

APPENDIX A

In this appendix, we provide a detailed derivation of CDLMRI described in Section III-B.

A. CDLMRI

Stage 1) Coupled Dictionary Learning (details). The dictionary update step solves the optimization problem (9) which is reproduced again for convenience:

$$\begin{aligned} & \underset{\Psi_c, \Psi, \Phi_c, \Phi}{\text{minimize}} \sum_{ij} \left\{ \|\mathbf{R}_{ij}\mathbf{x}^{(1)} - (\Psi_c \mathbf{z}_{ij} + \Psi \mathbf{u}_{ij})\|_2^2 \right. \\ & \quad \left. + \|\mathbf{R}_{ij}\mathbf{x}^{(2)} - (\Phi_c \mathbf{z}_{ij} + \Phi \mathbf{v}_{ij})\|_2^2 \right\} \\ & \text{subject to} \quad \left\| \begin{bmatrix} \psi_{ck} \\ \phi_{ck} \end{bmatrix} \right\|_2^2 \leq 1, \|\psi_k\|_2^2 \leq 1, \|\phi_k\|_2^2 \leq 1, \forall k. \end{aligned}$$

Given a subset of the patches that constitute the training dataset $\mathbf{X}^{(1)} = [\dots, \mathbf{x}_{ij}^{(1)}, \dots]$ and $\mathbf{X}^{(2)} = [\dots, \mathbf{x}_{ij}^{(2)}, \dots]$ in Stage 1), the optimization problem (9) is equivalent to:

$$\begin{aligned} & \underset{\Psi_c, \Psi, \Phi_c, \Phi}{\text{minimize}} \left\| \begin{bmatrix} \mathbf{X}^{(1)} - \Psi \mathbf{U} \\ \mathbf{X}^{(2)} - \Phi \mathbf{V} \end{bmatrix} - \begin{bmatrix} \Psi_c \\ \Phi_c \end{bmatrix} \mathbf{Z} \right\|_F^2 \\ & \text{subject to} \quad \left\| \begin{bmatrix} \psi_{ck} \\ \phi_{ck} \end{bmatrix} \right\|_2^2 \leq 1, \|\psi_k\|_2^2 \leq 1, \|\phi_k\|_2^2 \leq 1, \forall k, \end{aligned} \quad (17)$$

where, $\mathbf{Z} = [\dots, \mathbf{z}_{ij}^{(1)}, \dots]$, $\mathbf{U} = [\dots, \mathbf{u}_{ij}^{(1)}, \dots]$, $\mathbf{V} = [\dots, \mathbf{v}_{ij}^{(1)}, \dots]$.

Taking the dictionary update of Ψ_c and Φ_c for example, we update the atom pairs one by one. For the k -th atom pair ψ_{ck} and ϕ_{ck} , we have

$$\begin{aligned} & \min_{\mathbf{d}} \left\| \begin{pmatrix} \mathbf{X}^{(1)} - \Psi \mathbf{U} \\ \mathbf{X}^{(2)} - \Phi \mathbf{V} \end{pmatrix} - \begin{bmatrix} \Psi_c \\ \Phi_c \end{bmatrix} \mathbf{Z} + \begin{bmatrix} \psi_{ck} \\ \phi_{ck} \end{bmatrix}_{old} \mathbf{z}^k - \mathbf{d} \mathbf{z}^k \right\|_F^2 \\ & \text{s.t. } \|\mathbf{d}\|_2^2 \leq 1, \end{aligned}$$

where \mathbf{z}^k denotes the k -th row of \mathbf{Z} .⁴ The term in the parenthesis represents the residual without the contribution from the old k -th atom pair $\begin{bmatrix} \psi_{ck} \\ \phi_{ck} \end{bmatrix}_{old}$. We want to find a

new atom pair $\mathbf{d} = \begin{bmatrix} \psi_{ck} \\ \phi_{ck} \end{bmatrix}_{new}$ to minimize the residual. By expanding the Frobenius norm and removing the constant term, the problem is equivalent to

