Unsupervised Cross-modal Synthesis of Subject-specific Scans

Raviteja Vemulapalli
Center for Automation Research, UMIACS
University of Maryland, College Park
raviteja@umd.edu

Hien Van Nguyen, Shaohua Kevin Zhou
Siemens Healthcare Technology Center
Princeton, New Jersey
{hien.nguyen, shaohua.zhou}@siemens.com

Abstract

Recently, cross-modal synthesis of subject-specific scans has been receiving significant attention from the medical imaging community. Though various synthesis approaches have been introduced in the recent past, most of them are either tailored to a specific application or proposed for the supervised setting, i.e., they assume the availability of training data from the same set of subjects in both source and target modalities. But, collecting multiple scans from each subject is undesirable. Hence, to address this issue, we propose a general unsupervised cross-modal medical image synthesis approach that works without paired training data. Given a source modality image of a subject, we first generate multiple target modality candidate values for each voxel independently using cross-modal nearest neighbor search. Then, we select the best candidate values jointly for all the voxels by simultaneously maximizing a global mutual information cost function and a local spatial consistency cost function. Finally, we use coupled sparse representation for further refinement of synthesized images. Our experiments on generating T1-MRI brain scans from T2-MRI and vice versa demonstrate that the synthesis capability of the proposed unsupervised approach is comparable to various state-of-the-art supervised approaches in the literature.

1. Introduction

Currently, a multitude of imaging modalities such as Computed Tomography, T1-weighted and T2-weighted Magnetic Resonance Imaging (MRI), X-ray, Ultrasound, Diffusion MRI, etc., are being used for medical imaging research. Each of these modalities captures different characteristics of the underlying anatomy, and the relationship between any two modalities is highly nonlinear.

Due to variations in the image characteristics across modalities, medical image analysis algorithms trained with data from one modality may not work well when applied to data from a different modality. A straightforward way to address this issue is to collect a large amount of training data in each modality. But, this is impractical since collecting medical images is both time consuming and expensive. Hence, it is highly desirable to have a general cross-modal synthesis approach that generates subject-specific scans in the desired target modality from the given source modality images. The ability to synthesize medical images without real acquisitions has many potential applications like super-resolution [13, 26, 32], building virtual models [20], atlas construction [7], multimodal registration [5, 23, 24, 28], segmentation [12, 25] and virtual enhancement [18].

Though various cross-modal medical image synthesis approaches have been proposed in the recent past, most of them work under supervised setting, requiring training data in both source and target modalities from the same set of subjects. For example, a coupled sparse representation-based image synthesis approach was proposed in [5], which used paired training data to learn coupled dictionaries in source and target modalities. Various regression-based synthesis approaches were proposed in [14, 17], which used paired training data to train either regression forests [14] or deep networks [17]. Recently, various supervised approaches [25, 31] based on image analogies [10] and label propagation [22] have also been proposed.

Though some recent supervised approaches have shown promising synthesis capabilities, the availability of paired data from source and target modalities is limited in many cases. Also, collecting multiple scans from each subject is not desirable. Hence, to address these issues, we introduce an unsupervised cross-modal image synthesis approach that can generate subject-specific target modality scans without using paired training data. Given a source modality image from a subject, the proposed approach generates the corresponding target modality image by using the target modality images from a different set of subjects. Since the proposed approach does not use any supervision, i.e., paired training data, for learning a mapping between the modalities, we call it *unsupervised*.

Since synthesizing a full medical image (that may have millions of voxels) is a fairly complex task, we propose a novel two-step synthesis approach in this paper. In the first

step, we generate multiple target modality candidate intensity values independently for each voxel using patch-based cross-modal nearest neighbor search. In the second step, we synthesize a full target modality image by selecting the best candidate values jointly for all the voxels based on the following two criteria:

- **Global mutual information maximization:** Since we are interested in generating subject-specific scans, the synthesized target modality image should have high mutual information with the given source modality image.
- **Local spatial consistency maximization:** The best candidate selected for each voxel should be spatially consistent with the best candidates selected for its neighbors.

Using these two criteria, we formulate the image synthesis (or the candidate selection) step as an optimization problem and solve it using reduced gradient ascent approach. Finally, after synthesizing a full target modality image, we refine it further using coupled sparse representation.

The proposed unsupervised approach shows promising synthesis capabilities when evaluated using the NAMIC brain multimodality database by generating T1-MRI scans from T2-MRI scans and vice versa. Though our main focus is on unsupervised image synthesis, the proposed approach can also be used in the supervised setting by replacing the cross-modal nearest neighbor search with source-modal nearest neighbor search. In fact, when evaluated under the supervised setting using the NAMIC brain database, the proposed synthesis approach outperforms state-of-the-art methods including modality-propagation [31] and location-sensitive deep network [17].

Contributions:

- **Unsupervised image synthesis:** We propose a general unsupervised approach for cross-modal synthesis of subject-specific scans. The proposed approach does not require paired training data from the source and target modalities. To the best of our knowledge, this is the first approach that addresses the cross-modal medical image synthesis problem in an unsupervised setting.
- **Optimization-based synthesis:** We formulate the cross-modal image synthesis task as an optimization problem that jointly maximizes the mutual information between the given source modality and the synthesized target modality images, and the spatial consistency among neighboring voxels in the synthesized target image.
- **Supervised extension:** We show that the proposed approach can also be used in the supervised setting by replacing the cross-modal nearest neighbor search with source-modal nearest neighbor search. In fact, under the supervised setting, the proposed approach clearly outperforms state-of-the-art modality propagation [31] and location-sensitive deep network [17] appraoches.

Organization: Section 2 provides a brief overview of the existing cross-modal medical image synthesis approaches and section 3 describes the proposed synthesis approach. Section 4 presents the experimental results and section 5 concludes the paper.

