Simultaneous Super-Resolution and Cross-Modality Synthesis of 3D Medical Images using Weakly-Supervised Joint Convolutional Sparse Coding

Yawen Huang[1], Ling Shao[2], Alejandro F. Frangi[1]

[1]Department of Electronic and Electrical Engineering, The University of Sheffield, UK

[2]School of Computing Sciences, University of East Anglia, UK

{yhuang36, a.frangi}@sheffield.ac.uk, ling.shao@uea.ac.uk

Abstract

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 paper, 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. 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

---

[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.

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 [24] 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 [20, 28]. State-of-the-art methods mostly adopt either external data or internal data to guide the learning algorithms [25, 30]. 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 [8], nonlinear regression forests [19], coupled dictionary learning [26], and convolutional neural network (CNN) [10], 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 paper, 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 <u>WE</u>akly-sup<u>E</u>rvised joi<u>N</u>t convolut<u>I</u>onal spars<u>E</u> coding (WEENIE), and perform extensive experiments to verify its performance.

The main contributions of this paper 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.

2. Related Work

With the goal to transfer the modality information from the source domain to the target domain, recent developments in CMS, such as texture synthesis [6, 10, 13], face photo-sketch synthesis [9, 36], and multi-modal retrieval [23, 29], have shown promising results. In this paper, we focus on the problems of image super-resolution and cross-modality synthesis, so only review related methods on these two aspects.

Image Super-Resolution: The purpose of image SR is to reconstruct an HR image from its 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 [21], bicubic [20], and some follow-up methods [28, 41], 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 [3, 7, 38] exploit a mapping relationship between LR and HR image pairs from a specified external dataset. In the pioneer work of Freeman *et al.* [7], 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.* [3] 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 [27, 38] were explored to generate a pair of dictionaries for LR and HR patch pairs to address the image SR problem. Wang *et al.* [35] and Huang *et al.* [14] 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 [5], 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 [25] and/or scale-space pyramid of the given image itself [15].

Cross-Modality Synthesis: In parallel, various CMS methods have been proposed for synthesizing unavailable modality data from available source images, especially in the medical imaging community [26, 33, 34]. One of the well-established modality transformation approaches is the example-based learning method generated by Freeman *et al.* [8]. 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. Roy *et al.* [26] 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. Sim-
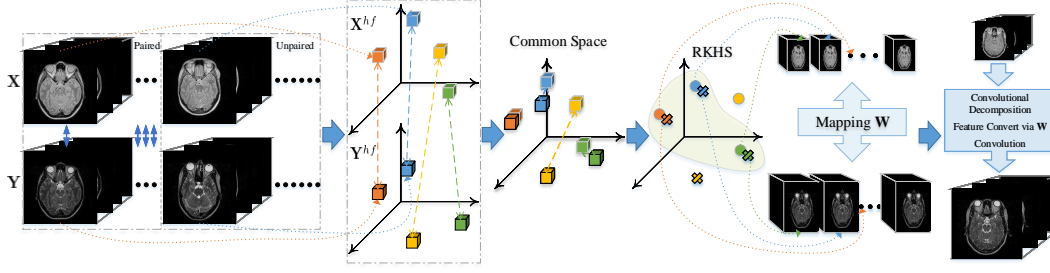
Figure 1. Flowchart of the proposed method (WEENIE) for simultaneous SR and cross-modality synthesis.

ilar work was also used in [17]. In [1], 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. More recently, a location-sensitive deep network [33] 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. Gatys *et al.* [10] introduced a CNN algorithm of artistic style, that new images can be generated by performing a pre-image search in high-level image content to match generic feature representations of example images. In addition to the aforementioned methods, most CMS algorithms rely on the strictly registered pairs to train models. As argued in [34], it would be preferable to use an unsupervised approach to deal with input data instead of ensuring data to be coupled invariably.

3. Weakly-Supervised Joint Convolutional Sparse Coding

3.1. 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:

$$\arg\min_{\mathbf{f},\mathbf{z}} \frac{1}{2} \left\| \mathbf{x} - \sum_{k=1}^{K} \mathbf{f}_k * \mathbf{z}_k \right\|_2^2 + \lambda \sum_{k=1}^{K} \|\mathbf{z}_k\|_1 \quad (1)$$
$$s.t.\ \|\mathbf{f}_k\|_2^2 \leq 1\ \forall k = \{1, ..., K\},$$

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. $\mathbf{f} = \begin{bmatrix} \mathbf{f}_1^T, ..., \mathbf{f}_K^T \end{bmatrix}^T$ and $\mathbf{z} = \begin{bmatrix} \mathbf{z}_1^T, ..., \mathbf{z}_K^T \end{bmatrix}^T$ 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.* [39] proposed to solve the CSC in Eq. (1) through alternatively optimizing one variable while fixing the other one in the spatial domain. Advances in recent fast convolutional sparse coding (FCSC) [2] 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.

