This is a repository copy of *Cross-Modality Image Synthesis via Weakly Coupled and Geometry Co-Regularized Joint Dictionary Learning*.

White Rose Research Online URL for this paper:
http://eprints.whiterose.ac.uk/135012/

Version: Accepted Version

Cross-Modality Image Synthesis via Weakly-Coupled and Geometry Co-Regularized Joint Dictionary Learning

Yawen Huang, *Student Member, IEEE,* Ling Shao, *Senior Member, IEEE,* and Alejandro F. Frangi, *Fellow, IEEE*

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, multi-modal 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. In this paper, 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.**

Index Terms—**Dictionary Learning, Sparse Representation, Image Synthesis, Domain Adaption, Manifold Learning, MRI.**

I. 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 T1weighted 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 [1] and histogram matching [2, 3] 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 [4, 5]. 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 [6, 7], segmentation [8], and atlas construction [9, 10].

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 [4, 11–13] or that learn the mapping from a large set of training image pairs [14–17]. Although these approaches have shown great promise, they are supervised and require labeled data sets.

In this paper, instead, we propose a single-image cross-modality synthesis method with an application to T1w, T2w and PDw brain MRI that utilizes a few registered multi-modality image pairs[1] 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

Y. Huang and A.F. Frangi are with the Center for Computational Imaging and Simulation Technologies in Biomedicine (CISTIB), Department of Electronic and Electrical Engineering, The University of Sheffield, Sheffield, United Kingdom (e-mail: yhuang36@sheffield.ac.uk, a.frangi@sheffield.ac.uk).

L. Shao is with the School of Computing Sciences, University of East Anglia, Norwich, United Kingdom (e-mail: ling.shao@ieee.org).

[1]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.

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 W̲eakly-coupled A̲nd G̲eometry (WAG) co-regularized joint dictionary learning, and perform extensive experiments to verify its performance.

The contributions of this work are threefold:

(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. T1w, T2w and PDw MRI) by simultaneously minimizing the distribution discrepancy of similar instances and preserving geometric structures in each domain.

A preliminary version [5] 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.

The remainder of this paper is organized as follows. Section II reviews related work. Section III defines the cross-modality synthesis problem and introduces our proposed method. The experimental results are demonstrated in Section IV with discussions. The discussion of this work is given in Section V. Section VI concludes the paper.

II. RELATED WORK

To synthesize a target modality image from a source modality image, several approaches have been suggested in the literature with promising results [13, 15, 16]. Most of these methods can be broadly referred to as *example-based methods* and roughly subdivided based on the size of the training set.

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 Hertzmann *et al.*'s image analogies [11], which transfers the texture information from a source modality space onto a target modality space. The same strategy is also applied to facilitate multi-modal image registration in correlative microscopy [6]. Kroon *et al.* [18] mapped between T1w and T2w 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 [4]. Recently, Jog *et al.* [13] 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.* [14] 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 [19] 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.* [16] introduced another framework called pseudo CT synthesis for generating CT-like image from the T1w or T2w input using multi-atlas deformable registration and tissue contrast fusion. In [17], a location-sensitive deep learning-based method was proposed to explicitly utilize voxel image coordinates by incorporating image intensities and spatial information into a deep network
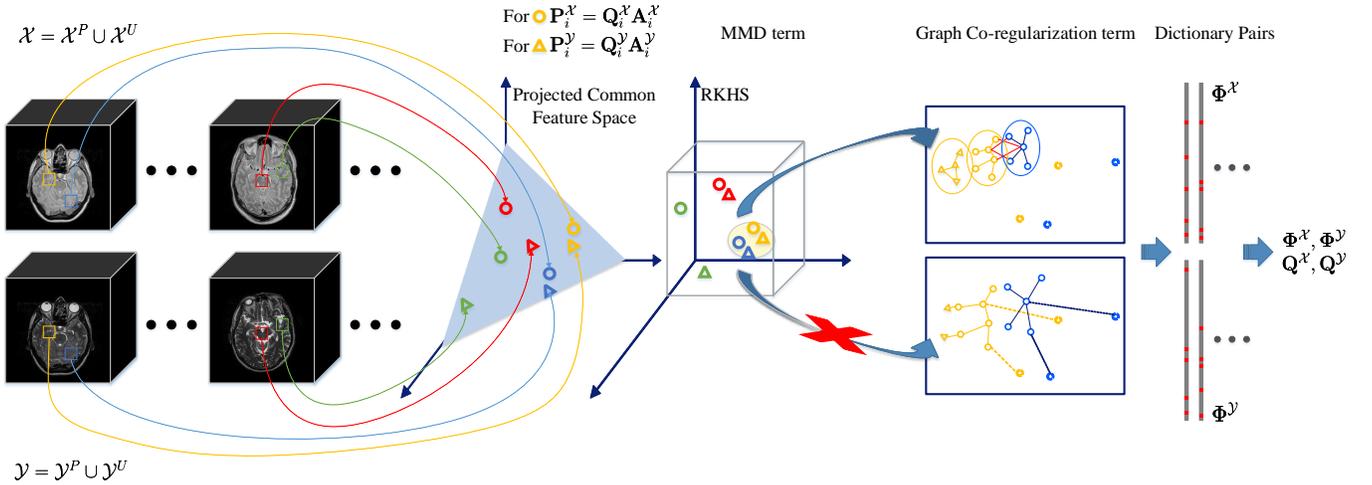
Fig. 1. Overview of the proposed cross-modality image synthesis method. Squares within the 3D images represent the extracted 3D patches with size of $5 \times 5 \times 5$. The first step is to project the extracted paired/unpaired patches into a common feature space denoted by circles for source modality 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.

for synthesis. Instead of using coupled image pairs as training data, matching feature representations and learning spatial relations with joint sparse coding [15] has shown great potential in synthesizing images across modalities. To improve the quality of the synthesized images across different modalities, Huang *et al.* [20] 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 [15], collecting a large number of multi-modality images is both time-consuming and expensive, and sometimes even impractical in medical imaging. 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 a more practical cross-modality image synthesis solution that links source-target domains in a weakly-coupled fashion, which outperforms existing state-of-the-art methods on several experimental scenarios.

III. 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. 1.

A. Problem Definition

Let $\mathcal{X} = \{\mathbf{X}_1, ..., \mathbf{X}_S\}$ be the source modality images of S subjects using modality $\mathcal{M}_1$, and $\mathcal{Y} = \{\mathbf{Y}_1, ..., \mathbf{Y}_T\}$ be the target modality images of T subjects imaged using modality $\mathcal{M}_2$. 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.*, $\mathcal{X}^P = \{\mathbf{X}_1, ..., \mathbf{X}_R\}$, $\mathcal{Y}^P = \{\mathbf{Y}_1, ..., \mathbf{Y}_R\}$, and an unregistered/unpaired domain subset of size $T - R$ or $S - R$, respectively, *i.e.* $\mathcal{X}^U = \{\mathbf{X}_{R+1}, ..., \mathbf{X}_S\}$, $\mathcal{Y}^U = \{\mathbf{Y}_{R+1}, ..., \mathbf{Y}_T\}$ so $\mathcal{X} = \mathcal{X}^P \cup \mathcal{X}^U$ and $\mathcal{Y} = \mathcal{Y}^P \cup \mathcal{Y}^U$. 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 $\mathcal{X}$ and $\mathcal{Y}$ 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 $\mathcal{M}_1$, and will be the input to synthesize the corresponding 3D patches in modality $\mathcal{M}_2$.

Problem Statement: We first denote the coding coefficients $\mathbf{A}^{\mathcal{X}}$, $\mathbf{A}^{\mathcal{Y}}$ of $\mathbf{X}$, $\mathbf{Y}$ over the learned dictionaries $\mathbf{\Phi}^{\mathcal{X}}$, $\mathbf{\Phi}^{\mathcal{Y}}$, the projected data $\mathbf{P}^{\mathcal{X}}$, $\mathbf{P}^{\mathcal{Y}}$ of $\mathbf{X}$, $\mathbf{Y}$ in a defined common space, and a mapping function $\mathcal{F}(\cdot)$ to represent the relationship between the sparse codes $\mathbf{A}^{\mathcal{X}}$, $\mathbf{A}^{\mathcal{Y}}$ 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 $\mathbf{X} = \begin{bmatrix} \mathbf{X}^P \, \mathbf{X}^U \end{bmatrix}$ and $\mathbf{Y} = \begin{bmatrix} \mathbf{Y}^P \, \mathbf{Y}^U \end{bmatrix}$, our goal is: 1) to learn a pair of dictionaries $\{\boldsymbol{\Phi}^{\mathcal{X}}, \boldsymbol{\Phi}^{\mathcal{Y}}\}$, their sparse codes $\{\mathbf{A}^{\mathcal{X}}, \mathbf{A}^{\mathcal{Y}}\}$, and an association function $\mathcal{F}(\cdot) : \mathcal{M}_1 \rightarrow \mathcal{M}_2$ using the projected data $\mathbf{P}^{\mathcal{X}}$ and $\mathbf{P}^{\mathcal{Y}}$; and 2) to minimize the inter-modality divergence between $\mathbf{P}^{\mathcal{X}}$ and $\mathbf{P}^{\mathcal{Y}}$, and 3) to preserve the domain-specific local geometric structure.

