to-image translation. *arXiv preprint arXiv:1704.02510*.

[174] Yosinski, J., Clune, J., Bengio, Y., and Lipson, H. (2014). How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems*, pages 3320–3328.

[175] Zeiler, M. D., Krishnan, D., Taylor, G. W., and Fergus, R. (2010). Deconvolutional networks. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2528–2535. IEEE.

[176] Zeyde, R., Elad, M., and Protter, M. (2010). On single image scale-up using sparse-representations. In *International Conference on Curves and Surfaces*, pages 711–730. Springer.

[177] Zhang, H., Xu, T., Li, H., Zhang, S., Huang, X., Wang, X., and Metaxas, D. (2016). Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. *arXiv preprint arXiv:1612.03242*.

[178] Zhang, K., Gao, X., Tao, D., and Li, X. (2012). Multi-scale dictionary for single image super-resolution. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1114–1121. IEEE.

[179] Zhang, L. and Wu, X. (2006). An edge-guided image interpolation algorithm via directional filtering and data fusion. *IEEE Transactions on Image Processing*, 15(8):2226–2238.

[180] Zhang, Q. and Li, B. (2010). Discriminative k-svd for dictionary learning in face recognition. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2691–2698. IEEE.

[181] Zhang, Y., Jiang, Z., and Davis, L. S. (2013). Learning structured low-rank representations for image classification. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 676–683. IEEE.

[182] Zhen, X., Zhang, H., Islam, A., Bhaduri, M., Chan, I., and Li, S. (2017). Direct and simultaneous estimation of cardiac four chamber volumes by multioutput sparse regression. *Medical Image Analysis*, 36:184–196.

[183] Zheng, F., Tang, Y., and Shao, L. (2016). Hetero-manifold regularisation for cross-modal hashing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

[184] Zheng, M., Bu, J., Chen, C., Wang, C., Zhang, L., Qiu, G., and Cai, D. (2011). Graph regularized sparse coding for image representation. *IEEE Transactions on Image Processing*, 20(5):1327–1336.

[185] Zhou, N., Shen, Y., Peng, J., and Fan, J. (2012). Learning inter-related visual dictionary for object recognition. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 3490–3497. IEEE.

[186] Zhu, F. and Shao, L. (2014). Weakly-supervised cross-domain dictionary learning for visual recognition. *International Journal of Computer Vision*, 109(1-2):42–59.

[187] Zhu, F., Shao, L., and Lin, M. (2013). Multi-view action recognition using local similarity random forests and sensor fusion. *Pattern Recognition Letters*, 34(1):20–24.

[188] Zhu, J.-Y., Krähenbühl, P., Shechtman, E., and Efros, A. A. (2016). Generative visual manipulation on the natural image manifold. In *European Conference on Computer Vision*, pages 597–613. Springer.

[189] Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. *arXiv preprint arXiv:1703.10593*.

[190] Zoran, D. and Weiss, Y. (2011). From learning models of natural image patches to whole image restoration. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 479–486. IEEE.

Appendix A

Parallel Contrast Experiment

To quantitatively evaluate all proposed methods in this thesis, we further examine the performance of our methods including SiSCS, DOTE, WEENIE, WAG, and T-GAN for image synthesis from the IXI dataset of the selected subjects using PSNR (dB) and SSIM. Specifically, we utilize the same experimental settings trained with the identical data for the parallel comparison.

Implementation details: The models are evaluated on whole subject of brain MR images acquired from the IXI dataset. We randomly select 100 subjects involving pairwise PD-w and T2-w data for training, and then divide the reminding 478 subjects into two parts: 400 subjects and 78 subjects. To verify the effectiveness of the weakly-supervised approaches, we simply separate 400 pairwise data to 200 unpaired images. That is, for each method, four groups of experiments (i.e., PD-w -> T2-w and T2-w -> PD-w training on the fixed 100 subject pairs; PD-w -> T2-w and T2-w -> PD-w training on the 100 subject pairs + 200 enhancing subjects) are performed. In particular, the proposed supervised algorithms cannot handle the image alignment procedure. We keep the original pairs and input them for the enhancement.

We generate 20 groups of experiments corresponding to various proposed methods in two cases (i.e., PD-w -> T2-w and T2-w -> PD-w). The algorithms we used are signed as: SiSCS-100, SiSCS-300, DOTE-100, DOTE-300, WEENIE-100, WEENIE-300, WAG-100, WAG-300, T-GAN-100, and T-GAN-300. Models are trained using 100 subjects are represented as'-100', and '-300' denotes that the enhanced 200 subjects are included in the training data. Table A.1 shows the quantitative results (averaged PSNR and SSIM) on the IXI dataset. The proposed method T-GAN yields the best quantitative results for all cases, 4.51 dB (training with 100 subjects for T2-w -> PD-w), 3.42 dB (training with 100 subjects for PD-w -> T2-w), 3.67 dB (training with 300 subjects for T2-w -> PD-w), and 3.06 dB (training with 300 subjects for PD-w -> T2-w) PSNR better than the worst case; 0.0578 (training with 100 subjects for T2-w -> PD-w), 0.0893 (training with 100 subjects for PD-w -> T2-w), 0.0637 (training with 300

Metric(avg.)	**SiSCS-100**	**DOTE-100**	**WEENIE-100**	**WAG-100**	**T-GAN-100**
PD-w → T2-w					
PSNR (dB)	33.25	34.43	34.89	34.32	36.67
SSIM	0.8611	0.8786	0.8801	0.8599	0.9492
T2-w → PD-w					
PSNR (dB)	30.27	32.91	33.00	32.03	34.78
SSIM	0.8523	0.8614	0.8617	0.8549	0.9098
Metric(avg.)	**SiSCS-300**	**DOTE-300**	**WEENIE-300**	**WAG-300**	**T-GAN-300**
PD-w → T2-w					
PSNR (dB)	33.98	37.09	37.02	35.01	37.04
SSIM	0.8700	0.9100	0.8927	0.8612	0.9523
T2-w → PD-w					
PSNR (dB)	31.45	35.07	34.93	32.96	35.12
SSIM	0.8597	0.9178	0.9050	0.8578	0.9215

Table A.1 Quantitative evaluation (PSNR (dB) and SSIM): SiSCS, DOTE, WEENIE, WAG, and T-GAN on the IXI dataset.

subjects for T2-w -> PD-w), and 0.0911 (training with 300 subjects for PD-w -> T2-w) SSIM better than the worst case.