3.2. Problem Formulation

The simultaneous SR and cross-modality synthesis problem can be formulated as: given a three-dimensional LR image $\mathbf{X}$ of modality $\mathcal{M}_1$, 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 $\mathcal{M}_2$. Suppose that we are given a group of LR images of modality $\mathcal{M}_1$, *i.e.*, $\mathcal{X} = [\mathbf{X}_1, ..., \mathbf{X}_P] \in \mathbb{R}^{m \times n \times t \times P}$, and a set of HR images of modality $\mathcal{M}_2$, *i.e.*, $\mathcal{Y} = [\mathbf{Y}_2, ..., \mathbf{Y}_Q] \in \mathbb{R}^{m \times n \times t \times Q}$. 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 $\mathcal{M}_1$ are mostly different from target modality $\mathcal{M}_2$, 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 $\mathcal{X}$ and $\mathcal{Y}$ firstly, then employ the aligned data to jointly learn a pair of filters, while assuming that there exists a mapping function $\mathcal{F}(\cdot)$ for associating and predicting cross-modality data in the projected common feature space. Particularly, we want to synthesize MRI of human brains in this paper. An overview of our proposed work is depicted in Fig. 1.

Notation: For simplicity, we denote matrices and 3D im-

ages 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 $\mathcal{M}_1$ called source modality belongs to the source domain, and with modality $\mathcal{M}_2$ called target modality belongs to the target domain.

3.3. Hetero-Domain Image Alignment

The design of an alignment $\mathcal{A}(\cdot)$ from $\mathcal{X}$ to $\mathcal{Y}$ 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 [3, 38]. In this paper, we adopt first- and second-order derivatives involving horizontal and vertical gradients as the features for LR images by $\mathbf{X}_p^{hf} = \mathcal{G} * \mathbf{X}_p$. $\mathcal{G} = \begin{bmatrix} \mathbf{G}_1^1, \mathbf{G}_1^2 \\ \mathbf{G}_2^1, \mathbf{G}_2^2 \end{bmatrix}$, and each gradient $\mathbf{G}$ has the same length of z-axis as input image while $\mathbf{g}_1^1 = [-1, 0, 1]$, $\mathbf{g}_1^2 = \mathbf{g}_1^{1T}$, and $\mathbf{g}_2^1 = [-2, -1, 0, 1, 2]$, $\mathbf{g}_2^2 = \mathbf{g}_2^{1T}$. 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 $\mathcal{A}(\cdot)$, 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 [16]). To this end, we define a subject-specific transformation matrix $\mathbb{A}$ as

$$\mathbb{A} = \begin{bmatrix} K(\mathbf{X}_1^{hf}, \mathbf{Y}_1^{hf}) & \cdots & K(\mathbf{X}_1^{hf}, \mathbf{Y}_Q^{hf}) \\ \vdots & \ddots & \vdots \\ K(\mathbf{X}_P^{hf}, \mathbf{Y}_1^{hf}) & \cdots & K(\mathbf{X}_P^{hf}, \mathbf{Y}_Q^{hf}) \end{bmatrix}, \quad (2)$$

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

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

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

$$\mathbb{A} = \begin{bmatrix} \max(K(1,:)) & & \\ & \ddots & \\ & & \max(K(P,:)) \end{bmatrix}, \quad (4)$$

where $\max(K(p,:))$ denotes the maximum element of the p-th row of $\mathbb{A}$. We further set $\max(K(p,:))$ 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 $\mathcal{X}$ 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.*, $\mathcal{P}_i = \{\mathbf{X}_i, \mathbf{Y}_i\}_{i=1}^P$, by constructing virtual correspondence: $\mathcal{A}(\mathcal{X}, \mathcal{Y}) = \left\| \mathbf{X}^{hf} - \mathbb{A}\mathbf{Y}^{hf} \right\|_2^2$.