2. Relevant Work

A model-based approach for generating ultrasound scans from CT scans was proposed in [28]. In [9], an approach was proposed for explicitly estimating the intrinsic tissue parameters (that cause MR contrast) using a set of MRI scans. The estimated parameters were later used to synthesize arbitrary MR contrast. Both these approaches rely on the underlying physics of the data acquisition process and cannot be applied to other modalities.

A simple modality transformation between T1-MRI and T2-MRI was proposed in [16] for image registration. This approach simply used the peaks in joint intensity histogram of the two images being registered, resulting in a coarse modality transformation. In [21], a model-based approach was proposed to generate a high resolution brain MRI image from its low resolution version, by using a high resolution image of the same brain with a different MR contrast.

In [15], a set of paired low/high resolution images was used to learn the joint probability distribution of low/high resolution patches. This probability distribution was then used for synthesizing a high resolution image from the given low resolution image. A supervised regression-based synthesis approach was proposed in [2, 14]. This approach used paired data from source and target modalities to train a random forest, which was later used for regressing the target modality patches from the source modality patches.

In [5], a supervised synthesis approach was proposed based on coupled sparse representation. While training, this approach used paired data from source and target modalities to learn coupled dictionaries that establish cross-modal correspondences. These learned dictionaries were then used for sparse coding-based image synthesis. Similar sparse coding-based approaches have also been used in [4, 24, 26, 27] for image synthesis applications.

In [25], a supervised approach was proposed for MR contrast synthesis using a codebook of paired training patches. For each test patch, few best matches were found from the codebook and the target patches corresponding to best matches were averaged to generate target MR contrast. Similar approaches were also used in [6, 12]. An iterative synthesis approach, called modality propagation, was proposed in [31] using the label propagation framework [22]. Recently, an efficient location-sensitive deep network-based approach was proposed in [17], which explicitly used the voxel image coordinates in the synthesis process.

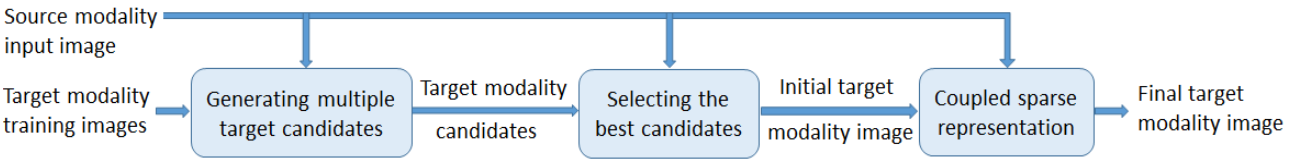

Figure 1: The proposed unsupervised cross-modal image synthesis approach.

3. Proposed Synthesis Approach

Given a source modality image from a subject, the proposed unsupervised approach synthesizes the corresponding target modality image by making use of target modality images from a different set of subjects.

Synthesizing an image with N voxels can be considered as estimating an N-dimensional quantity. The following are two possible (extreme) approaches for solving this problem: (i) Estimating all the voxels jointly, (ii) Estimating each voxel independently. Each approach has its own advantages and disadvantages. While the first approach takes interactions between voxels into account, it is fairly complex given the large set of possible values for each voxel. While the second approach simplifies the problem by considering each voxel independently, it does not take into account the image level context.

To take advantage of both these approaches, we follow a two-step synthesis strategy in this paper. In the first step, we generate multiple target modality candidate values for each voxel independently. In the second step, we synthesize a full target modality image by selecting the best candidate values jointly for all the voxels by taking the image level context into account. The rationale behind generating multiple candidates in the first step is that at least one of the top K candidates would be an appropriate value for synthesis when the image context is considered in the second step. This can also be interpreted as restricting the set of possible intensity values for each voxel so that the joint estimation step becomes more tractable.

Since we are working without paired training data, the quality of our synthesized images would be usually low when compared to the supervised approaches. Hence, to further improve the synthesis results, we use coupled sparse representation (CSR) as a refinement step. Our choice of CSR is motivated by its recent success in supervised image synthesis applications [4, 5, 11, 24]. Figure 1 shows the block diagram of the overall synthesis process.

Neighborhood: Let Φ^v denote the set consisting of voxel v and its neighbors. In this paper, we use all the voxels that are at unit distance from v as its neighbors. This would be six neighbors in the case of 3D volumes. If required, one can add more neighbors to the set Φ^v without any changes to the proposed approach.

Notations: We use the notation $\Phi^v(p,q,r)$ to represent the elements of Φ^v. Here, $\Phi^v(p,q,r)$ refers to the voxel $(v+(p,q,r))$. We denote the ℓ_0 and ℓ_2 norms using $\|\ \|_0$ and $\|\ \|_2$, respectively. The i^{th} column of a matrix A is denoted using $A(:,i)$. We use the notation $v \backsim v'$ to indicate that voxels v and v' are neighbors. We use $\mathbb{I}$ to denote the indicator function and P to denote probability.

3.1. Generating multiple target modality candidates

In the first step, given a source modality image I_s, we generate multiple target modality candidate intensity values for the set Φ^v at each voxel independently. To generate the target values for Φ^v, we make use of a $d_1 \times d_1 \times d_1$ patch centered on v extracted from the given source image I_s. If we had paired $Source-Target$ images during training, we could have possibly learned a predictor/regressor

$$f : (\text{Source modality patch at voxel } v) \longrightarrow (\text{Multiple target modality candidate values for } \Phi^v).$$

But, since we do not have such paired training data in the unsupervised setting, we obtain the target modality candidates using cross-modal nearest neighbor search. For each $d_1 \times d_1 \times d_1$ patch from the given source image I_s, we obtain K nearest $d_1 \times d_1 \times d_1$ target patches by searching across the target modality training images. We use the intensity values of the center voxel and its neighbors from these K nearest patches as target candidate values for the set Φ^v.