$$\min_{\mathbf{d}} \frac{1}{2} \|\mathbf{d} \mathbf{z}^k\|_F^2 - \mathbf{d}^H \mathbf{M} \mathbf{z}^{kH} \quad \text{s.t. } \frac{1}{2} \|\mathbf{d}\|_2^2 \leq \frac{1}{2}, \quad (18)$$

where,

$$\mathbf{M} = \begin{bmatrix} \mathbf{X}^{(1)} - \Psi \mathbf{U} \\ \mathbf{X}^{(2)} - \Phi \mathbf{V} \end{bmatrix} - \begin{bmatrix} \Psi_c \\ \Phi_c \end{bmatrix} \mathbf{Z} + \begin{bmatrix} \psi_{ck} \\ \phi_{ck} \end{bmatrix}_{old} \mathbf{z}^k.$$

The Lagrangian of (18) is given by

$$\mathcal{L}(\mathbf{d}, \mu) = \frac{1}{2} \|\mathbf{d} \mathbf{z}^k\|_F^2 - \mathbf{d}^H \mathbf{M} \mathbf{z}^{kH} + \mu \frac{1}{2} \|\mathbf{d}\|_2^2 - \mu \frac{1}{2}.$$

According to the KKT conditions, μ and \mathbf{d} must satisfy

$$\nabla_{\mathbf{d}} \mathcal{L}(\mathbf{d}, \mu) = 0 \Leftrightarrow \mathbf{d}(\mathbf{z}^k \mathbf{z}^{kH} + \mu) = \mathbf{M} \mathbf{z}^{kH}, \quad (19a)$$

$$\|\mathbf{d}\|_2 \leq 1, \quad (19b)$$

$$\mu \geq 0, \quad (19c)$$

$$\mu(\|\mathbf{d}\|_2 - 1) = 0. \quad (19d)$$

⁴Note that \mathbf{z}^k is a row vector resulting from the derivative w.r.t the k -th atom pair, and \mathbf{z}_{ij} is a column vector corresponding to the ij -th patch.

There are two situations:

- Assume $\|\mathbf{d}\|_2 < 1$. By the complementary slackness in (19d), this implies $\mu = 0$. Therefore, from (19a), provided $\mathbf{z} \neq 0$, \mathbf{d} is given by

$$\mathbf{d} = \frac{\mathbf{M} \mathbf{z}^{kH}}{\mathbf{z}^k \mathbf{z}^{kH}}. \quad (20)$$

Note that this case holds only if $\|\mathbf{M} \mathbf{z}^{kH}\| < \mathbf{z}^k \mathbf{z}^{kH}$.

- Assume $\mu > 0$. By the complementary slackness in (19d), this implies $\|\mathbf{d}\|_2 = 1$, and $\mathbf{d} = \mathbf{M} \mathbf{z}^{kH} / (\mathbf{z}^k \mathbf{z}^{kH} + \mu)$. To find μ , use the condition $\|\mathbf{d}\|_2 = 1$, which implies $\|\mathbf{M} \mathbf{z}^{kH}\|_2 = \|\mathbf{z}^k \mathbf{z}^{kH} + \mu\| = \mathbf{z}^k \mathbf{z}^{kH} + \mu$, that is, $\mu = \|\mathbf{M} \mathbf{z}^{kH}\|_2 - \mathbf{z}^k \mathbf{z}^{kH}$. The primal solution in this case is

$$\mathbf{d} = \frac{\mathbf{M} \mathbf{z}^{kH}}{\|\mathbf{M} \mathbf{z}^{kH}\|_2}. \quad (21)$$

Note that this case holds only if $\|\mathbf{M} \mathbf{z}^{kH}\| > \mathbf{z}^k \mathbf{z}^{kH}$ [by (19c)].