B. 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, $\mathbf{A}^{\mathcal{X}} = \begin{bmatrix} \boldsymbol{\alpha}_1^{\mathcal{X}}, ..., \boldsymbol{\alpha}_n^{\mathcal{X}} \end{bmatrix} \in \mathbb{R}^{k \times n}$ is associated to the dictionary $\boldsymbol{\Phi}^{\mathcal{X}} = \begin{bmatrix} \boldsymbol{\phi}_1^{\mathcal{X}}, ..., \boldsymbol{\phi}_k^{\mathcal{X}} \end{bmatrix} \in \mathbb{R}^{m \times k}$. Here, $k > m$ to make the dictionary over-complete [21]. Considering the reconstruction error for each data point, the problem of learning a dictionary $\boldsymbol{\Phi}^{\mathcal{X}}$ for sparse representation of $\mathbf{X}$ can be formulated as

$$\min_{\boldsymbol{\Phi}^{\mathcal{X}}, \mathbf{A}^{\mathcal{X}}} \left\| \mathbf{X} - \boldsymbol{\Phi}^{\mathcal{X}} \mathbf{A}^{\mathcal{X}} \right\|_F^2 + \lambda \left\| \mathbf{A}^{\mathcal{X}} \right\|_0, \quad (1)$$

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 [22], the minimization problem in Eq. (1) 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 [23]. The dictionary learning problem in Eq. (1) can be reformulated as

$$\min_{\boldsymbol{\Phi}^{\mathcal{X}}, \mathbf{A}^{\mathcal{X}}} \left\| \mathbf{X} - \boldsymbol{\Phi}^{\mathcal{X}} \mathbf{A}^{\mathcal{X}} \right\|_F^2 + \lambda \left\| \mathbf{A}^{\mathcal{X}} \right\|_1. \quad (2)$$

The above objective function is not simultaneously convex over $\boldsymbol{\Phi}$ and $\mathbf{A}$. A practical solution is to alternate between optimizing for the dictionary $\boldsymbol{\Phi}$ 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 [24]. When the dictionary is fixed, the algorithm is known as Lasso/LARS [25] with an l_1 penalty over the coefficients and can be solved by the feature-sign search approach [24]. 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 [24].

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 [21, 26] and Eq. (2), we can learn the dictionaries separately to obtain the two dictionaries, $\boldsymbol{\Phi}^{\mathcal{X}}$ and $\boldsymbol{\Phi}^{\mathcal{Y}}$, and the two corresponding sparse coefficients, $\mathbf{A}^{\mathcal{X}}$ and $\mathbf{A}^{\mathcal{Y}}$, respectively. The data of each modality can be reconstructed using the respective dictionary and associated sparse coefficients.

C. Cross-Modality Dictionary Learning

Cross-modality image synthesis is based on learning a joint sparse representation [19] with a common set of sparse codes shared between source and target image modalities, i.e. $\mathbf{A}^{\mathcal{X}} \equiv \mathbf{A}^{\mathcal{Y}}$. These sparse codes act on independent dictionaries for each modality, viz. $\boldsymbol{\Phi}^{\mathcal{X}}$ and $\boldsymbol{\Phi}^{\mathcal{Y}}$, 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. $\left\| \mathbf{X} - \boldsymbol{\Phi}^{\mathcal{X}} \mathbf{A}^{\mathcal{X}} \right\|_F^2$ and $\left\| \mathbf{Y} - \boldsymbol{\Phi}^{\mathcal{Y}} \mathbf{A}^{\mathcal{Y}} \right\|_F^2$) in a single objective function to be optimized:

$$\min_{\boldsymbol{\Phi}^{\mathcal{X}}, \boldsymbol{\Phi}^{\mathcal{Y}}, \mathbf{A}} \left\| \mathbf{X} - \boldsymbol{\Phi}^{\mathcal{X}} \mathbf{A} \right\|_F^2 + \left\| \mathbf{Y} - \boldsymbol{\Phi}^{\mathcal{Y}} \mathbf{A} \right\|_F^2 + \lambda \left\| \mathbf{A} \right\|_1$$
$$\text{s.t. } \left\| \boldsymbol{\phi}_i^{\mathcal{X}} \right\|_2^2 \leq 1, \left\| \boldsymbol{\phi}_i^{\mathcal{Y}} \right\|_2^2 \leq 1 \ \forall i = 1, ..., k, \quad (3)$$

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. (3) is convex regarding the learned dictionaries, $\boldsymbol{\Phi}^{\mathcal{X}}$ and $\boldsymbol{\Phi}^{\mathcal{Y}}$, for fixed sparse codes $\mathbf{A}$. Therefore, the computation of $\mathbf{A}$ and of the dictionary pairs can be alternated. Analyzing (3), we note that this objective function is suitable to collaboratively learn a pair of dictionaries, so the sparse codes in the source modality space $\mathcal{M}_1$ can directly reconstruct the target modality image $\mathcal{M}_2$ 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, $\boldsymbol{\Phi}^{\mathcal{X}}$ and $\boldsymbol{\Phi}^{\mathcal{Y}}$, 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 paper, 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.

D. 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; and (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.

1) 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 [19].

In this paper, we follow [19, 27] and adopt first- and second- order derivatives involving horizontal and vertical gradients as the HF features for each training data by $\mathbf{X}^h = \mathcal{H} * \mathbf{X}$, $\mathbf{Y}^h = \mathcal{H} * \mathbf{Y}$. Generally, $\mathcal{H}$ is a high-pass filter operator used to extract derivatives, Considering first and second order derivatives, $\mathcal{H}$ is one of the following operators: $\mathbf{H}_1^1, \mathbf{H}_1^2, \mathbf{H}_2^1, \mathbf{H}_2^2$, where $\mathbf{H}_1^1 = [-1, 0, 1]$, $\mathbf{H}_1^2 = \mathbf{H}_1^{1T}$, and $\mathbf{H}_2^1 = [-2, -1, 0, 1, 2]$, $\mathbf{H}_2^2 = \mathbf{H}_2^{1T}$. Once the features in both domains are computed, we can use them to optimize CMIM and define a mapping $\mathcal{C}(\cdot) : \mathcal{X} \rightarrow \mathcal{Y}$. 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. [28–30]). Given associated HF image feature sets, $\mathbf{X}^h$ and $\mathbf{Y}^h$, corresponding to both paired and unpaired training image data sets, $\mathcal{X}^P$, $\mathcal{X}^U$, $\mathcal{Y}^P$ and $\mathcal{Y}^U$, CMIM represents an ensemble of paired and unpaired cross-modality matching sub-problems. Images in $\mathcal{X}^P$ and $\mathcal{Y}^P$ are endowed with a natural correspondence, $\mathcal{X}^P \rightleftharpoons \mathcal{Y}^P$. In contrast, CMIM finds a mapping for multi-modality unpaired image data for $\mathcal{X}^U$ and $\mathcal{Y}^U$. Since $\mathcal{X}^P$ and $\mathcal{Y}^P$ 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 $\mathcal{C}(\mathcal{X}, \mathcal{Y}) = \left\| \mathbf{X}^h - \mathbb{T} \mathbf{Y}^h \right\|_2^2$:

$$\mathbb{T} = \begin{bmatrix} D(\mathbf{x}_1^h, \mathbf{y}_1^h) & \cdots & D(\mathbf{x}_1^h, \mathbf{y}_n^h) \\ \vdots & \ddots & \vdots \\ D(\mathbf{x}_n^h, \mathbf{y}_1^h) & \cdots & D(\mathbf{x}_n^h, \mathbf{y}_n^h) \end{bmatrix}, \quad (4)$$

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 $\mathcal{X}$ and $\mathcal{Y}$ using an g-dimensional Gaussian kernel

$$D(\mathbf{x}_i^h, \mathbf{y}_j^h) = \frac{1}{(\sqrt{2\pi}\sigma)^g} e^{-\frac{\left\| \mathbf{x}_i^h - \mathbf{y}_j^h \right\|^2}{2\sigma^2}}, \quad (5)$$

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 $\mathcal{Y}$ while discarding other 3D patches. In this way, from $\mathbb{T}$ we define $\hat{\mathbb{T}}$ as:

$$\hat{\mathbb{T}}(i, j) = \begin{cases} 1, & \text{if } j = j_i, \\ 0, & \text{otherwise.} \end{cases} \quad (6)$$

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.*, $\mathcal{X} \rightleftharpoons \mathcal{Y}$ for each $\mathbf{x}_i$ paired with $\mathbf{y}_{j_i}$ denoted as $\mathcal{P}_i = \{\mathbf{x}_i, \mathbf{y}_{j_i}\}$ for $i = 1 \ldots n$.