For cross-modal nearest neighbor search, we require a similarity measure that is robust to changes in modality. In this paper, we use voxel intensity-based mutual information, which has been successfully used in the past as a cross-modal similarity measure for medical image registration [19]. Given two image patches A and B, their mutual information is given by

$$MI(A,B) = H(X_a) + H(X_b) - H(X_a, X_b), \quad (1)$$

where H denotes the Shannon entropy function, X_a and X_b are random variables representing the voxel intensities in patches A and B, respectively.

Note that instead of mutual information, one can use any other cross-modal similarity measure that is appropriate for the modalities under consideration.

3.2. Full image synthesis using best candidates

In the second step, given K target modality candidate intensity values for the set Φ^v at each voxel, we synthesize a full target modality image $\tilde{I}_t$ by selecting one among the

K candidates at each voxel. We use the value of $\Phi^v(0,0,0)$ from the selected candidate to synthesize voxel v in $\tilde{I}_t$.

Let X_s and X_t be two random variables with support $\Psi = \{l_1, \ldots, l_L\}$, representing the voxel intensity values of images I_s and $\tilde{I}_t$, respectively. Let $I_s(v)$ denote the intensity value of voxel v in image I_s. Let V represent the set of all voxels with cardinality N. Let $\{\phi^{v1}, \ldots, \phi^{vK}\}$ denote the K target modality candidate values for the set Φ^v at voxel v. Let $w_{vk} = \mathbb{I}[\text{Candidate } \phi^{vk} \text{ is selected at voxel } v]$.

Since the candidates have been obtained for each voxel independently (using nearest neighbor search), we propose to solve the selection problem jointly for all the voxels based on the following two criteria: (i) Mutual information maximization, which is a global criterion, and (ii) Spatial consistency maximization, which is a local criterion.

3.2.1 Global mutual information maximization

Motivated by the assumption that regions of similar tissues (and hence similar grey values) in one modality image would correspond to regions of similar grey values in the other modality image (though the values could be different across modalities), mutual information has been successfully used in the past as a cost function for cross-modal medical image registration [19]. Motivated by this, in this paper, we use mutual information as a cost function for cross-modal medical image synthesis. Since we are interested in generating subject-specific scans, the synthesized target modality image $\tilde{I}_t$ should have high mutual information with the given source modality image I_s, i.e., the amount of information I_s and $\tilde{I}_t$ contain about each other should be maximal. This global criterion helps in transferring the image level structure across modalities.

The mutual information between images I_s and $\tilde{I}_t$ is given by $MI(X_s, X_t) = H(X_s) + H(X_t) - H(X_s, X_t)$. Since the entropy $H(X_s)$ is constant for a given source modality image, maximizing the mutual information is equivalent to maximizing $H(X_t) - H(X_s, X_t)$, where

$$
\begin{aligned}
H(X_t) &= -\sum_{b=1}^{L} P(X_t = l_b) \log[P(X_t = l_b)], \\
P(X_t = l_b) &= \frac{1}{N} \sum_{v \in V} \sum_{k=1}^{K} w_{vk} \mathbb{I}[\phi^{vk}(0,0,0) = l_b], \\
H(X_s, X_t) &= -\sum_{a,b=1}^{L} P_{ab} \log[P_{ab}], \qquad (2) \\
P_{ab} &= P(X_s = l_a, X_t = l_b) \\
&= \frac{1}{N} \sum_{v \in V} \sum_{k=1}^{K} w_{vk} \mathbb{I}[I_s(v) = l_a, \phi^{vk}(0,0,0) = l_b].
\end{aligned}
$$

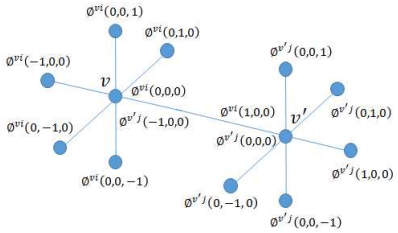

Figure 2: The values assigned by ϕ^{vi} to the set Φ^v and the values assigned by $\phi^{v'j}$ to the set $\Phi^{v'}$.

3.2.2 Local spatial consistency maximization

Let $v, v' \in V$ be two neighboring voxels. Note that if we select a candidate ϕ^{vi} at voxel v, along with assigning the value $\phi^{vi}(0,0,0)$ to voxel v, it also assigns the value $\phi^{vi}(v' - v)$ to the neighboring voxel v'. Similarly if we select a candidate $\phi^{v'j}$ at voxel v', along with assigning the value $\phi^{v'j}(0,0,0)$ to voxel v', it also assigns the value $\phi^{v'j}(v - v')$ to the neighboring voxel v. Figure 2 shows this pictorially. In this case, we would ideally want the selected candidates ϕ^{vi} and $\phi^{v'j}$ to be spatially consistent, i.e.,

$$
\phi^{vi}(0,0,0) = \phi^{v'j}(v - v'), \;\; \phi^{v'j}(0,0,0) = \phi^{vi}(v' - v). \qquad (3)
$$

Hence, in order to promote spatial consistency among the selected candidates, we maximize the following cost function (note the minus sign in the cost):

$$
SC(W) = -\sum_{\substack{v,v' \in V \\ v \backsim v'}} \begin{bmatrix} w_{v1} \ldots w_{vK} \end{bmatrix} \begin{bmatrix} C_{11}^{vv'} \ldots C_{1K}^{vv'} \\ \vdots \;\; \ddots \;\; \vdots \\ C_{K1}^{vv'} \ldots C_{KK}^{vv'} \end{bmatrix} \begin{bmatrix} w_{v'1} \\ \vdots \\ w_{v'K} \end{bmatrix},
$$

$$
\text{where } C_{ij}^{vv'} = \sqrt{\begin{aligned} &\Big(\phi^{vi}(0,0,0) - \phi^{v'j}(v - v')\Big)^2 \\ &\quad + \Big(\phi^{v'j}(0,0,0) - \phi^{vi}(v' - v)\Big)^2 \end{aligned}}. \qquad (4)
$$

Note that here $C_{ij}^{vv'}$ is the spatial consistency cost between neighbors v and v' when ϕ^{vi} is selected at v and $\phi^{v'j}$ is selected at v'.