Writing both situations in a more compact way, we obtain

$$\begin{bmatrix} \psi_{ck} \\ \phi_{ck} \end{bmatrix}_{new} \leftarrow \mathbf{d} = \frac{\mathbf{M} \mathbf{z}^{kH}}{\max \left\{ \mathbf{z}^k \mathbf{z}^{kH}, \|\mathbf{M} \mathbf{z}^{kH}\|_2 \right\}}, \quad (22)$$

which exists for $\mathbf{z}^k \neq 0$. If $\mathbf{z}^k = 0$, any \mathbf{d} with $\|\mathbf{d}\|_2 \leq 1$ is a solution. Also note that, if $\|\mathbf{M} \mathbf{z}^{kH}\| = \mathbf{z}^k \mathbf{z}^{kH}$, (20) is equivalent to (21), therefore, (22) is still valid.

The dictionary update of Ψ and Φ is performed in a similar way. In order to accelerate the training, the proposed algorithm is updated to online training version where we split the training dataset into small batches, feed them into the algorithm batch by batch, and introduce extra auxiliary variables to accumulate the contribution from each batch.

Stage 2) Coupled Sparse Denoising (details). In this stage, we update the sparse representations of all the patches by solving the following sparse coding problems:⁵

$$\begin{aligned} & \min_{\mathbf{z}_{ij}} \max \left\{ \|\mathbf{R}_{ij}\mathbf{x}^{(1)} - \Psi_c \mathbf{z}_{ij}\|_2^2 + \|\mathbf{R}_{ij}\mathbf{x}^{(2)} - \Phi_c \mathbf{z}_{ij}\|_2^2, \epsilon_c \right\} \\ & \text{s.t. } \|\mathbf{z}_{ij}\|_0 \leq s_c. \end{aligned}$$

$$\begin{aligned} & \min_{\mathbf{u}_{ij}} \max \left\{ \|\mathbf{R}_{ij}\mathbf{x}^{(1)} - \Psi_c \mathbf{z}_{ij} - \Psi \mathbf{u}_{ij}\|_2^2, \epsilon_1 \right\} \\ & \text{s.t. } \|\mathbf{u}_{ij}\|_0 \leq s_1, \end{aligned}$$

$$\begin{aligned} & \min_{\mathbf{v}_{ij}} \max \left\{ \|\mathbf{R}_{ij}\mathbf{x}^{(2)} - \Phi_c \mathbf{z}_{ij} - \Phi \mathbf{v}_{ij}\|_2^2, \epsilon_2 \right\} \\ & \text{s.t. } \|\mathbf{v}_{ij}\|_0 \leq s_2. \end{aligned}$$

Here, the combination of error thresholds ϵ_c , ϵ_1 and ϵ_2 with sparsity thresholds s_c , s_1 and s_2 allows early stopping in OMP. In particular, once the objective value for the patch (i, j) decreases below the expected error threshold, the OMP loop terminates with no need to search more non-zero elements until reaching the maximum sparsity thresholds. This combination of error thresholds with sparsity thresholds is very effective in the acceleration of sparse coding. In the early outer iterations, the reconstructed target contrast tends to be very noisy, so it is better to obtain a sparser representation that

⁵The objectives of (10), (11) and (12) are convex as minimizing the maximum of convex functions is a convex problem. Due to the non-convex ℓ_0 norm constraints, the optimizations are non-convex. Greedy algorithms, such as OMP, are able to provide satisfactory estimations.