3.4. Objective Function

For image modality transformation, coupled sparse coding [18, 38] 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 [39] 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 [18], 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 $\mathcal{X}$ and $\mathcal{Y}$ 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 $\mathcal{X}$ and $\mathcal{Y}$, 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 $\mathcal{F}(\cdot)$, where the feature maps of LR $\mathcal{M}_1$ modality images can be converted to their HR $\mathcal{M}_2$ versions. Given $\mathcal{X}$ and $\mathcal{Y}$, we propose to learn a pair of filters with corresponding feature maps and a mapping function together with the aligned term by

$$\arg \min_{\mathbf{F}^x, \mathbf{F}^y, \mathbf{Z}^x, \mathbf{Z}^y, \mathbf{W}} \frac{1}{2} \left\| \mathbf{X} - \sum_{k=1}^K \mathbf{F}_k^x * \mathbf{Z}_k^x \right\|_F^2$$
$$+ \frac{1}{2} \left\| \mathbf{Y} - \sum_{k=1}^K \mathbf{F}_k^y * \mathbf{Z}_k^y \right\|_F^2 + \beta \sum_{k=1}^K \| \mathbf{Z}_k^y - \mathbf{W}_k \mathbf{Z}_k^x \|_F^2 \quad (5)$$
$$+ \lambda \left(\sum_{k=1}^K \| \mathbf{Z}_k^x \|_1 + \sum_{k=1}^K \| \mathbf{Z}_k^y \|_1 \right) + \gamma \sum_{k=1}^K \| \mathbf{W}_k \|_F^2$$
$$+ \left\| \mathbf{X}^{hf} - \mathbb{A}\mathbf{Y}^{hf} \right\|_2^2 \ s.t. \ \| \mathbf{f}_k^x \|_2^2 \leq 1, \| \mathbf{f}_k^y \|_2^2 \leq 1 \ \forall k,$$

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 $\mathcal{P}$ with LR and $\mathcal{M}_1$ modality; $\mathbf{Y}$ denotes the aligned image from $\mathcal{P}$ containing HR and $\mathcal{M}_2$ 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. $\mathcal{A}(\mathcal{X}, \mathcal{Y})$ is combined to enforce the correspondence for unpaired auxiliary subjects. The mapping function $\mathcal{F}(\mathbf{Z}_k^x, \mathbf{W}_k) = \mathbf{W}_k \mathbf{Z}_k^x$ 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.*, $\min_{\mathbf{W}} \sum_{k=1}^{K} \|\mathbf{Z}_k^y - \mathbf{W}_k \mathbf{Z}_k^x\|_F^2$). Parameters λ, β and γ balance sparsity, feature representation and association mapping.

It is worth noting that $\mathcal{P}_i = \{\mathbf{X}_i, \mathbf{Y}_i\}$ may not be perfect since HF feature alignment in Eq. (4) is not good enough for very heterogeneous domain adaptation by matching the first- and second-order derivatives of $\mathcal{X}$ and means of $\mathcal{Y}$, 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 [4, 22, 42] 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:

$$\frac{1}{P} \sum_{i=1}^{P} \sum_{k=1}^{K} \|\mathbf{W}_k(i)\mathbf{Z}_k^x(i) - \mathbf{Z}_k^y(i)\|_{\mathcal{H}}^2$$
$$= \sum_{k=1}^{K} (\mathbf{W}_k \mathbf{Z}_k^x)^T M_i \mathbf{Z}_k^y = Tr(\sum_{k=1}^{K} \mathbf{Z}_k^y \mathbf{M}(\mathbf{W}_k \mathbf{Z}_k^x)^T), \quad (6)$$

where $\mathcal{H}$ indicates RKHS space, $\mathbf{Z}_k^x(i)$ and $\mathbf{Z}_k^y(i)$ are the paired sparse feature maps for $\mathcal{P}_i = \{\mathbf{X}_i, \mathbf{Y}_i\}$ 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

$$M_i = \begin{cases} \frac{1}{P}, & \mathbf{Z}_k^x(i), \mathbf{Z}_k^y(i) \in \mathcal{P}_i, \\ -\frac{1}{P^2}, & \text{otherwise.} \end{cases}, \quad (7)$$

By regularizing Eq. (5) with Eq. (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:

$$\arg \min_{\mathbf{F}^x, \mathbf{F}^y, \mathbf{Z}^x, \mathbf{Z}^y, \mathbf{W}} \frac{1}{2} \left\| \mathbf{X} - \sum_{k=1}^{K} \mathbf{F}_k^x * \mathbf{Z}_k^x \right\|_F^2 + \gamma \sum_{k=1}^{K} \|\mathbf{W}_k\|_F^2$$

$$+ \frac{1}{2} \left\| \mathbf{Y} - \sum_{k=1}^{K} \mathbf{F}_k^y * \mathbf{Z}_k^y \right\|_F^2 + \beta \sum_{k=1}^{K} \|\mathbf{Z}_k^y - \mathbf{W}_k \mathbf{Z}_k^x\|_F^2$$