2) Computing the mapping function: Starting off by Eq. (2), by minimizing the reconstruction error, the corresponding sparse codes $\mathbf{A}^{\mathcal{X}}$ and $\mathbf{A}^{\mathcal{Y}}$ 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 $\mathcal{F}: \mathcal{M}_1 \rightarrow \mathcal{M}_2$ with $\mathbf{Y} = \mathcal{F}(\mathbf{X})$. Accordingly, the sparse codes of $\mathbf{X}$ and $\mathbf{Y}$ over the dictionaries will be related by such a mapping function $\mathcal{F}(\mathbf{A}^{\mathcal{X}}, \mathbf{A}^{\mathcal{Y}})$. To build a stable mapping between two domains, Wang *et al.* [31] 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 [32], 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 $\mathbf{P}^{\mathcal{X}}$, $\mathbf{P}^{\mathcal{Y}}$ of $\mathbf{X}$, $\mathbf{Y}$, and replace $\mathcal{F}(\mathbf{A}^{\mathcal{X}}, \mathbf{A}^{\mathcal{Y}})$ by $\mathcal{F}(\mathbf{P}^{\mathcal{X}}, \mathbf{P}^{\mathcal{Y}})$, and then incorporate the projected features into CMIM-driven coupled dictionary learning. The objective function of this learning model is:

$$\min_{\mathbf{\Phi}^{\mathcal{X}}, \mathbf{\Phi}^{\mathcal{Y}}, \mathbf{A}^{\mathcal{X}}, \mathbf{A}^{\mathcal{Y}}} \left\| \mathbf{X} - \mathbf{\Phi}^{\mathcal{X}} \mathbf{A}^{\mathcal{X}} \right\|_F^2 + \left\| \mathbf{Y} - \mathbf{\Phi}^{\mathcal{Y}} \mathbf{A}^{\mathcal{Y}} \right\|_F^2$$
$$+ \lambda \left(\left\| \mathbf{A}^{\mathcal{X}} \right\|_1 + \left\| \mathbf{A}^{\mathcal{Y}} \right\|_1 \right) \quad (7)$$
$$+ \left\| \mathbf{X}^h - \hat{\mathbb{T}} \mathbf{Y}^h \right\|_2^2 + \nu \mathcal{F}\left(\mathbf{P}^{\mathcal{X}}, \mathbf{P}^{\mathcal{Y}}\right),$$

where $\mathbf{P}^{\mathcal{X}} = \mathbf{Q}^{\mathcal{X}} \mathbf{A}^{\mathcal{X}} \in \mathbb{R}^{k \times n}$ and $\mathbf{P}^{\mathcal{Y}} = \mathbf{Q}^{\mathcal{Y}} \mathbf{A}^{\mathcal{Y}} \in \mathbb{R}^{k \times n}$ 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, $\mathbf{Q}^{\mathcal{X}} \in \mathbb{R}^{k \times k}$ and $\mathbf{Q}^{\mathcal{Y}} \in \mathbb{R}^{k \times k}$ are the projection matrices for $\mathbf{A}^{\mathcal{X}}$ and $\mathbf{A}^{\mathcal{Y}}$, respectively. Generally, $\mathcal{F}(\mathbf{P}^{\mathcal{X}}, \mathbf{P}^{\mathcal{Y}})$ can be applied to any joint dictionary learning scheme with $\mathcal{F}(\mathbf{P}^{\mathcal{X}}, \mathbf{P}^{\mathcal{Y}}) = \left\| \mathbf{P}^{\mathcal{X}} - \mathbf{P}^{\mathcal{Y}} \right\|_F^2 = \left\| \mathbf{Q}^{\mathcal{X}} \mathbf{A}^{\mathcal{X}} - \mathbf{Q}^{\mathcal{Y}} \mathbf{A}^{\mathcal{Y}} \right\|_F^2$. For example, in Eq. (3) of [19], $\mathcal{F}$ is defined with an infinitely large ν having $\mathbf{Q}^{\mathcal{X}} = \mathbf{Q}^{\mathcal{Y}} = \mathbf{I}$, while in [31] $\mathcal{F}$ is defined so $\mathbf{Q}^{\mathcal{X}} = \mathbf{I}$ and $\mathbf{Q}^{\mathcal{Y}} = \mathbf{W}$, where $\mathbf{I}$ is the identity matrix. The solutions of $\mathbf{Q}^{\mathcal{X}}$ and $\mathbf{Q}^{\mathcal{Y}}$ are not unique. Following [32], 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 $\mathbf{P}^{\mathcal{X}}$ and $\mathbf{P}^{\mathcal{Y}}$, we minimize their distance in the projected common space considering the projections separately, *viz.* $\nu \left(\left\| \mathbf{A}^{\mathcal{X}} \mathbf{Q}^{\mathcal{X}} - \mathbf{P}^{\mathcal{Y}} \right\|_F^2 + \left\| \mathbf{A}^{\mathcal{Y}} \mathbf{Q}^{\mathcal{Y}} - \mathbf{P}^{\mathcal{X}} \right\|_F^2 \right)$. Solving this objective function, we obtain $\mathbf{A}^{\mathcal{X}} = \mathbf{Q}^{\mathcal{X}^{-1}} \mathbf{P}^{\mathcal{Y}}$ and $\mathbf{A}^{\mathcal{Y}} = \mathbf{Q}^{\mathcal{Y}^{-1}} \mathbf{P}^{\mathcal{X}}$, where $\mathbf{P}^{\mathcal{X}} = \mathbf{Q}^{\mathcal{X}} \mathbf{A}^{\mathcal{X}}$ and $\mathbf{P}^{\mathcal{Y}} = \mathbf{Q}^{\mathcal{Y}} \mathbf{A}^{\mathcal{Y}}$ denote the projected data of $\mathbf{X}$ and $\mathbf{Y}$, respectively, in the constructed common feature space.

3) 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) [33, 34] over all matched image pairs. MMD is a nonparametric statistic utilized to assess whether two samples are drawn from the same distribution. In this paper, we seek that the probability distributions of the projected data $\mathbf{P}^{\mathcal{X}}$ and $\mathbf{P}^{\mathcal{Y}}$ are identical in the common HF feature space. To this effect, we follow [33, 35, 36] and estimate the largest difference of $\mathbf{P}^{\mathcal{X}}$ and $\mathbf{P}^{\mathcal{Y}}$ in expectations over functions in the unit ball of a reproducing kernel Hilbert space:

$$
\frac{1}{n^2} \left\| \sum_{i=1}^n \mathbf{p}_i^{\mathcal{X}} + \sum_{j=1}^n \mathbf{p}_j^{\mathcal{Y}} \right\|_H^2 = \sum_{i,j=1}^n \mathbf{p}_i^{\mathcal{X}^{\mathrm{T}}} m_{i,j} \mathbf{p}_j^{\mathcal{Y}}
$$
$$
= \mathrm{Tr} \left(\mathbf{P}^{\mathcal{X}^{\mathrm{T}}} \mathbf{M} \mathbf{P}^{\mathcal{Y}} \right), \tag{8}
$$

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

$$
\mathbf{m}_{i,j} = \begin{cases} 1/n^2, & \text{if } j = j_i, \text{ hence, } \{\mathbf{p}_i, \mathbf{p}_{j_i}\} \in \mathcal{P}_i, \\ -1/n^2, & \text{otherwise.} \end{cases} \tag{9}
$$

The objective function is then rewritten by incorporating the MMD regularization term into Eq. (7).