3.2.3 Combined formulation

Combining the global mutual information cost and the local spatial consistency cost, we formulate the candidate selection step as the following optimization problem:

$$
\begin{aligned}
&\underset{\{w_{vk}\}}{\text{maximize}} \;\; H(X_t) - H(X_s, X_t) + \lambda SC(W) \\
&\text{subject to} \;\; \sum_{k=1}^{K} w_{vk} = 1, \;\; \forall v \in V, \qquad (5) \\
&\qquad\qquad w_{vk} \in \{0,1\}, \text{ for } k = 1, \ldots, K, \forall v \in V,
\end{aligned}
$$

where λ is a trade-off parameter.

The optimization problem (5) is combinatorial is nature due to the binary integer constraints on w_{vk} and is difficult to solve. Hence, we relax the binary integer constraints to positivity constraints to get the following relaxed problem:

$$\begin{aligned} \underset{\{w_{vk}\}}{\text{maximize}} \;\; & H(X_t) - H(X_s, X_t) + \lambda SC(W) \\ \text{subject to} \;\; & \sum_{k=1}^{K} w_{vk} = 1, \;\; \forall v \in V, \\ & w_{vk} \geq 0, \text{ for } \; k = 1, \ldots, K, \forall v \in V. \end{aligned} \tag{6}$$

3.2.4 Optimization

The cost function $H(X_t) - H(X_s, X_t) + \lambda SC(W)$ is differentiable and its derivative with respect to w_{vk} can be computed using:

$$\begin{aligned} \frac{dH(X_t)}{dw_{vk}} &= -\sum_{b=1}^{L} \left(1 + \log[P(X_t = l_b)]\right) \frac{d}{dw_{vk}} P(X_t = l_b) \\ &= -\frac{1}{N} \sum_{b=1}^{L} \left(1 + \log[P(X_t = l_b)]\right) \mathbb{I}[\phi^{vk}(0,0,0) = l_b] \\ &= -\frac{1}{N} \left(1 + \log[P(X_t = \phi^{vk}(0,0,0))]\right), \\ \frac{dH(X_s, X_t)}{dw_{vk}} &= -\sum_{a,b=1}^{L} \left(1 + \log[P_{ab}]\right) \frac{dP_{ab}}{dw_{vk}} \\ &= -\frac{1}{N} \sum_{a,b=1}^{L} \left(1 + \log[P_{ab}]\right) \mathbb{I}[I_s(v) = l_a, \phi^{vk}(0,0,0) = l_b] \\ &= -\frac{1}{N} \left(1 + \log[P(X_s = I_s(v), X_t = \phi^{vk}(0,0,0))]\right), \\ \frac{dSC(W)}{dw_{vk}} &= \sum_{v' \backsim v} \left(\sum_{p=1}^{K} C_{kp}^{vv'} w_{v'p} \right). \end{aligned} \tag{7}$$

The optimization problem (6) has a differentiable cost function with linear equality and inequality constraints. Hence, we solve it using reduced gradient ascent approach, in which the gradient computed from (7) is projected onto the constraint set in each iteration. Once we obtain w_{vk}, we use $\phi^{vk^*}(0,0,0)$ to synthesize voxel v in $\tilde{I}_t$, where $k^* = \underset{k}{\text{argmax}}\, w_{vk}$.

Note that, though the optimization problem (6) is a relaxation of problem (5), the unit ℓ_1-norm constraints on the weights promote a sparse solution [3, 8] pushing most of w_{vk} towards zero. Since the cost function in (6) is non-convex, the reduced gradient ascent approach is not guaranteed to find the global optimum. In our experiments, we use the local optimum obtained by initializing all the variables w_{vk} with a value of $\frac{1}{K}$. This initialization can be interpreted as giving equal weight to all the K candidates at the beginning of the optimization. During the optimization, in each iteration, along with projecting the gradient on to the constraint set, we also adjust the ascent direction such that the variables satisfying $w_{vk} = 0$ remain as zero. In each iteration, the learning rate is chosen as the maximum possible value such that none of the variables w_{vk} goes below zero.

3.3. Refinement with coupled sparse representation

Recently, coupled sparse representation has been shown to be a powerful model when dealing with coupled signal spaces in applications like cross-modal image synthesis [4, 5, 11, 24], super-resolution [26, 29, 30], etc. Sparse representations are robust to noise and artifacts present in the data. Hence, we use coupled sparse representation to refine the synthesized target modality image $\tilde{I}_t$ and generate the final target modality image I_t.

At each voxel $v \in V$, we extract small $d_2 \times d_2 \times d_2$ patches from the given source modality image I_s and the synthesized target modality image $\tilde{I}_t$. Let Z_v^s and Z_v^t denote the patches at voxel v from images I_s and $\tilde{I}_t$, respectively. Using $\{(Z_v^s, Z_v^t) \mid v \in V\}$ as signal pairs from the source and target modalities, coupled sparse representation can be formulated as the following optimization problem:

$$\begin{aligned} \underset{D_s, D_t, \{\alpha_v\}}{\text{minimize}} \;\; & \sum_{v \in V} \left(\|Z_v^s - D_s \alpha_v\|_2^2 + \|Z_v^t - D_t \alpha_v\|_2^2 \right) \\ \text{subject to} \;\; & \|\alpha_v\|_0 \leq T_0 \;\; \forall v \in V, \\ & \|D_s(:,j)\|_2 = 1, \;\; \|D_t(:,j)\|_2 = 1 \;\; \forall j, \end{aligned} \tag{8}$$

where D_s and D_t are over-complete dictionaries with M atoms in the source and target modalities respectively, α_v is the coupled sparse code for signals Z_v^s and Z_v^t in their respective dictionaries, and T_0 is the sparsity parameter.