$$+ \lambda \left(\sum_{k=1}^{K} \|\mathbf{Z}_k^x\|_1 + \sum_{k=1}^{K} \|\mathbf{Z}_k^y\|_1 \right) + Tr(\sum_{k=1}^{K} \mathbf{Z}_k^y \mathbf{M}(\mathbf{W}_k \mathbf{Z}_k^x)^T)$$

$$+ \|\mathbf{X}^{hf} - \mathbb{A}\mathbf{Y}^{hf}\|_2^2 \quad s.t. \ \|\mathbf{f}_k^x\|_2^2 \le 1, \|\mathbf{f}_k^y\|_2^2 \le 1 \ \forall k. \quad (8)$$

3.5. Optimization

We propose a three-step optimization strategy for efficiently tackling the objective function in Eq. (8) (termed (WEENIE), summarized in Algorithm 1) considering that such multi-variables and unified framework cannot be jointly convex to $\mathbf{F}$, $\mathbf{Z}$, and $\mathbf{W}$. Instead, it is convex with respect to each of them while fixing the remaining variables.

3.5.1 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 $\mathbf{W}$ ($\mathbf{W}$ 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. (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 [2, 12] have been proposed to work around this problem on the theoretical derivation by introducing two auxiliary variables $\mathbf{U}$ and $\mathbf{S}$ to enforce the constraint inherent in the splitting. To facilitate component-wise multiplications, we exploit the convolution subproblem [2] in the Fourier domain[2] derived within the ADMMs framework:

$$\min_{\mathbf{Z}^x} \frac{1}{2} \left\| \hat{\mathbf{X}} - \sum_{k=1}^{K} \hat{\mathbf{F}}_k^x \odot \hat{\mathbf{Z}}_k^x \right\|_F^2 + \|\mathbf{X}^{hf} - \mathbb{A}\mathbf{Y}^{hf}\|_2^2$$

$$+ Tr(\sum_{k=1}^{K} \hat{\mathbf{Z}}_k^y \mathbf{M}(\mathbf{W}_k \hat{\mathbf{Z}}_k^x)^T) + \beta \sum_{k=1}^{K} \left\| \hat{\mathbf{Z}}_k^y - \mathbf{W}_k \hat{\mathbf{Z}}_k^x \right\|_F^2$$

$$+ \lambda \sum_{k=1}^{K} \|\mathbf{U}_k^x\|_1 \quad s.t. \ \|\mathbf{S}_k^x\|_2^2 \le 1, \mathbf{S}_k^x = \mathbf{\Phi}^T \hat{\mathbf{F}}_k^x, \mathbf{U}_k^x = \mathbf{Z}_k^x \ \forall k,$$

---

[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.

$$\min_{\mathbf{Z}^y} \frac{1}{2} \left\| \hat{\mathbf{Y}} - \sum_{k=1}^{K} \hat{\mathbf{F}}_k^y \odot \hat{\mathbf{Z}}_k^y \right\|_F^2 + \left\| \mathbf{X}^{hf} - \mathbb{A}\mathbf{Y}^{hf} \right\|_2^2$$

$$+ Tr(\sum_{k=1}^{K} \hat{\mathbf{Z}}_k^y \mathbf{M}(\mathbf{W}_k \hat{\mathbf{Z}}_k^x)^T) + \beta \sum_{k=1}^{K} \left\| \hat{\mathbf{Z}}_k^x - \mathbf{W}_k \hat{\mathbf{Z}}_k^y \right\|_F^2$$

$$+ \lambda \sum_{k=1}^{K} \|\mathbf{U}_k^y\|_1 \ \ s.t. \ \|\mathbf{S}_k^y\|_2^2 \le 1, \mathbf{S}_k^y = \mathbf{\Phi}^T \hat{\mathbf{F}}_k^y, \mathbf{U}_k^y = \mathbf{Z}_k^y \ \forall k, \tag{9}$$

where $\hat{}$ 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), $\mathbf{\Phi}^T$ 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.