4) 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 paper, 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.* [37] and Zheng *et al.* [38] 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 $\mathbf{P}^{\mathcal{X}}$ and $\mathbf{P}^{\mathcal{Y}}$ of $\mathbf{X}$ and $\mathbf{Y}$, respectively, one can construct two q-nearest neighbor graphs, $\mathcal{G}^{\mathcal{X}}$ and $\mathcal{G}^{\mathcal{Y}}$, with n vertices each based on prior work by [38]. The weight matrices $\mathbf{W}^{\mathcal{X}}$ and $\mathbf{W}^{\mathcal{Y}}$ of $\mathbf{G}^{\mathcal{X}}$ and $\mathbf{G}^{\mathcal{Y}}$ are then defined as the matrices with elements $w_{i,j}^{\mathcal{X}} = 1$ and $w_{i,j}^{\mathcal{Y}} = 1$ if and only if for any two features $\mathbf{p}_i^{\mathcal{X}}, \mathbf{p}_j^{\mathcal{X}}$ or $\mathbf{p}_i^{\mathcal{Y}}, \mathbf{p}_j^{\mathcal{Y}}$ satisfying: $\mathbf{p}_i^{\mathcal{X}}$ or $\mathbf{p}_i^{\mathcal{Y}}$ is among the q-nearest neighbors of $\mathbf{p}_j^{\mathcal{X}}$ or $\mathbf{p}_j^{\mathcal{Y}}$, otherwise $w_{i,j}^{\mathcal{X}} = 0$ or $w_{i,j}^{\mathcal{Y}} = 0$. Let $\mathbf{D}^{\mathcal{X}} = \mathrm{diag}\left(d_1^{\mathcal{X}}, \cdots, d_n^{\mathcal{X}} \right)$ and $\mathbf{D}^{\mathcal{Y}} = \mathrm{diag}\left(d_1^{\mathcal{Y}}, \cdots, d_n^{\mathcal{Y}} \right)$ be the degree matrices of $\mathbf{P}^{\mathcal{X}}$ and $\mathbf{P}^{\mathcal{Y}}$, with elements $d_j^{\mathcal{X}} = \sum_{i=1}^n w_{i,j}^{\mathcal{X}}$ and $d_j^{\mathcal{Y}} = \sum_{i=1}^n w_{i,j}^{\mathcal{Y}}$. Based on the graph Laplacian [39], we can define $\mathbf{G}^{\mathcal{X}} = \mathbf{D}^{\mathcal{X}} - \mathbf{W}^{\mathcal{X}}$ and $\mathbf{G}^{\mathcal{Y}} = \mathbf{D}^{\mathcal{Y}} - \mathbf{W}^{\mathcal{Y}}$, respectively. Considering the case of mapping the graphs $\mathbf{G}^{\mathcal{X}}$ and $\mathbf{G}^{\mathcal{Y}}$ to the projected features $\mathbf{P}^{\mathcal{X}}$ and $\mathbf{P}^{\mathcal{Y}}$, a reasonable criterion [40] for preserving the domain-specific geometrical strictures is designed by minimizing the following objective function:

$$
\frac{1}{2} \sum_{i,j=1}^n \left(w_{i,j}^{\mathcal{X}} \left\| \mathbf{p}_i^{\mathcal{X}} - \mathbf{p}_j^{\mathcal{X}} \right\|_2^2 + w_{i,j}^{\mathcal{Y}} \left\| \mathbf{p}_i^{\mathcal{Y}} - \mathbf{p}_j^{\mathcal{Y}} \right\|_2^2 \right)
$$
$$
= \frac{1}{2} \sum_{i,j=1}^n \left(\mathbf{p}_i^{\mathcal{X}} \mathbf{p}_i^{\mathcal{X}^T} d_{i,i} - \mathbf{p}_i^{\mathcal{X}} \mathbf{p}_j^{\mathcal{X}^T} w_{i,j}^{\mathcal{Y}} \right.
$$
$$
\left. + \mathbf{p}_i^{\mathcal{Y}} \mathbf{p}_i^{\mathcal{Y}^T} d_{i,i} - \mathbf{p}_i^{\mathcal{Y}} \mathbf{p}_j^{\mathcal{Y}^T} w_{i,j}^{\mathcal{Y}} \right) \tag{10}
$$
$$
= \frac{1}{2} \mathrm{Tr} \left(\mathbf{P}^{\mathcal{X}} \mathbf{G}^{\mathcal{X}} \mathbf{P}^{\mathcal{X}^T} + \mathbf{P}^{\mathcal{Y}} \mathbf{G}^{\mathcal{Y}} \mathbf{P}^{\mathcal{Y}^T} \right).
$$

The regularization criterion in Eq. (10) guarantees that the projected data varies smoothly along the geodesics of the manifold defined by the corresponding graph.

5) 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. (7) leading to the final objective function:

$$
\min_{\boldsymbol{\Phi}, \mathbf{A}, \mathbf{Q}} \left\| \mathbf{X} - \boldsymbol{\Phi}^{\mathcal{X}} \mathbf{A}^{\mathcal{X}} \right\|_F^2 + \left\| \mathbf{Y} - \boldsymbol{\Phi}^{\mathcal{Y}} \mathbf{A}^{\mathcal{Y}} \right\|_F^2 + \left\| \mathbf{X}^h - \hat{\mathbf{T}} \mathbf{Y}^h \right\|_2^2
$$
$$
+ \nu \left(\left\| \mathbf{A}^{\mathcal{X}} \mathbf{Q}^{\mathcal{X}} - \mathbf{P}^{\mathcal{Y}} \right\|_F^2 + \left\| \mathbf{A}^{\mathcal{Y}} \mathbf{Q}^{\mathcal{Y}} - \mathbf{P}^{\mathcal{X}} \right\|_F^2 \right)
$$
$$
+ \lambda \left(\left\| \mathbf{A}^{\mathcal{X}} \right\|_1 + \left\| \mathbf{A}^{\mathcal{Y}} \right\|_1 \right) + \gamma \, \mathrm{Tr} \left(\mathbf{P}^{\mathcal{X}^T} \mathbf{M} \mathbf{P}^{\mathcal{Y}} \right)
$$
$$
+ \frac{\mu}{2} \mathrm{Tr} \left(\mathbf{P}^{\mathcal{X}} \mathbf{G}^{\mathcal{X}} \mathbf{P}^{\mathcal{X}^T} + \mathbf{P}^{\mathcal{Y}} \mathbf{G}^{\mathcal{Y}} \mathbf{P}^{\mathcal{Y}^T} \right), \tag{11}
$$

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

E. Optimization

Similarly to existing joint dictionary learning methods [31, 32, 37], the optimization problem of Eq. (11) 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.

1) Computing Sparse Codes: We initialize the dictionary pair $\boldsymbol{\Phi}^{\mathcal{X}}, \boldsymbol{\Phi}^{\mathcal{Y}}$ and the projection matrices $\mathbf{Q}^{\mathcal{X}}, \mathbf{Q}^{\mathcal{Y}}$, fix them, and solve for $\mathbf{A}^{\mathcal{X}}$ and $\mathbf{A}^{\mathcal{Y}}$. Particularly, $\boldsymbol{\Phi}^{\mathcal{X}}$ and $\boldsymbol{\Phi}^{\mathcal{Y}}$ can be simply initialized as two random matrices (or use PCA or DCT bases), and $\mathbf{Q}^{\mathcal{X}}, \mathbf{Q}^{\mathcal{Y}}$ can be initialized to two identity matrices. Unlike conventional sparse coding, two additional terms are related to the projected feature space. Given $\boldsymbol{\Phi}^{\mathcal{X}}, \boldsymbol{\Phi}^{\mathcal{Y}}$ and $\mathbf{Q}^{\mathcal{X}}, \mathbf{Q}^{\mathcal{Y}}$, we can rewrite Eq. (11) as follows:

$$
\min_{\mathbf{A}^{\mathcal{X}}} \left\| \mathbf{X} - \boldsymbol{\Phi}^{\mathcal{X}} \mathbf{A}^{\mathcal{X}} \right\|_F^2 + \left\| \mathbf{X}^h - \hat{\mathbf{T}} \mathbf{Y}^h \right\|_2^2 + \nu \left\| \mathbf{A}^{\mathcal{X}} \mathbf{Q}^{\mathcal{X}} - \mathbf{P}^{\mathcal{Y}} \right\|_F^2
$$
$$
+ \lambda \left\| \mathbf{A}^{\mathcal{X}} \right\|_1 + \mathrm{Tr} \left(\gamma \mathbf{P}^{\mathcal{X}^T} \mathbf{M} \mathbf{P}^{\mathcal{Y}} + \frac{\mu}{2} \mathbf{P}^{\mathcal{X}} \mathbf{G}^{\mathcal{X}} \mathbf{P}^{\mathcal{X}^T} \right),
$$
$$
\min_{\mathbf{A}^{\mathcal{Y}}} \left\| \mathbf{Y} - \boldsymbol{\Phi}^{\mathcal{Y}} \mathbf{A}^{\mathcal{Y}} \right\|_F^2 + \left\| \mathbf{X}^h - \hat{\mathbf{T}} \mathbf{Y}^h \right\|_2^2 + \nu \left\| \mathbf{A}^{\mathcal{Y}} \mathbf{Q}^{\mathcal{Y}} - \mathbf{P}^{\mathcal{X}} \right\|_F^2
$$
$$
+ \lambda \left\| \mathbf{A}^{\mathcal{Y}} \right\|_1 + \mathrm{Tr} \left(\gamma \mathbf{P}^{\mathcal{X}^T} \mathbf{M} \mathbf{P}^{\mathcal{Y}} + \frac{\mu}{2} \mathbf{P}^{\mathcal{Y}} \mathbf{G}^{\mathcal{Y}} \mathbf{P}^{\mathcal{Y}^T} \right). \tag{12}
$$