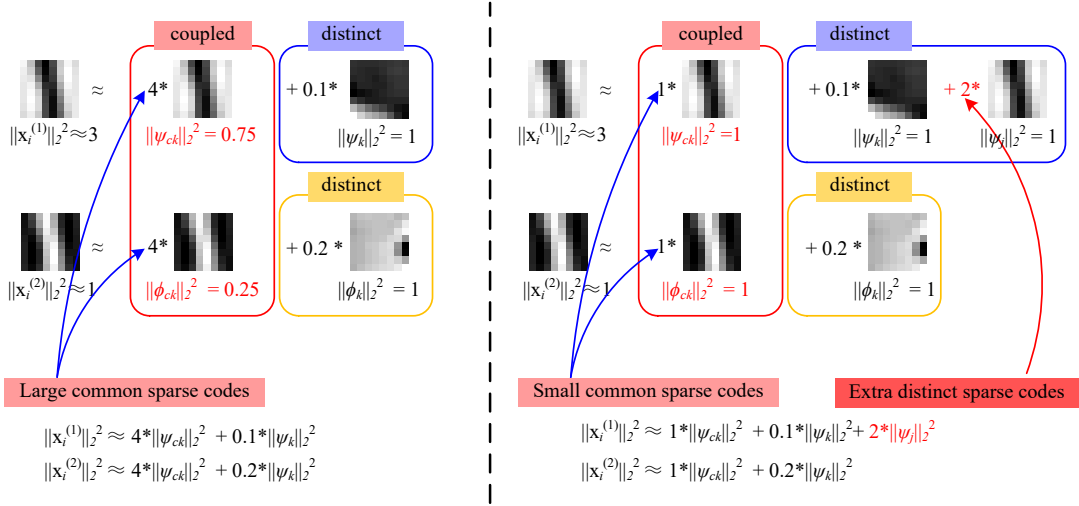


Figure 9: Illustration of the rationale for enforcing joint unit ℓ_2 norm constraint for coupled atoms (left), instead of separate unit ℓ_2 norm constraints (right). For a pair of patches with power ratio $\|\mathbf{x}_i^{(1)}\|_2^2 / \|\mathbf{x}_i^{(2)}\|_2^2 \approx 3$, the joint unit ℓ_2 norm constraint enforces learned coupled atoms having similar power ratio, e.g. $\|\psi_{ck}\|_2^2 / \|\phi_{ck}\|_2^2 \approx 3$ in order to use the same common coefficient to well represent such patch pairs, as shown in the left subfigure. In contrast, the separate unit ℓ_2 norm constraints may find more (maybe large) distinct coefficients. The joint unit ℓ_2 norm constraint setting allows us to promote the contribution of the coupled dictionaries in order to better capture the relationship between different contrasts.

has less non-zero elements in order to enhance the denoising effect. In addition, this combination of error thresholds with sparsity thresholds also allows flat or smooth patches to be represented very sparsely with only a few non-zero elements, thus avoiding possible artifacts due to dense representations.

On the other hand, considering that the quality of the reconstructed target image improves along the outer iterations and very sparse representation may not give a good approximation to the patches, it is better to add more non-zero elements in the sparse representations in order to provide a good approximation and ensure the data fidelity. This motivates us to decrease the thresholds ϵ_c , ϵ_1 and ϵ_2 linearly along with the outer iterations to control the expected sparsity effectively. In this way, the very sparse representation in the beginning enhances the denoising effect and less sparse representation at the end encourages a good approximation. In practice, this strategy improves the speed of our algorithm greatly, as well as contributes to improved performance.

Stage 3) Enforcing k -space Consistency (details). In this stage, we enforce the consistency between the denoised image and its measurements in the k -space domain. In particular, given the estimated patches $\hat{\mathbf{x}}_{ij}^{(1)}$ from Stage 2), this step is formulated as the least squares problem in (13):

$$\min_{\mathbf{x}^{(1)}} \sum_{ij} \left\| \mathbf{R}_{ij} \mathbf{x}^{(1)} - \hat{\mathbf{x}}_{ij}^{(1)} \right\|_2^2 + \nu_1 \left\| \mathbf{F}_{u1} \mathbf{x}^{(1)} - \mathbf{y}^{(1)} \right\|_2^2,$$

where $\hat{\mathbf{x}}_{ij}^{(1)}$ is a denoised patch obtained from Stage 2.