3.5.2 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

$$\min_{\mathbf{F}^x, \mathbf{F}^y} \frac{1}{2} \left\| \hat{\mathbf{X}} - \sum_{k=1}^{K} \hat{\mathbf{F}}_k^x \odot \hat{\mathbf{Z}}_k^x \right\|_F^2 + \frac{1}{2} \left\| \hat{\mathbf{Y}} - \sum_{k=1}^{K} \hat{\mathbf{F}}_k^y \odot \hat{\mathbf{Z}}_k^y \right\|_F^2$$
$$s.t. \ \|\mathbf{f}_k^x\|_2^2 \le 1, \|\mathbf{f}_k^y\|_2^2 \le 1 \ \forall k, \tag{10}$$

The optimization with respect to Eq. (10) can be solved by a one-by-one update strategy [35] through an augmented Lagrangian method [2].

3.5.3 Learning Mapping Function

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

$$\min_{\mathbf{W}} \sum_{k=1}^{K} \|\mathbf{Z}_k^y - \mathbf{W}_k \mathbf{Z}_k^x\|_F^2 + \left(\frac{\gamma}{\beta}\right) \sum_{k=1}^{K} \|\mathbf{W}_k\|_F^2$$
$$+ Tr(\sum_{k=1}^{K} \mathbf{Z}_k^y \mathbf{M}(\mathbf{W}_k \mathbf{Z}_k^x)^T), \tag{11}$$

where Eq. (11) is a ridge regression problem with a regularization term. We simplify the regularization term $\mathcal{R}(tr) = Tr(\sum_{k=1}^{K} \mathbf{Z}_k^y \mathbf{M}(\mathbf{W}_k \mathbf{Z}_k^x)^T)$ and analytically derive the solution as $\mathbf{W} = (\mathbf{Z}_k^y \mathbf{Z}_k^{xT} - \mathcal{R}(tr))(\mathbf{Z}_k^x \mathbf{Z}_k^{xT} + \frac{\gamma}{\beta}\mathbf{I})^{-1}$, where $\mathbf{I}$ is an identity matrix.

Algorithm 1: WEENIE Algorithm

Input: Training data $\mathbf{X}$ and $\mathbf{Y}$, parameters λ, γ, σ.
1   Initialize $\mathbf{F}_0^x, \mathbf{F}_0^y, \mathbf{Z}_0^x, \mathbf{Z}_0^y, \mathbf{U}_0^x, \mathbf{U}_0^y, \mathbf{S}_0^x, \mathbf{S}_0^y, \mathbf{W}_0$.
2   Perform FFT $\mathbf{Z}_0^x \to \hat{\mathbf{Z}}_0^x, \mathbf{Z}_0^y \to \hat{\mathbf{Z}}_0^y, \mathbf{F}_0^x \to \hat{\mathbf{F}}_0^x$,
   $\mathbf{F}_0^y \to \hat{\mathbf{F}}_0^y, \mathbf{U}_0^x \to \hat{\mathbf{U}}_0^x, \mathbf{U}_0^y \to \hat{\mathbf{U}}_0^y, \mathbf{S}_0^x \to \hat{\mathbf{S}}_0^x$,
   $\mathbf{S}_0^y \to \hat{\mathbf{S}}_0^y$.
3   Let $\hat{\mathbf{Z}}_0^y \leftarrow \mathbf{W}\hat{\mathbf{Z}}_0^x$.
4   **while** *not converged* **do**
5   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 (9).
6   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 (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
     $\mathbf{W}_k$.
7   Fix other variables, update $\mathbf{W}_k$ by (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
     $\hat{\mathbf{S}}_{k+1}^x, \hat{\mathbf{S}}_{k+1}^y$.
8   Inverse FFT $\hat{\mathbf{F}}_{k+1}^x \to \mathbf{F}_{k+1}^x, \hat{\mathbf{F}}_{k+1}^y \to \mathbf{F}_{k+1}^y$.
9   **end**
   Output: $\mathbf{F}^x, \mathbf{F}^y, \mathbf{W}$.

3.6. 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 $\mathcal{X}$, we can synthesize its desirable HR version of style $\mathcal{Y}$. 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 $\mathcal{M}_2$ 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 2):

$$\mathbf{Y}^t = \sum_{k=1}^{K} \mathbf{F}_k^y \mathbf{W}_k \mathbf{Z}_k^t = \sum_{k=1}^{K} \mathbf{F}_k^y \hat{\mathbf{Z}}_k^t. \tag{12}$$