However, the problem in Eq. (12) is non-differentiable when the sparse codes take zero values. Coordinate Descent is usually adopted [21, 37, 38] to solve this l_1-regularized least squares problem. This is done by updating each vector $\alpha_i^{\mathcal{X}}$ or $\alpha_i^{\mathcal{Y}}$ individually while considering constant all other vectors $\alpha_j^{\mathcal{X}}$ or $\alpha_j^{\mathcal{Y}}$ where $j \neq i$. To optimize over each $\alpha_i^{\mathcal{X}}$ or $\alpha_i^{\mathcal{Y}}$, Eq. (12) can be expanded using vector-wise manipulations. Sparse representations in vector form can be solved by the feature-sign search algorithm [41].

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

$$\min_{\mathbf{\Phi}^{\mathcal{X}}, \mathbf{\Phi}^{\mathcal{Y}}} \left\| \mathbf{X} - \mathbf{\Phi}^{\mathcal{X}} \mathbf{A}^{\mathcal{X}} \right\|_F^2 + \left\| \mathbf{Y} - \mathbf{\Phi}^{\mathcal{Y}} \mathbf{A}^{\mathcal{Y}} \right\|_F^2 + \left\| \mathbf{X}^h - \hat{\mathbb{T}} \mathbf{Y}^h \right\|_2^2$$
$$\text{s.t. } \left\| \phi_i^{\mathcal{X}} \right\|_2^2 \leq 1, \left\| \phi_i^{\mathcal{Y}} \right\|_2^2 \leq 1 \ \forall i = \{1, ..., k\}. \tag{13}$$

The optimization in Eq. (13) can be solved by the Lagrange dual method [42].

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

$$\min_{\mathbf{Q}^{\mathcal{X}}, \mathbf{Q}^{\mathcal{Y}}} \nu(\left\| \mathbf{A}^{\mathcal{X}} \mathbf{Q}^{\mathcal{X}} - \mathbf{P}^{\mathcal{Y}} \right\|_F^2 + \left\| \mathbf{A}^{\mathcal{Y}} \mathbf{Q}^{\mathcal{Y}} - \mathbf{P}^{\mathcal{X}} \right\|_F^2). \tag{14}$$

Eq. (14) can be solved using simple ridge regression. Following [32], additional constraints, *viz.* $\delta \left(\left\| \mathbf{Q}^{\mathcal{X}} \right\|_F^2 + \left\| \mathbf{Q}^{\mathcal{Y}} \right\|_F^2 \right)$ regarding the projection matrices $\mathbf{Q}^{\mathcal{X}}$ and $\mathbf{Q}^{\mathcal{Y}}$, are imposed to avoid over-fitting. We can rewrite Eq. (14) by combining the constraints as:

$$\min_{\mathbf{Q}^{\mathcal{X}}} \nu \left\| \mathbf{A}^{\mathcal{X}} \mathbf{Q}^{\mathcal{X}} - \mathbf{P}^{\mathcal{Y}} \right\|_F^2 + \delta \left\| \mathbf{Q}^{\mathcal{X}} \right\|_F^2,$$
$$\min_{\mathbf{Q}^{\mathcal{Y}}} \nu \left\| \mathbf{A}^{\mathcal{Y}} \mathbf{Q}^{\mathcal{Y}} - \mathbf{P}^{\mathcal{X}} \right\|_F^2 + \delta \left\| \mathbf{Q}^{\mathcal{Y}} \right\|_F^2. \tag{15}$$

The solution of Eq. (15) can be analytically derived as

$$\mathbf{Q}^{\mathcal{X}} = \mathbf{P}^{\mathcal{Y}} \mathbf{A}^{\mathcal{X}T} \left(\mathbf{A}^{\mathcal{X}} \mathbf{A}^{\mathcal{X}T} + (\delta/\nu)\mathbf{I} \right)^{-1},$$
$$\mathbf{Q}^{\mathcal{Y}} = \mathbf{P}^{\mathcal{X}} \mathbf{A}^{\mathcal{Y}T} \left(\mathbf{A}^{\mathcal{Y}} \mathbf{A}^{\mathcal{Y}T} + (\delta/\nu)\mathbf{I} \right)^{-1}, \tag{16}$$

where $\mathbf{I}$ indicates an identity matrix. Algorithm 1 summarizes the proposed method.

F. 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 $\mathbf{A}^{t\mathcal{X}}$ of $\mathbf{X}^t$ related to $\mathbf{\Phi}^{\mathcal{X}}$ by solving a single sparse coding problem in Eq. (2). After that, we associate $\mathbf{A}^{t\mathcal{X}}$ to the expected sparse codes $\mathbf{A}^{t\mathcal{Y}}$ via $\mathbf{Q}^{\mathcal{X}}$ and $\mathbf{Q}^{\mathcal{Y}}$ leading to

$$\mathbf{A}^{t\mathcal{Y}} \approx \mathbf{Q}^{\mathcal{Y}-1} \mathbf{P}^{t\mathcal{X}} = \mathbf{Q}^{\mathcal{Y}-1} \mathbf{Q}^{\mathcal{X}} \mathbf{A}^{t\mathcal{X}}, \tag{17}$$

where $\mathbf{P}^{t\mathcal{X}}$ is the projected data of $\mathbf{X}^t$. Finally, the data in the target $\mathcal{M}_2$ modality, $\mathbf{Y}^t$, can be synthesized by $\mathbf{Y}^t = \mathbf{A}^{t\mathcal{Y}} \mathbf{\Phi}^{\mathcal{Y}}$.

Algorithm 1: WAG Algorithm

Input: Training data $\mathbf{X}$ and $\mathbf{Y}$, parameters $\lambda, \mu, \sigma, \gamma$.

1 Initialize $\mathbf{\Phi}_0^{\mathcal{X}}, \mathbf{\Phi}_0^{\mathcal{Y}}, \mathbf{A}_0^{\mathcal{X}}, \mathbf{A}_0^{\mathcal{Y}}, \mathbf{Q}_0^{\mathcal{X}}, \mathbf{Q}_0^{\mathcal{Y}}$.
2 Let $\mathbf{Q}_0^{\mathcal{X}} = \mathbf{I}$, $\mathbf{Q}_0^{\mathcal{Y}} = \mathbf{I}$, $\mathbf{P}_0^{\mathcal{X}} \leftarrow \mathbf{A}_0^{\mathcal{X}} \mathbf{Q}_0^{\mathcal{X}}$, $\mathbf{P}_0^{\mathcal{Y}} \leftarrow \mathbf{A}_0^{\mathcal{Y}} \mathbf{Q}_0^{\mathcal{Y}}$.
3 **while** *not converged* **do**
4   Fix other variables, update $\mathbf{A}_{i+1}^{\mathcal{X}}$ and $\mathbf{A}_{i+1}^{\mathcal{Y}}$ by sparse coding according to Eq. (12).
5   Fix other variables, update $\mathbf{\Phi}_{i+1}^{\mathcal{X}}$ and $\mathbf{\Phi}_{i+1}^{\mathcal{Y}}$ by dictionary learning according to Eq. (13).
6   Fix other variables, update $\mathbf{Q}_{i+1}^{\mathcal{X}}$ and $\mathbf{Q}_{i+1}^{\mathcal{Y}}$ according to Eq. (16) based on $\mathbf{A}_{i+1}^{\mathcal{X}}$, $\mathbf{A}_{i+1}^{\mathcal{Y}}$ and $\mathbf{\Phi}_{i+1}^{\mathcal{X}}$, $\mathbf{\Phi}_{i+1}^{\mathcal{Y}}$.
7   Update $\mathbf{P}_{i+1}^{\mathcal{X}} \leftarrow \mathbf{A}_{i+1}^{\mathcal{X}} \mathbf{Q}_{i+1}^{\mathcal{X}}$, $\mathbf{P}_{i+1}^{\mathcal{Y}} \leftarrow \mathbf{A}_{i+1}^{\mathcal{Y}} \mathbf{Q}_{i+1}^{\mathcal{Y}}$.
8 **end**

Output: $\mathbf{\Phi}^{\mathcal{X}}$, $\mathbf{\Phi}^{\mathcal{Y}}$ and $\mathbf{Q}^{\mathcal{X}}$, $\mathbf{Q}^{\mathcal{Y}}$.