For sampling schemes on a uniform grid of the k -space, such as Gaussian random 2D sampling and Cartesian 1D sampling, there exists an analytical solution satisfying the

normal equation:

$$\left(\sum_{ij} \mathbf{R}_{ij}^H \mathbf{R}_{ij} + \nu_1 \mathbf{F}_{u1}^H \mathbf{F}_{u1} \right) \mathbf{x}^{(1)} = \sum_{ij} \mathbf{R}_{ij}^H \hat{\mathbf{x}}_{ij}^{(1)} + \nu_1 \mathbf{F}_{u1}^H \mathbf{y}^{(1)}, \quad (23)$$

where the superscript $()^H$ denotes the Hermitian transpose operation. The term $\sum_{ij} \mathbf{R}_{ij}^H \mathbf{R}_{ij} \in \mathbb{R}^{N \times N}$ is a diagonal matrix where each diagonal entry is the number of overlapping patches at the corresponding pixel location in $\mathbf{x}^{(1)}$. Assuming that patches wrap around at image boundaries [2], the number of overlapping patches at each pixel is the same, denoted by β .⁶ Thus, the term $\frac{1}{\beta} \sum_{ij} \mathbf{R}_{ij}^H \hat{\mathbf{x}}_{ij}^{(1)}$ represents the denoised image $\hat{\mathbf{x}}^{(1)}$, where the intensity value of each pixel is the average of all the overlapping patches that cover this pixel.

Multiplying by the normalized full Fourier transform matrix \mathbf{F} on both sides of (23) leads to

$$\begin{aligned} & \left(\mathbf{F} \sum_{ij} \mathbf{R}_{ij}^H \mathbf{R}_{ij} \mathbf{F}^H + \nu_1 \mathbf{F} \mathbf{F}_{u1}^H \mathbf{F}_{u1} \mathbf{F}^H \right) \mathbf{F} \mathbf{x}^{(1)} \\ &= \mathbf{F} \sum_{ij} \mathbf{R}_{ij}^H \hat{\mathbf{x}}_{ij}^{(1)} + \nu_1 \mathbf{F} \mathbf{F}_{u1}^H \mathbf{y}^{(1)}. \end{aligned} \quad (24)$$

The matrix $\mathbf{F} \mathbf{F}_{u1}^H \mathbf{F}_{u1} \mathbf{F}^H$ is a diagonal matrix consisting of ones (corresponding to sampling locations in k -space) and zeros. Under the "wrap around" assumption, $\mathbf{F} \sum_{ij} \mathbf{R}_{ij}^H \mathbf{R}_{ij} \mathbf{F}^H = \beta \mathbf{I}_P$. Thus, the matrix pre-multiplying $\mathbf{F} \mathbf{x}^{(1)}$ in (24) is diagonal and trivially invertible. The vector $\mathbf{F} \mathbf{F}_{u1}^H \mathbf{y}^{(1)}$ denotes the zero-filled Fourier measurements. Di-

⁶In particular, $\beta = n$ when the overlap stride $r = 1$, where the *overlap stride* is defined as the distance in pixels between corresponding pixel locations in adjacent image patches.

viding both sides of (24) by the constant β , leads to

$$\tilde{\mathbf{y}}_{[p]}^{(1)} = \begin{cases} (\mathbf{F}\hat{\mathbf{x}}^{(1)})_{[p]}, & [p] \notin \Omega^{(1)} \\ \frac{1}{1 + \tilde{\nu}_1} (\mathbf{F}\hat{\mathbf{x}}^{(1)} + \tilde{\nu}_1 \mathbf{F}\mathbf{F}_{u1}^H \mathbf{y}^{(1)})_{[p]}, & [p] \in \Omega^{(1)} \end{cases}$$

where $\tilde{\nu}_1 = \nu_1/\beta$, $\hat{\mathbf{x}}^{(1)} = \frac{1}{\beta} \sum_{ij} \mathbf{R}_{ij}^H \hat{\mathbf{x}}_{ij}^{(1)}$ is the denoised image, β denotes the number of overlapping patches at the corresponding pixel location in $\mathbf{x}^{(1)}$, $\Omega^{(1)}$ is the subset of k -space that has been sampled. We denote by $\tilde{\mathbf{y}}_{[p]}^{(1)}$ the updated value at location $[p]$ in the vectorized k -space. Specifically, if the location $[p]$ is not sampled, i.e. $[p] \notin \Omega^{(1)}$, we use the Fourier transform results of the denoised image $\hat{\mathbf{x}}^{(1)}$ to interpolate this location. Otherwise, if the location $[p]$ is sampled, i.e. $[p] \in \Omega^{(1)}$, we consider both the Fourier transform results of the denoised image $\hat{\mathbf{x}}^{(1)}$ and original samples, then use their weighted sum to interpolate this location. This immediately results in the solution:

$$\hat{\mathbf{x}}^{(1)} = \mathbf{F}^H \tilde{\mathbf{y}}^{(1)}$$

where \mathbf{F}^H denotes the conjugate of the Fourier transform matrix and $\tilde{\mathbf{y}}^{(1)}$ is the estimated k -space samples as in (15).

For sampling schemes on a non-uniform grid of the k -space, such as radial sampling, we propose to perform data consistency via a gradient step. Specifically, Equation (13) contains two data fidelity terms: the first one enforcing fidelity to denoised image $\hat{\mathbf{x}}^{(1)}$ which results from sparse coding, and the second one enforcing fidelity to k -space samples $\mathbf{y}^{(1)}$.

$$\min_{\mathbf{x}^{(1)}} \sum_{ij} \left\| \mathbf{R}_{ij} \mathbf{x}^{(1)} - \hat{\mathbf{x}}_{ij}^{(1)} \right\|_2^2 + \nu_1 \left\| \mathbf{F}_{u1} \mathbf{x}^{(1)} - \mathbf{y}^{(1)} \right\|_2^2$$

One needs to balance between these two terms to get an appropriate reconstruction. Due to non-uniform sampling, simultaneously solving both terms, like Equation (15), is not applicable here. Therefore, we address two terms sequentially. In particular, we enforce the first data fidelity term via setting $\mathbf{x}^{(1)} = \hat{\mathbf{x}}^{(1)}$. Then we combine it with the gradient descent for the second term $\mathbf{x}^{(1)} - \eta \mathbf{F}_{u1}^H (\mathbf{F}_{u1} \mathbf{x}^{(1)} - \mathbf{y}^{(1)})$. Finally the result is assigned to $\hat{\mathbf{x}}^{(1)}$. These operations lead to Equation (16), i.e. $\hat{\mathbf{x}}^{(1)} = \hat{\mathbf{x}}^{(1)} - \eta \mathbf{F}_{u1}^H (\mathbf{F}_{u1} \hat{\mathbf{x}}^{(1)} - \mathbf{y}^{(1)})$.

Finally, we comment on the rationale for enforcing unit ℓ_2 norm for coupled atoms jointly in the formulation (7). The intuition of this joint unit ℓ_2 norm constraint is based on the observation that patches from different contrasts generally have different power (measured by ℓ_2 norm), thus the learned coupled atoms should also capture corresponding powers. For example, as shown in Fig. 9, assuming that the black-edge patch $\mathbf{x}_i^{(1)}$ with a white background has larger power (squared ℓ_2 norm approximate 3), while the white-edge patch $\mathbf{x}_i^{(2)}$ with a black background has smaller power (squared ℓ_2 norm approximate 1), this leads to a power ratio around $\|\mathbf{x}_i^{(1)}\|_2^2 / \|\mathbf{x}_i^{(2)}\|_2^2 \approx 3$. If we train our coupled dictionaries using a set of such patch pairs, it is natural to expect the learned coupled atoms to have similar power ratio, like $\|\psi_{ck}\|_2^2 / \|\phi_{ck}\|_2^2 \approx 3$ in order to use the same common coefficient to well represent such patch pairs, as shown in the left subfigure of Fig. 9. The joint unit ℓ_2 norm constraint helps achieve this via enforcing a pair of coupled atoms to share unit