Algorithm 2: SRCMS

Input: Test image $\mathbf{X}^t$, filter pairs $\mathbf{F}^x$ and $\mathbf{F}^y$,
     mapping $\mathbf{W}$.
1   Initialize $\mathbf{Z}_0^t$.
2   Let $\hat{\mathbf{Z}}_0^t \leftarrow \mathbf{W}\mathbf{Z}_0^t, \mathbf{Y}_0^t \leftarrow \mathbf{F}^y \mathbf{W}\mathbf{Z}_0^t$.
3   **while** *not converged* **do**
4   Update $\mathbf{Z}_{k+1}^t$ and $\hat{\mathbf{Z}}_{k+1}^t$ by (9) with $\mathbf{Y}_k^t$, and $\mathbf{W}$.
5   Update $\mathbf{Y}_{k+1}^t \leftarrow \mathbf{W}\hat{\mathbf{Z}}_{k+1}^t$.
6   **end**
7   Synthesize $\mathbf{Y}^t$ by (12).
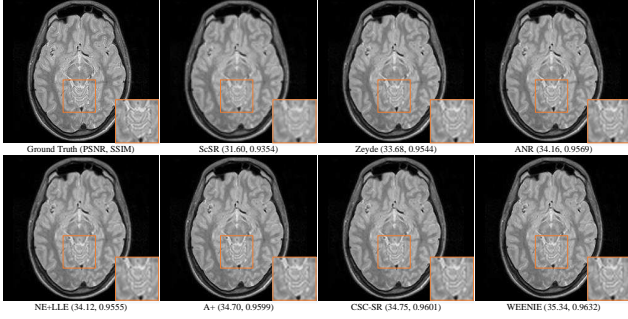   **Output**: Synthesized image $\mathbf{Y}^t$.

Figure 2. Example SR results and corresponding PSNRs, SSIMs (zoom in for details).

Figure 3. Performance comparisons of different SR approaches.

Metric(avg.)	ScSR [38]	Zeyde [40]	ANR [31]	NE+LLE [3]	A+ [32]	CSC-SR [11]	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 1. Quantitative evaluation (PSNR and SSIM): WEENIE vs. other SR methods on 95 subjects of the IXI dataset.

4. Experimental Results

We conduct the experiments using two datasets, *i.e.*, IXI[3] and NAMIC brain mutlimodality[4] datasets. Following [11, 35, 38], 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 **F**, **S**, **Z**, and **U** are randomly initialized with Gaussian noise considering [2]. Generally, a larger number of filters leads to better results. To balance between computation complexity and result quality, we learn 800 filters following [11]. 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 [37] indices to objectively assess the quality of the synthesized images.

Experimental Data: The IXI dataset consists of 578 $256 \times 256 \times n$ 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 SR-CMS. Further, we conduct SRCMS experiments on the processed NAMIC dataset, which consists of 20 $128 \times 128 \times m$ 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
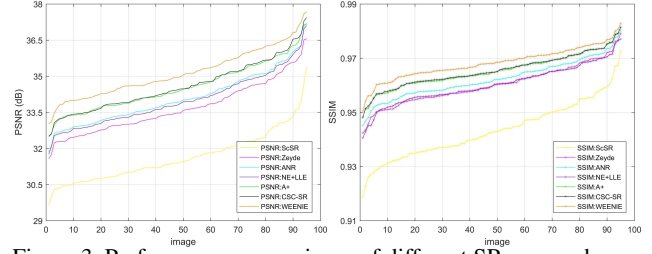
for IXI, and 3 registered subject pairs for NAMIC, respectively, from the half of the corresponding dataset for training purposes, and process the reminding training data to be unpaired. Particularly, all the existing methods with respect to cross-modality synthesis in brain imaging request a pre-processing, *i.e.*, skull stripping and/or bias corrections, as done in [34, 26]. We follow such processes and further validate whether pre-processing (especially skull stripping) is always helpful for brain image synthesis.

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) [38], anchored neighborhood regression method (ANR) [31], neighbor embedding + locally linear embedding method (NE+LLE) [3], Zeyde's method [40], convolutional sparse coding-based SR method (CSC-SR) [11], and adjusted anchored neighborhood regression method (A+) [32]. We perform image SR with scaling factor 2, and show visual results on an example slice in Fig. 2. The quantitative results for different methods are shown in Fig. 3, and the average PSNR and SSIM for all 95 test subjects are listed in Table 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.

4.2. 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 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 correc-
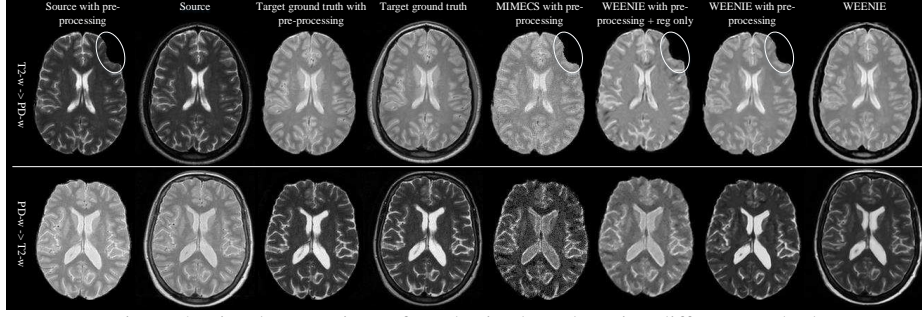
---

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

Figure 4. Visual comparison of synthesized results using different methods.