Algorithm 2: Cross-Modality Image Synthesis

Input: Test image $\mathbf{X}^t$, dictionary pairs $\mathbf{\Phi}^{\mathcal{X}}$ and $\mathbf{\Phi}^{\mathcal{Y}}$, projection matrices $\mathbf{Q}^{\mathcal{X}}$ and $\mathbf{Q}^{\mathcal{Y}}$.

1 Initialize $\mathbf{A}_0^{t\mathcal{X}}$, $\mathbf{A}_0^{t\mathcal{Y}}$ by Eq. (17).
2 Let $\mathbf{A}_0^{t\mathcal{Y}} \leftarrow \mathbf{Q}^{\mathcal{Y}-1} \mathbf{Q}^{\mathcal{X}} \mathbf{A}_0^{t\mathcal{X}}$, $\mathbf{Y}_0^t \leftarrow \mathbf{A}_0^{t\mathcal{Y}} \mathbf{\Phi}_0^{\mathcal{Y}}$.
3 **while** *not converged* **do**
4   Solve $\mathbf{A}_{i+1}^{t\mathcal{X}}$, $\mathbf{A}_{i+1}^{t\mathcal{Y}}$ using Eq. (12) with $\mathbf{Q}^{\mathcal{X}}$, $\mathbf{Q}^{\mathcal{Y}}$ and $\mathbf{Y}_i^t$.
5   Update $\mathbf{Y}_{i+1}^t \leftarrow \mathbf{Q}^{\mathcal{Y}-1} \mathbf{Q}^{\mathcal{X}} \mathbf{A}_{i+1}^{t\mathcal{X}} \mathbf{\Phi}^{\mathcal{Y}} = \mathbf{A}_{i+1}^{t\mathcal{Y}} \mathbf{\Phi}^{\mathcal{Y}}$.
6 **end**

Output: Synthesized image $\mathbf{Y}^t$.

Algorithm 2 summarizes the process for cross-modality image synthesis.

IV. 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.

A. Databases and Pre-processing

We validate our method on two public multi-modality brain datasets, *viz.* IXI[2] and NAMIC[3] databases, respectively. The IXI database involves 578 healthy subjects each imaged using a matrix of $256 \times 256 \times v$ ($v = 112 \sim 136$) scanned with a Magnetic Resonance Imaging (MRI) system. The NAMIC database, instead, contains 20 subjects (ten are normal controls and the other ten are schizophrenic) each imaged using a matrix of $128 \times 128 \times z$ ($z = 88$) scanned with a 3T MRI system. For our experiments, we adopt PDw, T2w MRI scans from the IXI dataset, and T1w, T2w acquisitions form the NAMIC dataset. Following [4, 14, 15], 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

---

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

TABLE I
THE NUMBER OF SELECTED PAIRED/UNPAIRED IMAGES.

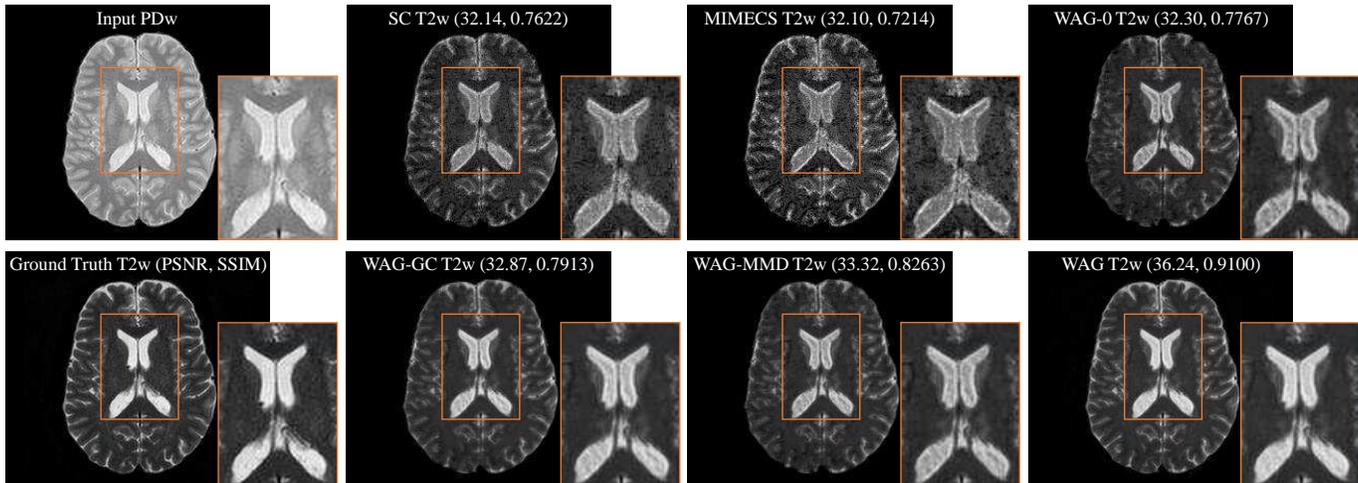|	IXI		NAMIC		RATIO
	PAIRED SETS	UNPAIRED SETS	PAIRED SETS	UNPAIRED SETS	PAIRED/FULL SET
Scenario #1	289	–	10	–	100%
Scenario #2	145	72	6	2	50.2%
Scenario #3	73	108	4	3	25.3%
Scenario #4	37	126	2	4	12.8%

Fig. 2. Synthesized results generated using SC, MIMECS, WAG-0, WAG-GC, WAG-MMD and WAG (zoom in for details).

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 I for two datasets separately. Specifically, Table I 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 PDw and T2w images.
- Unpaired sets: B = 72 subjects with PDw images.
- Unpaired sets: C = 72 subjects with T2w images.
- $A \cap B \cap C = \emptyset$

B. Experimental Setup

We evaluate our method in two scenarios. First, we use the IXI dataset for synthesizing the T2w images from the PDw acquisitions and *vice versa*. Second, we adopt the NAMIC dataset for generating the T1w scans from the T2w inputs and *vice versa*. In our experiments, we 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 [32, 35], 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 [19]. 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 IV-D). For the evaluation metrics, we adopt the widely used Peak Signal to Noise Ratio (PSNR) and Structural Similarity Index (SSIM) [43] to objectively assess the quality of the synthesized images.

C. 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:

- SC: Sparse Coding-based method [19]
- MIMECS: MRI example-based contrast synthesis [4]
- Ve-S: Vemulapalli's supervised [15]
- Ve-US: Vemulapalli's unsupervised [15]
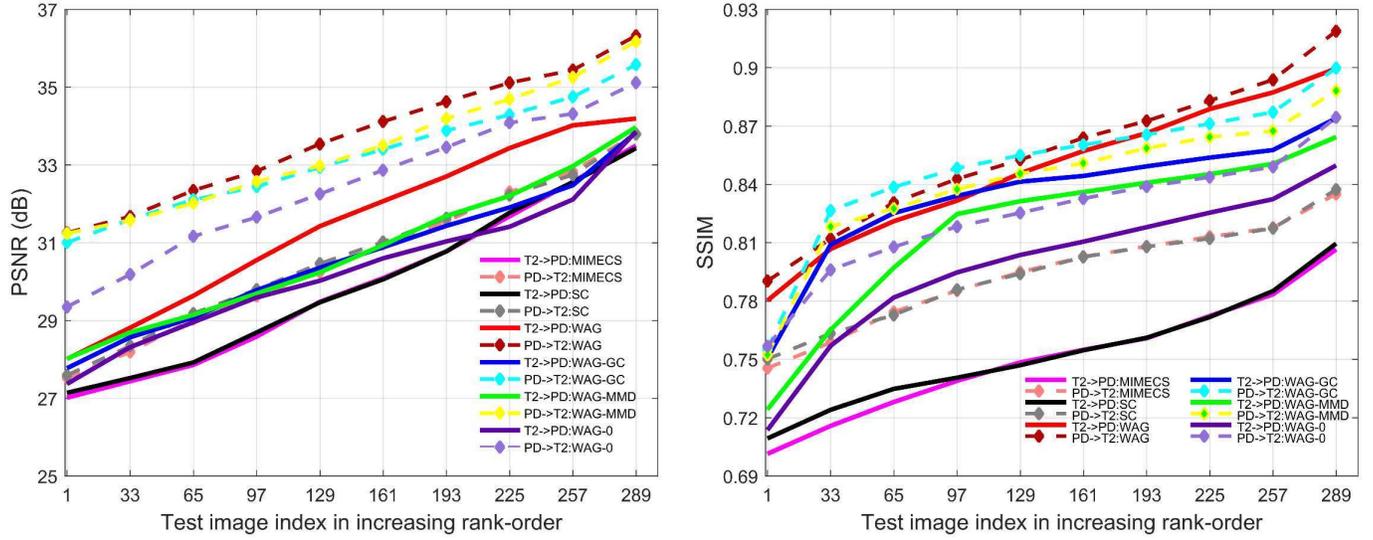- WAG-0: WAG without any regularization terms