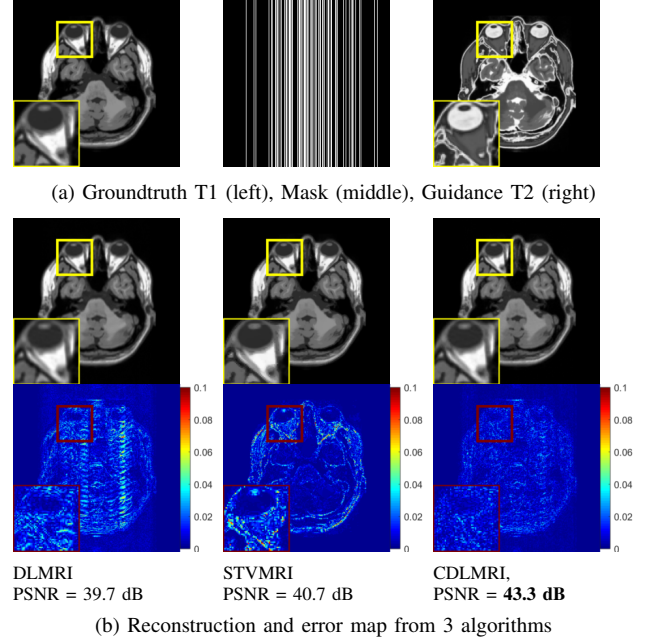


Figure 10: Guided reconstruction for T1, with fully-sampled T2 as reference with 4 fold Cartesian 1D random under-sampling. Sub-figure (a) shows the groundtruth T1-weighted contrast, sampling mask and guidance T2-weighted contrast. Sub-figure (b) shows the reconstructed images and the corresponding residual error for DLMRI [2], STVMRI [21], and the proposed CDLMRI.

power according to their contribution in the representation of a set of patches. Otherwise, if we use separate unit ℓ_2 norm constraints, as shown in the right subfigure of Fig. 9, it is more likely to have more larger coefficients in the distinct sparse representations, which is not our desire.

APPENDIX B. EXPERIMENTS

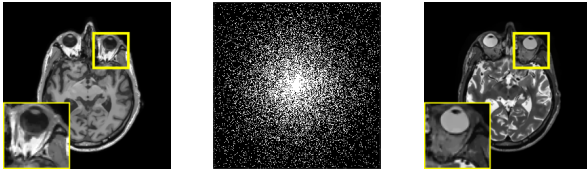
In this appendix, we provide additional experiments complementing those in Section IV.

A. Guided Reconstruction

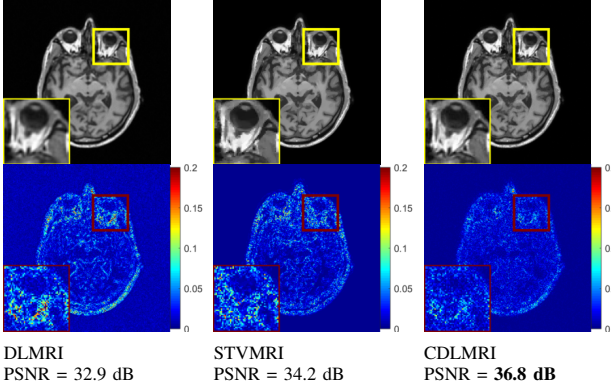
1) *Guided Reconstruction for T1, with T2 as reference:* Figures 10, 11 and 12 show reconstruction results for the scenario where a variable density Cartesian mask is employed for under-sampling on the target T1-weighted contrast, with a fully sampled T2-weighted MRI for guidance contrast.

2) *Guided Reconstruction for FLAIR, with T2 as reference:* In this experiment, the task is to reconstruct the target contrast, a Fluid Attenuated Inversion Recovery (FLAIR) image, from under-sampled measurements, by capitalizing on similarity to a fully-sampled T2-weighted image. Fig. 13 shows the groundtruth FLAIR contrast and fully sampled T2-weighted version, as well as the reconstructed images and corresponding residual error maps from RefMRI [19], FJGP [24] and CDLMRI.