Metric(avg.)	IXI							
	PD−>T2	T2−>PD	PD−>T2+PRE			T2−>PD+PRE		
	WEENIE	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**

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 2. Quantitative evaluation (PSNR and SSIM): WEENIE vs. other synthesis methods on IXI and NAMIC datasets.
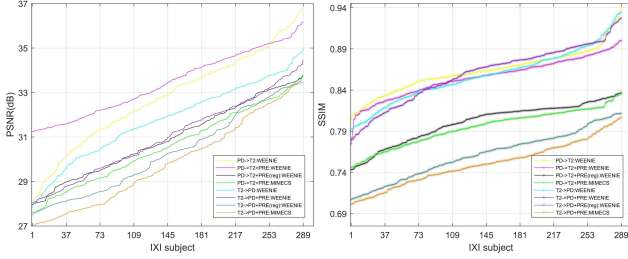

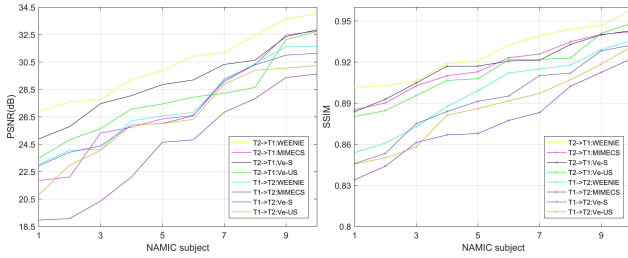Figure 5. SRCMS results: WEENIE vs. MIMECS on IXI dataset.

Figure 6. SRCMS: WEENIE vs. MIMECS on NAMIC dataset.

tions) 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) [34], Vemulapalli's unsupervised approach (V-US) [34] and MR image exampled-based contrast synthesis (MIMECS) [26] 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 Vemulapallis 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 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. 5, and Fig. 6, and summarize the averaged values in Table 2, 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.

5. Conclusion

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.

References

[1] K. Bahrami, F. Shi, X. Zong, H. W. Shin, H. An, and D. Shen. 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, 2015. 3

[2] H. Bristow, A. Eriksson, and S. Lucey. Fast convolutional sparse coding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 391–398, 2013. 3, 5, 6, 7

[3] H. Chang, D.-Y. Yeung, and Y. Xiong. 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, 2004. 2, 4, 7

[4] L. Chen, W. Li, and D. Xu. Recognizing rgb images by learning from rgb-d data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1418–1425, 2014. 5

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

[6] A. A. Efros and W. T. Freeman. Image quilting for texture synthesis and transfer. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 341–346. ACM, 2001. 2

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

[8] W. T. Freeman, E. C. Pasztor, and O. T. Carmichael. Learning low-level vision. *International journal of computer vision*, 40(1):25–47, 2000. 2

[9] X. Gao, N. Wang, D. Tao, and X. Li. Face sketch–photo synthesis and retrieval using sparse representation. *IEEE Transactions on circuits and systems for video technology*, 22(8):1213–1226, 2012. 2

[10] L. A. Gatys, A. S. Ecker, and M. Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2414–2423, 2016. 2, 3

[11] S. Gu, W. Zuo, Q. Xie, D. Meng, X. Feng, and L. Zhang. Convolutional sparse coding for image super-resolution. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1823–1831, 2015. 7

[12] F. Heide, W. Heidrich, and G. Wetzstein. Fast and flexible convolutional sparse coding. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5135–5143. IEEE, 2015. 5

[13] A. Hertzmann, C. E. Jacobs, N. Oliver, B. Curless, and D. H. Salesin. Image analogies. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 327–340. ACM, 2001. 2

[14] D.-A. Huang and Y.-C. Frank Wang. 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, 2013. 2

[15] J.-B. Huang, A. Singh, and N. Ahuja. Single image super-resolution from transformed self-exemplars. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5197–5206. IEEE, 2015. 2

[16] Y. Huang, F. Zhu, L. Shao, and A. F. Frangi. 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, 2016. 4