We jointly learn the dictionaries D_s, D_t and the coupled sparse codes α_v by solving the optimization problem (8) using the K-SVD [1] algorithm with explicitly re-normalizing the columns of D_s and D_t to unit norm after each iteration. Once we have obtained the dictionaries and sparse codes, we reconstruct the target modality patches at every voxel using $\hat{Z}_v^t = D_t \alpha_v$, and use the center voxel value from $\hat{Z}_v^t$ to synthesize voxel v in the final target modality image I_t.

3.4. Extension to supervised setting

To extend the proposed unsupervised synthesis approach to the supervised setting, we simply replace the cross-modal nearest neighbor search in the candidate generation step with source-modal nearest neighbor search. For each voxel $v \in V$, we extract a $d_3 \times d_3 \times d_3$ patch centered on v from the given source modality test image I_s and find K nearest $d_3 \times d_3 \times d_3$ patches from source modality training images using standard Euclidean distance. Note that we need to perform a nearest neighbor search (even within source

modality) because the training and test images are from different subjects. Once we find the K nearest source modality training patches, we use the corresponding target modality training patches for generating the target modality candidates for Φ^v.

For the CSR step, we first use the paired training data to learn coupled dictionaries, and then use the learned dictionaries for refining the synthesized target modality images.

4. Experimental Evaluation

In this section, we evaluate the proposed image synthesis approach by generating T1-MRI scans from T2-MRI scans and vice versa. We use the same dataset, evaluation setting and evaluation metrics that were recently used in [17, 31]. This allows us to compare our results with state-of-the-art supervised synthesis approaches proposed in [17, 31]. To the best of our knowledge the proposed approach is the first unsupervised cross-modal synthesis approach, and hence there is no existing state-of-the-art to compare with under the unsupervised setting.

Dataset: For our experiments, we used T1 and T2 MRI scans of 19 subjects from the NAMIC brain multimodality database [1]. Along with the MRI scans, this database also provides brain masks for skull-stripping.

Data pre-processing: Following [31], all the images were skull stripped, linearly registered, inhomogeneity corrected, histogram matched within each modality, and resampled to 2 mm resolution. For registration, we used the first subject as reference. First, the T2 scan of the reference subject was registered to the corresponding T1 scan, and then the T1 and T2 scans of the remaining subjects were registered to the reference subject. We used 3D Slicer [2] software for data pre-processing.

Evaluation setting: We use leave-one-out cross-validation setting, in which the target modality image of a subject is synthesized using his/her source modality image and the images (target modality in the case of unsupervised setting; both source and target modalities in the case of supervised setting) of the remaining 18 subjects.

Evaluation metric: Since we have both T1 and T2 MRI scans for each subject in this dataset, we can directly compare the synthesized and ground truth target modality scans for evaluation. Since our focus in this work is on image synthesis, we use normalized cross correlation which was used in [31], and signal to noise ratio (SNR) which was used in [17], as evaluation metrics. Using the synthesized images for improving the performance of image analysis tasks like detection, segmentation, tracking, etc., will be considered in our future work.

[1] http://hdl.handle.net/1926/1687
[2] www.slicer.org

Implementation details: Since exhaustively searching the images (to find nearest neighbors) is highly computational, we restricted the search in each image to a $h \times h \times h$ (with $h = 7$) region around the voxel of interest. The patch sizes d_1 and d_3 used for cross-modal and source-modal nearest neighbor searches were chosen as 9 and 3, respectively. The patch size used for cross-modal search is much larger than the patch size used for source-modal search because for reliable estimation of mutual information, the patch should have sufficient number of voxels. The number of nearest neighbors K was chosen as 10. Since MRI scans have a high dynamic range, the mutual information computed using the original intensity values would be highly unreliable. Hence, we quantized the intensity values to $L = 32$ levels for computing the mutual information.

Note that the spatial consistency cost (4) is the sum of errors over all pairs of neighboring voxels. As the number of pairs in an image is very large, the actual value of (4) will be much higher than the mutual information cost. Hence, we chose the value of parameter λ in (6) such that the mutual information and spatial consistency costs have values that are of the same order of magnitude. For the unsupervised setting, we used $\lambda = 10^{-8}$ and for the supervised setting we used $\lambda = 10^{-7}$. For the CSR step, we used patches with $d_2 = 3$. The sparsity parameter T_0 and the number of dictionary atoms M were chosen as 5 and 500, respectively.

4.1. Synthesis results

Table 1 shows the correlation and SNR values of the images synthesized using the proposed unsupervised approach. Figure 3 shows some visual examples comparing the unsupervised synthesis results with the ground truth. We can make the following observations from these results:

- Images synthesized using the proposed unsupervised approach capture most of the structural information.
- Synthesis of T1-MRI from T2-MRI produces much better results compared to the synthesis of T2-MRI from T1-MRI.
- CSR improves the results while synthesizing T2-MRI from T1-MRI. Images without CSR look a bit noisy compared to the images with CSR (please zoom figures 3 and 4).

CSR contribution: To report the results for coupled sparse representation, we ran the CSR step twenty times for each image. The results reported for CSR in table 1 are averaged over twenty runs. To quantify the statistical significance of the contribution of CSR while synthesizing T2-MRI scans from T1-MRI scans, we also report the minimum (among 20 runs) improvement in the correlation and SNR values along with the corresponding p-values in table 1 of the supplementary material.

Table 1: Correlation and SNR values for the proposed unsupervised synthesis approach.