Fig. 3. Cross-modality synthesis results: MIMECS, SC, WAG, WAG-MMD, WAG-GC and WAG-0 on the IXI dataset.

TABLE II
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 | |
	T2w $\mapsto$ PDw	PDw $\mapsto$ T2w	T2w $\mapsto$ PDw	PDw $\mapsto$ T2w	T2w $\mapsto$ PDw	PDw $\mapsto$ T2w	T2w $\mapsto$ PDw	PDw $\mapsto$ T2w
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 III
ERROR MEASURES OF THE WAG-SYNTHESIZED IMAGES RESULTING FROM DIFFERENT PAIRED/FULL SET RATIOS DURING DICTIONARY TRAINING.

| | Fixing the number of paired data as 145 | | | | | | | |
| Metric (mean) | no unpaired data | | 36 unpaired data | | 48 unpaired data | | 72 unpaired data | |
	T2w $\mapsto$ PDw	PDw $\mapsto$ T2w	T2w $\mapsto$ PDw	PDw $\mapsto$ T2w	T2w $\mapsto$ PDw	PDw $\mapsto$ T2w	T2w $\mapsto$ PDw	PDw $\mapsto$ T2w
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**

TABLE IV
ERROR MEASURES OF THE WAG-SYNTHESIZED IMAGES RESULTING FROM DIFFERENT PAIRED/FULL SET RATIOS DURING DICTIONARY TRAINING.

| | Fixing the number of unpaired data as 72 | | | | | |
| Metric (mean) | 37 paired data | | 73 paired data | | 145 paired data | |
	T2w $\mapsto$ PDw	PDw $\mapsto$ T2w	T2w $\mapsto$ PDw	PDw $\mapsto$ T2w	T2w $\mapsto$ PDw	PDw $\mapsto$ T2w
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**

- 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 III-D3 and III-D4, respectively.

D. Experimental Results

As we mentioned in Section IV-B, we first address cross-modality synthesis on the IXI dataset. In this scenario, we investigate both PDw and T2w 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 results in Fig. 2, while reporting all quantitative results in Fig. 3. From Fig. 3, 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 I. The averaged PSNRs and SSIMs are shown in Table II. Generally, a larger number of paired
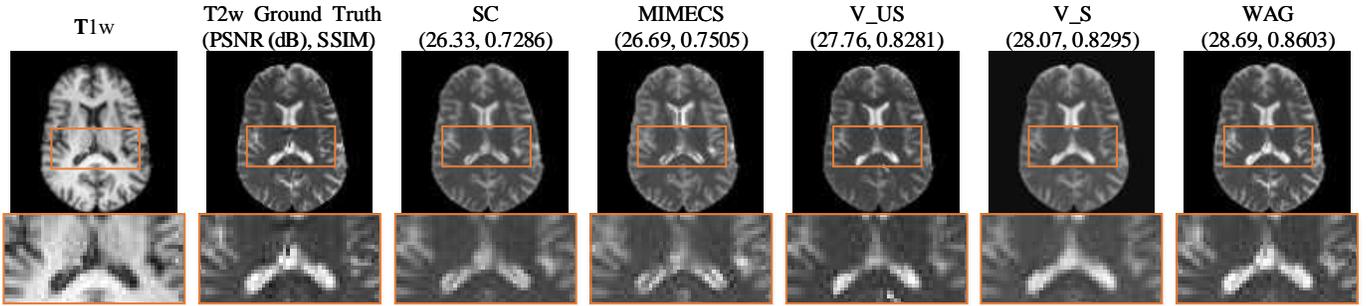
Fig. 4. Example cross-modality synthesis results generated by MIMECS, SC, Ve-S Ve-US and WAG on the NAMIC dataset.
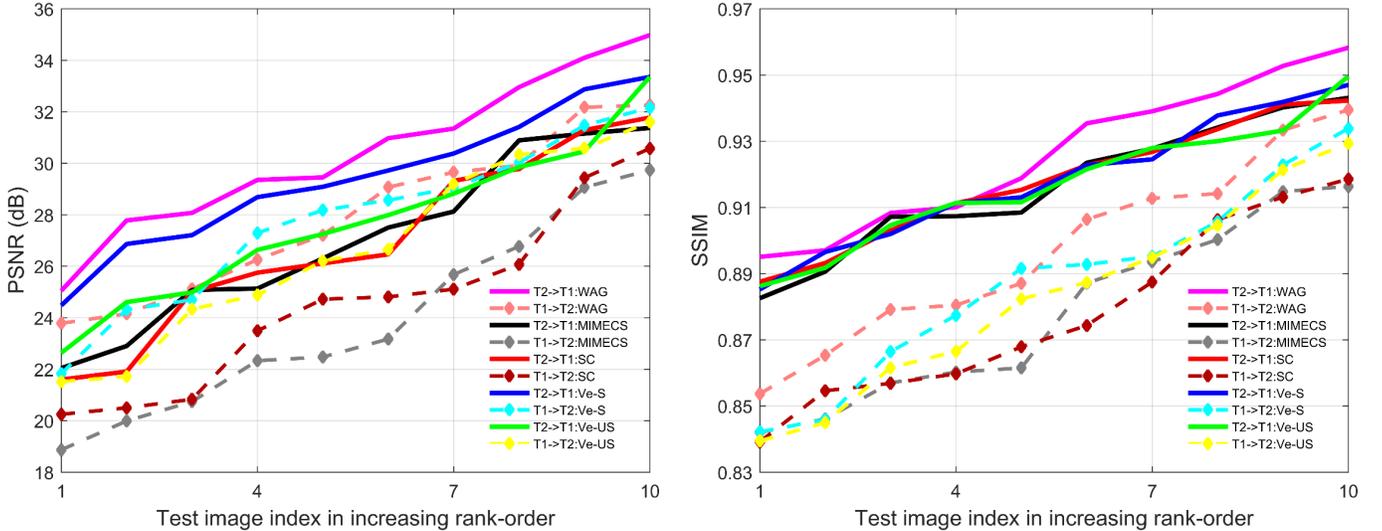

Fig. 5. Cross-modality synthesis results: MIMECS, SC, WAG, Ve-S and Ve-US on the NAMIC dataset.

TABLE V
AVERAGED PSNRs AND SSIMs OF THE SYNTHESIZED IMAGES USING DIFFERENT METHODS ON THE NAMIC DATASET.

	NAMIC Dataset									
Metric (mean)	T1w ↦ T2w					T2w ↦ T1w				
	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**

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 III and IV, 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. 4, Fig. 5 and summarize the averaged values in Table V, 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, about 30 minutes using a Matlab R2013a code. Execution time

TABLE VI
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: T2w ↦ PDw			
p-value (PSNR)	< 0.001	< 0.001	< 0.001
p-value (SSIM)	< 0.001	< 0.001	< 0.001
IXI: PDw ↦ T2w			
p-value (PSNR)	< 0.001	< 0.001	< 0.001
p-value (SSIM)	< 0.001	< 0.001	< 0.001

TABLE VII
INDEPENDENT t-TEST ON THE PERFORMANCE BENEFITS USING THE IXI DATASET.