[17] J. E. Iglesias, E. Konukoglu, D. Zikic, B. Glocker, K. Van Leemput, and B. Fischl. Is synthesizing mri contrast useful for inter-modality analysis? In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 631–638. Springer, 2013. 3

[18] K. Jia, X. Wang, and X. Tang. Image transformation based on learning dictionaries across image spaces. *IEEE transactions on pattern analysis and machine intelligence*, 35(2):367–380, 2013. 4

[19] A. Jog, S. Roy, A. Carass, and J. L. Prince. Magnetic resonance image synthesis through patch regression. In *2013 IEEE 10th International Symposium on Biomedical Imaging*, pages 350–353. IEEE, 2013. 2

[20] R. Keys. Cubic convolution interpolation for digital image processing. *IEEE transactions on acoustics, speech, and signal processing*, 29(6):1153–1160, 1981. 1, 2

[21] X. Li and M. T. Orchard. New edge-directed interpolation. *IEEE transactions on image processing*, 10(10):1521–1527, 2001. 2

[22] M. Long, G. Ding, J. Wang, J. Sun, Y. Guo, and P. S. Yu. Transfer sparse coding for robust image representation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 407–414, 2013. 5

[23] F. Monay and D. Gatica-Perez. Modeling semantic aspects for cross-media image indexing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(10):1802–1817, 2007. 2

[24] S. G. Mueller, M. W. Weiner, L. J. Thal, R. C. Petersen, C. Jack, W. Jagust, J. Q. Trojanowski, A. W. Toga, and L. Beckett. The alzheimer's disease neuroimaging initiative. *Neuroimaging Clinics of North America*, 15(4):869–877, 2005. 1

[25] F. Rousseau, A. D. N. Initiative, et al. A non-local approach for image super-resolution using intermodality priors. *Medical image analysis*, 14(4):594–605, 2010. 1, 2

[26] S. Roy, A. Carass, and J. L. Prince. Magnetic resonance image example-based contrast synthesis. *IEEE transactions on medical imaging*, 32(12):2348–2363, 2013. 2, 7, 8

[27] A. Rueda, N. Malpica, and E. Romero. Single-image super-resolution of brain mr images using overcomplete dictionaries. *Medical image analysis*, 17(1):113–132, 2013. 2

[28] L. Shao and M. Zhao. Order statistic filters for image interpolation. In *2007 IEEE International Conference on Multimedia and Expo*, pages 452–455. IEEE, 2007. 1, 2

[29] A. W. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Transactions on pattern analysis and machine intelligence*, 22(12):1349–1380, 2000. 2

[30] Y. Tang and L. Shao. Pairwise operator learning for patch-based single-image super-resolution. *IEEE Transactions on Image Processing*, 26(2):994–1003, 2017. 1

[31] R. Timofte, V. De Smet, and L. Van Gool. Anchored neighborhood regression for fast example-based super-resolution. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1920–1927, 2013. 7

[32] R. Timofte, V. De Smet, and L. Van Gool. A+: Adjusted anchored neighborhood regression for fast super-resolution. In *Asian Conference on Computer Vision*, pages 111–126. Springer, 2014. 7

[33] H. Van Nguyen, K. Zhou, and R. Vemulapalli. 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, 2015. 2, 3

[34] R. Vemulapalli, H. Van Nguyen, and S. Kevin Zhou. Unsupervised cross-modal synthesis of subject-specific scans. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 630–638, 2015. 2, 3, 7, 8

[35] S. Wang, L. Zhang, Y. Liang, and Q. Pan. 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, 2012. 2, 6, 7

[36] X. Wang and X. Tang. Face photo-sketch synthesis and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(11):1955–1967, 2009. 2

[37] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 7

[38] J. Yang, J. Wright, T. S. Huang, and Y. Ma. Image super-resolution via sparse representation. *IEEE transactions on image processing*, 19(11):2861–2873, 2010. 2, 4, 7

[39] M. D. Zeiler, D. Krishnan, G. W. Taylor, and R. Fergus. Deconvolutional networks. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2528–2535. IEEE, 2010. 3, 4

[40] R. Zeyde, M. Elad, and M. Protter. On single image scale-up using sparse-representations. In *International conference on curves and surfaces*, pages 711–730. Springer, 2010. 7

[41] L. Zhang and X. Wu. An edge-guided image interpolation algorithm via directional filtering and data fusion. *IEEE transactions on Image Processing*, 15(8):2226–2238, 2006. 2

[42] F. Zheng, Y. Tang, and L. Shao. Hetero-manifold regularisation for cross-modal hashing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2016. 5