Subject	Source: T1-MRI, Target: T2-MRI				Source: T2-MRI, Target: T1-MRI			
	Correlation		SNR (dB)		Correlation		SNR (dB)	
	No CSR	CSR	No CSR	CSR	No CSR	CSR	No CSR	CSR
1	0.862	0.877	13.31	13.82	0.932	0.936	16.38	16.68
2	0.839	0.858	13.06	13.77	0.932	0.935	16.39	16.66
3	0.881	0.894	13.77	14.35	0.934	0.939	16.55	16.89
4	0.841	0.855	12.72	13.18	0.933	0.937	16.48	16.77
5	0.814	0.832	11.81	12.44	0.873	0.871	14.99	14.95
6	0.841	0.861	13.24	13.96	0.939	0.943	16.92	17.21
7	0.792	0.811	11.76	12.33	0.900	0.905	14.76	15.00
8	0.833	0.845	12.56	12.98	0.941	0.944	17.09	17.40
9	0.856	0.876	13.21	13.93	0.933	0.938	16.27	16.63
10	0.848	0.863	13.33	13.96	0.936	0.941	16.89	17.28
11	0.871	0.887	13.53	14.22	0.935	0.939	16.73	17.00
12	0.822	0.837	12.34	12.82	0.925	0.930	15.89	16.20
13	0.838	0.852	12.53	13.00	0.926	0.930	16.13	16.40
14	0.861	0.871	13.12	13.55	0.940	0.944	16.93	17.25
15	0.791	0.810	11.96	12.57	0.915	0.921	15.50	15.81
16	0.830	0.847	12.23	12.72	0.936	0.940	16.53	16.86
17	0.851	0.868	13.01	13.65	0.929	0.934	16.13	16.53
18	0.859	0.874	13.19	13.73	0.923	0.928	15.86	16.18
19	0.811	0.824	12.14	12.59	0.924	0.929	15.92	16.15
average	**0.839**	**0.855**	**12.78**	**13.35**	**0.927**	**0.931**	**16.23**	**16.52**

Comparisons: To show the effectiveness of the proposed candidate selection approach, we compare our results with the following methods:

1. **First nearest neighbor (F-NN):** We use the center voxel value of the first nearest neighbor for synthesis.

2. **Average of nearest neighbors (A-NN):** We use the average of the center voxel values of all the K nearest neighbors for synthesis.

3. **Candidate selection using only mutual information (MI-only):** We use the center voxel value of the best candidate selected by optimizing only the global mutual information cost. This is equivalent to removing $SC(W)$ from optimization problem (6).

4. **Candidate selection using only spatial consistency (SC-only):** We use the center voxel value of the best candidate selected by optimizing only the local spatial consistency cost. This is equivalent to removing $H(X_t) - H(X_s, X_t)$ from optimization problem (6).

Table 2 compares the proposed unsupervised approach (MI+SC, no CSR) with the above-described methods in terms of average correlation and SNR values. Figure 4 shows some visual examples comparing the synthesis results of various methods. We can clearly see that the proposed approach gives the best synthesis results. The low correlation and SNR values of F-NN and A-NN methods indicate that directly using the first nearest neighbor or the average of K nearest neighbors is not sufficient for obtaining good synthesis results. While the F-NN method produces very noisy images with spurious structures, the A-NN method produces blurred images. The low correlation and SNR values of MI-only and SC-only methods suggest that using only the global mutual information criterion or only the local spatial consistency criterion would produce inferior synthesis results compared to the proposed approach that uses both criteria. While the images synthesized by the MI-only method are corrupted by salt and pepper type noise, the images synthesized by the SC-only method are missing a lot of structural details (see the circled areas in figure 4). The proposed approach, which uses both MI and SC criteria, is able to get rid of the noise without loosing the structural details.

Table 2: Average correlation and SNR values for various candidate selection approaches.

Measure	Source/ Target	F-NN	A-NN	MI only	SC only	MI+SC no CSR
Correlation	T1 / T2	0.717	0.815	0.808	0.809	**0.839**
	T2 / T1	0.858	0.910	0.903	0.906	**0.927**
SNR (dB)	T1 / T2	10.10	12.41	11.72	12.11	**12.78**
	T2 / T1	13.30	15.45	14.88	15.19	**16.23**

Table 3: Average correlation and SNR values for various synthesis approaches.

Method	Source: T1, Target: T2		Source: T2, Target: T1	
	Correlation	SNR (dB)	Correlation	SNR (dB)
MP [31]	0.875	13.64	0.931	15.13
LSDN [17]	0.892	14.93	0.941	17.39
Proposed (Unsupervised)	0.855	13.35	0.931	16.52
Proposed (Supervised)	**0.908**	**15.30**	**0.953**	**18.33**

Supervised synthesis results: Table 3 compares the synthesis results of the proposed approach under supervised and unsupervised settings with state-of-the-art modality propagation (MP) [31] and location-sensitive deep network (LSDN) [17] methods. We can clearly see that the proposed approach outperforms both the methods under supervised setting. In fact, the proposed unsupervised approach is able to match the performance of supervised modality propagation method while synthesizing T1-MRI from T2-MRI.

Computation time: When ran on a machine with Intel X5650 processor (2.66 GHz, 20 cores), the candidate generation step took 17 minutes (in c++ using OpenMP), the candidate selection step took 15 minutes (with Matlab), and the sparse coding step took 7 minutes [3].

[3] We used the KSVD and OMP toolboxes provided by http://www.cs.technion.ac.il/~ronrubin/software.html.