Independent t-test	WAG vs. MIMECS	WAG vs. SC
IXI: T2w ↦ PDw		
p-value (PSNR)	< 0.001	< 0.001
p-value (SSIM)	< 0.001	< 0.001
IXI: PDw ↦ T2w		
p-value (PSNR)	< 0.001	< 0.001
p-value (SSIM)	< 0.001	< 0.001

TABLE VIII
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: T1w ↦ T2w				
p-value (PSNR)	0.0319	0.0308	0.0450	0.0363
p-value (SSIM)	0.0347	0.0396	0.0468	0.0392
NAMIC: T2w ↦ T1w				
p-value (PSNR)	0.0168	0.0361	**0.0809**	0.041
p-value (SSIM)	0.0143	0.0138	0.0464	0.0345

for the synthesis of one 3D representative image with size $256 \times 256 \times 100$ pixels took about 7 minutes.

E. 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 VI lists the results of paired t-test for case (1), which shows our improvements are all statistically significant. Tables VII and VIII 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 T1w images from T2w data on the NAMIC dataset using Ve-S method.

V. 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 II and Table V, 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 T2w or PDw images from the opposite modality. For the synthesis from T2w data to PDw 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 PDw data to T2w 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. 4, 5 and Table V, for generating T2w from T1w 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 T1w images from T2w 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 [44]. 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.

VI. CONCLUSION

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.

REFERENCES

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

[2] L. G. Nyúl, J. K. Udupa *et al.*, "On standardizing the mr image intensity scale," *image*, vol. 1081, 1999.

[3] A. Madabhushi and J. K. Udupa, "New methods of mr image intensity standardization via generalized scale," *Medical Physics*, vol. 33, no. 9, pp. 3426–3434, 2006.

[4] S. Roy, A. Carass, and J. L. Prince, "Magnetic resonance image example-based contrast synthesis," *IEEE Trans. Med. Imaging*, vol. 32, no. 12, pp. 2348–2363, 2013.

[5] 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 *International Workshop on Simulation and Synthesis in Medical Imaging*. Springer, 2016, pp. 118–126.

[6] T. Cao, C. Zach, S. Modla, D. Powell, K. Czymmek, and M. Niethammer, "Multi-modal registration for correlative microscopy using image analogies," *Med. Image Anal.*, vol. 18, no. 6, pp. 914–926, 2014.

[7] J. Woo, M. Stone, and J. L. Prince, "Multimodal registration via mutual information incorporating geometric and spatial context," *IEEE Trans. Image Process.*, vol. 24, no. 2, pp. 757–769, 2015.

[8] J. E. Iglesias and M. R. Sabuncu, "Multi-atlas segmentation of biomedical images: a survey," *Med. Image Anal.*, vol. 24, no. 1, pp. 205–219, 2015.

[9] N. Cordier, H. Delingette, M. Lê, and N. Ayache, "Extended modality propagation: Image synthesis of pathological cases," *IEEE Trans. Med. Imaging*, vol. 35, no. 12, pp. 2598–2608, 2016.

[10] S. K. W. Olivier Commowick and G. Malandain, "Using Frankensteins creature paradigm to build a patient specific atlas," in *MICCAI*. Springer, 2009, pp. 993–1000.

[11] A. Hertzmann, C. E. Jacobs, N. Oliver, B. Curless, and D. H. Salesin, "Image analogies," in *ACM SIGGRAPH*, 2001, pp. 327–340.

[12] Z. Liu, Z. Zhang, and Y. Shan, "Image-based surface detail transfer," *IEEE Comput. Graphics Appl.*, vol. 24, no. 3, pp. 30–35, 2004.

[13] A. Jog, A. Carass, S. Roy, D. L. Pham, and J. L. Prince, "Random forest regression for magnetic resonance image synthesis," *Med. Image Anal.*, vol. 35, pp. 475–488, 2017.

[14] D. H. Ye, D. Zikic, B. Glocker, A. Criminisi, and E. Konukoglu, "Modality propagation: coherent synthesis of subject-specific scans with data-driven regularization," in *MICCAI*. Springer, 2013, pp. 606–613.

[15] R. Vemulapalli, H. Van Nguyen, and S. Kevin Zhou, "Unsupervised cross-modal synthesis of subject-specific scans," in *IEEE ICCV*, 2015, pp. 630–638.

[16] N. Burgos, M. J. Cardoso, K. Thielemans, M. Modat, S. Pedemonte, J. Dickson, A. Barnes, R. Ahmed, C. J. Mahoney, J. M. Schott *et al.*, "Attenuation correction synthesis for hybrid pet-mr scanners: application to brain studies," *IEEE Trans. Med. Imaging*, vol. 33, no. 12, pp. 2332–2341, 2014.

[17] H. Van Nguyen, K. Zhou, and R. Vemulapalli, "Cross-domain synthesis of medical images using efficient location-sensitive deep network," in *MICCAI*. Springer, 2015, pp. 677–684.

[18] D.-J. Kroon and C. H. Slump, "MRI modalitiy transfor-

mation in demon registration," in *IEEE ISBI*. IEEE, 2009, pp. 963–966.

[19] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE Trans. Image Process.*, vol. 19, no. 11, pp. 2861–2873, 2010.

[20] 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 *IEEE CVPR*, July 2017.

[21] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4311–4322, 2006.

[22] G. Davis, S. Mallat, and M. Avellaneda, "Adaptive greedy approximations," *Constructive approximation*, vol. 13, no. 1, pp. 57–98, 1997.

[23] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM Rev.*, vol. 43, no. 1, pp. 129–159, 2001.

[24] H. Lee, A. Battle, R. Raina, and A. Y. Ng, "Efficient sparse coding algorithms," in *NIPS*, 2006, pp. 801–808.

[25] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. R. Stat. Soc. Series B Stat. Methodol.*, pp. 267–288, 1996.

[26] J. Mairal, M. Elad, and G. Sapiro, "Sparse representation for color image restoration," *IEEE Trans. Image Process.*, vol. 17, no. 1, pp. 53–69, 2008.

[27] H. Chang, D.-Y. Yeung, and Y. Xiong, "Super-resolution through neighbor embedding," in *IEEE CVPR*.

[28] Y. Huang, F. Zhu, L. Shao, and A. F. Frangi, "Color object recognition via cross-domain learning on RGB-D images," in *IEEE ICRA*, 2016, pp. 1672–1677.

[29] F. Zheng, Y. Tang, and L. Shao, "Hetero-manifold regularisation for cross-modal hashing," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2016.

[30] F. Zhu and L. Shao, "Weakly-supervised cross-domain dictionary learning for visual recognition," *Int. J. Comput. Vision*, vol. 109, no. 1-2, pp. 42–59, 2014.

[31] S. Wang, L. Zhang, Y. Liang, and Q. Pan, "Semi-coupled dictionary learning with applications to image super-resolution and photo-sketch synthesis," in *IEEE CVPR*, 2012, pp. 2216–2223.

[32] D.-A. Huang and Y.-C. Frank Wang, "Coupled dictionary and feature space learning with applications to cross-domain image synthesis and recognition," in *IEEE ICCV*, 2013, pp. 2496–2503.

[33] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *J. Mach. Learn. Res.*, vol. 13, no. Mar, pp. 723–773, 2012.

[34] K. M. Borgwardt, A. Gretton, M. J. Rasch, H.-P. Kriegel, B. Schölkopf, and A. J. Smola, "Integrating structured biological data by kernel maximum mean discrepancy," *Bioinformatics*, vol. 22, no. 14, pp. e49–e57, 2006.

[35] M. Long, J. Wang, G. Ding, D. Shen, and Q. Yang, "Transfer learning with graph co-regularization," *IEEE Trans. Knowl. Data Eng.*, vol. 26.

[36] I. Steinwart, "On the influence of the kernel on the consistency of support vector machines," *J. Mach. Learn. Res.*, vol. 2, no. Nov, pp. 67–93, 2001.

[37] X. Lu, H. Yuan, P. Yan, Y. Yuan, and X. Li, "Geometry constrained sparse coding for single image super-resolution," in *IEEE CVPR*, 2012, pp. 1648–1655.

[38] M. Zheng, J. Bu, C. Chen, C. Wang, L. Zhang, G. Qiu, and D. Cai, "Graph regularized sparse coding for image representation," *IEEE Trans. Image Process.*, vol. 20, no. 5, pp. 1327–1336, 2011.

[39] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," in *NIPS*, vol. 14, no. 14, 2001, pp. 585–591.

[40] M. Long, G. Ding, J. Wang, J. Sun, Y. Guo, and P. S. Yu, "Transfer sparse coding for robust image representation," in *IEEE CVPR*, 2013, pp. 407–414.

[41] H. Lee, A. Battle, R. Raina, and A. Y. Ng, "Efficient sparse coding algorithms," *NIPS*, vol. 19, p. 801, 2007.

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

[43] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, 2004.

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