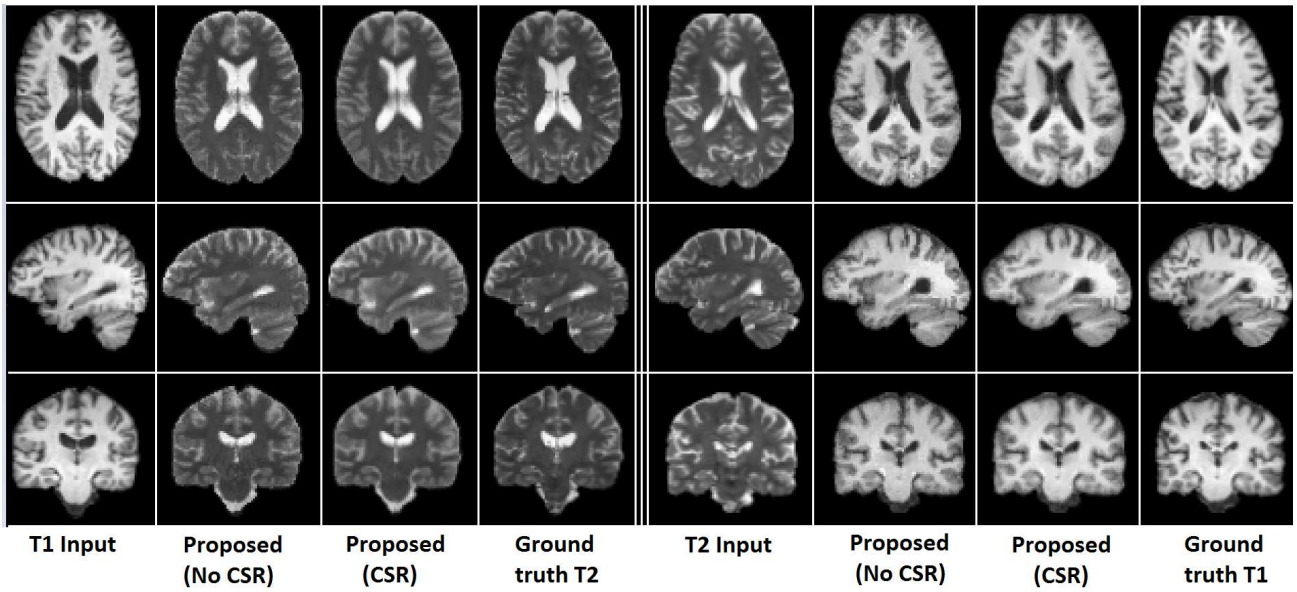

Figure 3: Comparison of the unsupervised synthesis results with ground truth: Left - T2-MRI synthesis from T1-MRI; Right - T1-MRI synthesis from T2-MRI.

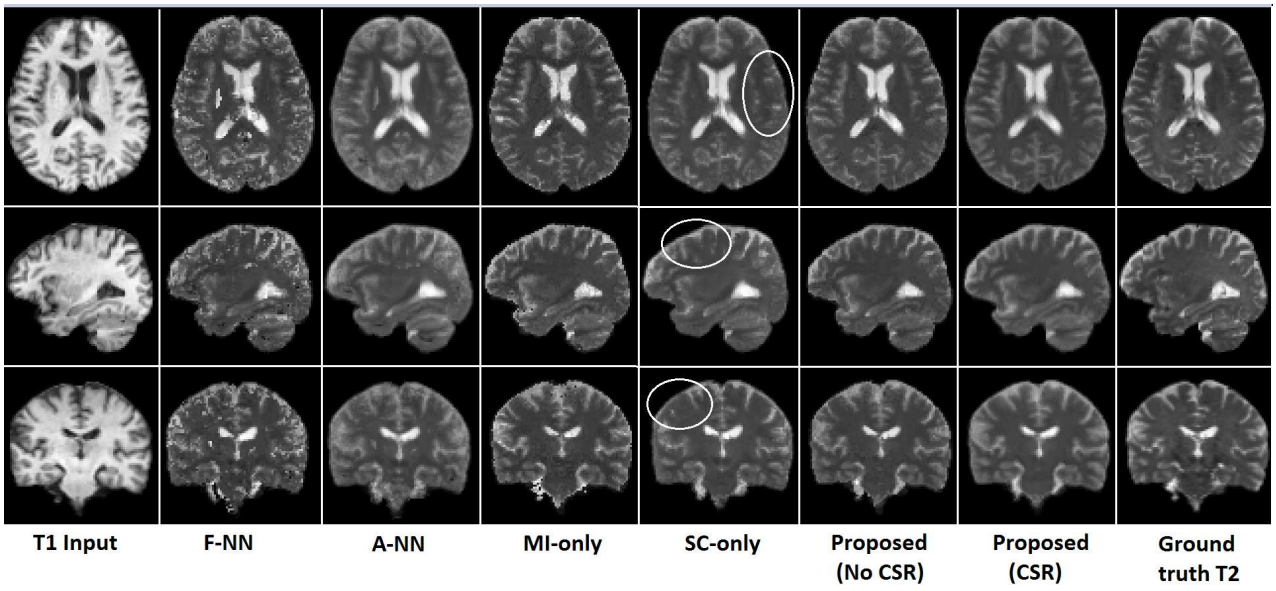

Figure 4: Comparison of various unsupervised synthesis approaches (T2-MRI synthesis from T1-MRI).

5. Conclusions and Future Work

In this paper, we proposed a general unsupervised approach for cross-modal synthesis of subject-specific scans. The proposed approach works without paired training data from source and target modalities. Given a source modality image, we first generated multiple target modality candidate values independently for each voxel using cross-modal nearest neighbor search. Then, we selected the best candidate values jointly for all the voxels by simultaneously maximizing a global mutual information cost and a local spatial consistency cost. Finally, we used coupled sparse representation to further refine the synthesized images. We extended the proposed unsupervised approach to supervised setting by replacing the cross-modal nearest neighbor search with source-modal nearest neighbor search. We experimentally demonstrated the synthesis capabilities of the proposed approach by generating T1-MRI scans from T2-MRI scans and vice versa.

In this work, we mainly focused on synthesizing MRI contrast. In the future, we will apply this approach to other medical imaging modalities. We also plan to use the synthesized images for improving image analysis algorithms like detection and segmentation.

References

[1] M. Aharon, M. Elad, and A. Bruckstein. K-SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation. *IEEE Transactions on Signal Processing*, 54(11):4311–4322, 2006. 5

[2] D. C. Alexander, D. Zikic, J. Zhang, H. Zhang, and A. Criminisi. Image Quality Transfer via Random Forest Regression: Applications in Diffusion MRI. In *MICCAI*, 2014. 2

[3] E. J. Candès, J. K. Romberg, and T. Tao. Robust Uncertainty Principles: Exact Signal Reconstruction from Highly Incomplete Frequency Information. *IEEE Transactions on Information Theory*, 52(2):489–509, 2006. 5

[4] T. Cao, V. Jojic, S. Modla, D. Powell, K. Czymmek, and M. Niethammer. Robust Multimodal Dictionary Learning. In *MICCAI*, 2013. 2, 3, 5

[5] T. Cao, C. Zach, S. Modla, D. Powell, K. Czymmek, and M. Niethammer. Multi-modal Registration for Correlative Microscopy Using Image Analogies. *Medical Image Analysis*, 18(6):914–926, 2014. 1, 2, 3, 5

[6] J. Chen, D. Yi, J. Yang, G. Zhao, S. Z. Li, and M. Pietikäinen. Learning Mappings for Face Synthesis from Near Infrared to Visual Light Images. In *CVPR*, 2009. 2

[7] O. Commowick, S. K. Warfield, and G. Malandain. Using Frankenstein's Creature Paradigm to Build a Patient Specific Atlas. In *MICCAI*, 2013. 1

[8] D. L. Donoho. Compressed Sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006. 5

[9] B. Fischl, D. H. Salat, A. J. W. van der Kouwe, N. Makris, F. Ségonne, B. T. Quinn, and A. M. Dale. Sequence-independent Segmentation of Magnetic Resonance Images. *NeuroImage*, 23:S69 – S84, 2004. 2

[10] A. Hertzmann, C. E. Jacobs, N. Oliver, B. Curless, and D. H. Salesin. Image Analogies. In *Annual Conference on Computer Graphics and Interactive Techniques*, 2001. 1

[11] D. Huang and Y. F. Wang. Coupled Dictionary and Feature Space Learning with Applications to Cross-domain Image Synthesis and Recognition. In *ICCV*, 2013. 3, 5

[12] J. E. Iglesias, E. Konukoglu, D. Zikic, B. Glocker, K. V. Leemput, and B. Fischl. Is Synthesizing MRI Contrast Useful for Inter-modality Analysis? In *MICCAI*, 2013. 1, 2

[13] A. Jog, A. Carass, and J. L. Prince. Improving Magnetic Resonance Resolution with Supervised Learning. In *ISBI*, 2014. 1

[14] A. Jog, S. Roy, A. Carass, and J. L. Prince. Magnetic Resonance Image Synthesis through Patch Regression. In *ISBI*, 2013. 1, 2

[15] E. Konukoglu, A. J. W. van der Kouwe, M. R. Sabuncu, and B. Fischl. Example-based Restoration of High Resolution Magnetic Resonance Image Acquisitions. In *MICCAI*, 2013. 2

[16] D. Kroon and K. Slump. MRI Modalitiy Transformation in Demon Registration. In *ISBI*, 2009. 2

[17] H. V. Nguyen, S. K. Zhou, and R. Vemulapalli. Cross-Domain Synthesis of Medical Images Using Efficient Location-Sensitive Deep Network. In *MICCAI*, 2015. 1, 2, 6, 7

[18] J. Nuyts, G. Bal, F. Kehren, M. Fenchel, C. Michel, and C. Watson. Completion of a Truncated Attenuation Image from the Attenuated PET Emission Data. *IEEE Transactions on Medical Imaging*, 32(2):237–246, 2013. 1

[19] J. P. W. Pluim, J. B. A. Maintz, and M. A. Viergever. Mutual Information-based Registration of Medical Images: A Survey. *IEEE Transactions on Medical Imaging*, 22(8):986–1004, 2003. 3, 4

[20] A. Prakosa, M. Sermesant, H. Delingette, S. Marchesseau, E. Saloux, P. Allain, N. Villain, and N. Ayache. Generation of Synthetic but Visually Realistic Time Series of Cardiac Images Combining a Biophysical Model and Clinical Images. *IEEE Transactions on Medical Imaging*, 32(1):99–109, 2013. 1

[21] F. Rousseau. Brain Hallucination. In *ECCV*, 2008. 2

[22] F. Rousseau, P. A. Habas, and C. Studholme. A Supervised Patch-based Approach for Human Brain Labeling. *IEEE Transactions on Medical Imaging*, 30(10):1852–1862, 2011. 1, 2

[23] S. Roy, A. Carass, A. Jog, J. L. Prince, and J. Lee. MR to CT Registration of Brains Using Image Synthesis. In *SPIE Medical Imaging*, 2014. 1

[24] 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. 1, 2, 3, 5

[25] S. Roy, A. Carass, N. Shiee, D. L. Pham, and J. L. Prince. MR Contrast Synthesis for Lesion Segmentation. In *ISBI*, 2010. 1, 2

[26] 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. 1, 2, 5

[27] S. Wang, L. Zhang, Y. Liang, and Q. Pan. Semi-coupled Dictionary Learning with Applications to Image Super-resolution and Photo-sketch Synthesis. In *CVPR*, 2012. 2

[28] W. Wein, S. Brunke, A. Khamene, M. R. Callstrom, and N. Navab. Automatic CT-Ultrasound Registration for Diagnostic Imaging and Image-guided Intervention. *Medical Image Analysis*, 12(5):577–585, 2008. 1, 2

[29] J. Yang, Z. Wang, Z. Lin, X. Shu, and T. S. Huang. Bilevel Sparse Coding for Coupled Feature Spaces. In *CVPR*, 2012. 5

[30] 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. 5

[31] 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*, 2013. 1, 2, 6, 7

[32] Y. Zhang, G. Wu, P.-T. Yap, Q. Feng, J. Lian, W. Chen, and D. Shen. Hierarchical Patch-based Sparse Representation - A New Approach for Resolution Enhancement of 4D-CT Lung Data. *IEEE Transactions on Medical Imaging*, 31(11):1993–2005, 2012. 1