3) *Guided Reconstruction for adjacent MRI slices:* In this application, we extend our CDLMRI to reconstruct one MRI slice by capitalizing on similarity to an adjacent slice. The T2-weighted slices for a brain are collected from available datasets

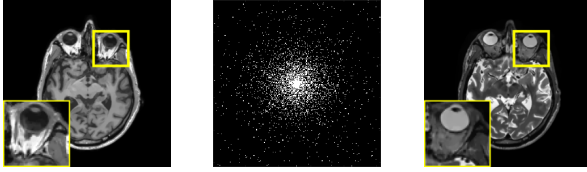


(a) Groundtruth T1 (left), Mask (middle), Guidance T2 (right)

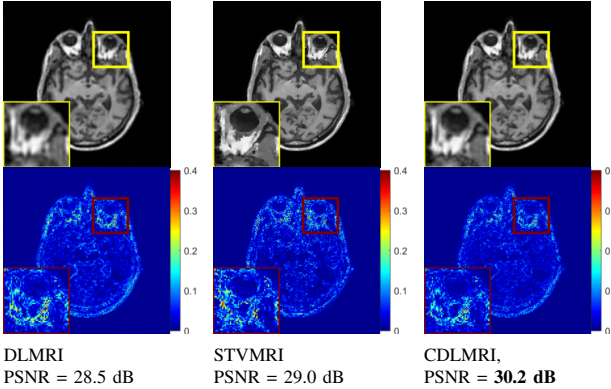


(b) Reconstruction and error map of 3 algorithms

Figure 11: Guided reconstruction for T1, with fully-sampled T2 as reference with 5 fold 2D random under-sampling, using DLMRI [2], STVMRI [21], and the proposed CDLMRI.



(a) Groundtruth T1 (left), Mask (middle), Guidance T2 (right)

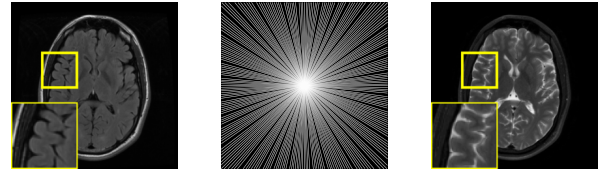


(b) Reconstruction and error map from 3 algorithms

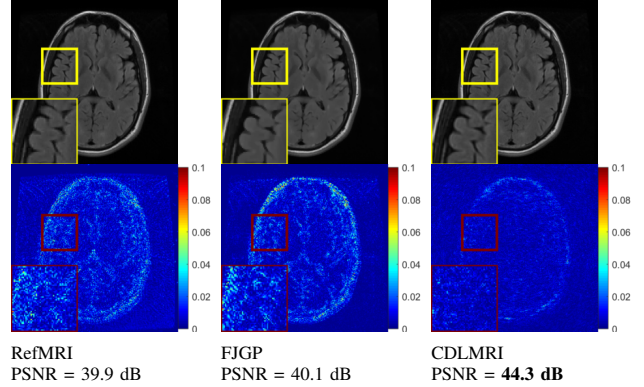
Figure 12: Guided reconstruction for T1, with fully-sampled T2 as reference with 20 fold 2D random under-sampling, using DLMRI [2], STVMRI [21], and the proposed CDLMRI.

of [19]. We acquire 15% (i.e. 6.67 fold under-sampling) k -space data of the target slice with radial sampling. We compare CDLMRI with RefMRI [19] and FJGP [24] which use the same guidance slice as ours.

Fig. 14 shows the groundtruth T2-weighted slices, as well as the reconstructed slices and corresponding residual error maps from RefMRI [19], FJGP [24] and CDLMRI. It is noticed that the reconstruction with RefMRI [19] and FJGP [24] cannot restore fine structure details well and also result in noticeable

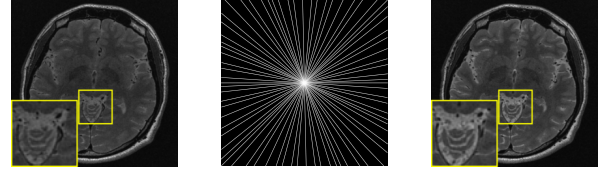


(a) Groundtruth FLAIR (left), Mask (middle), Guidance T2 (right)

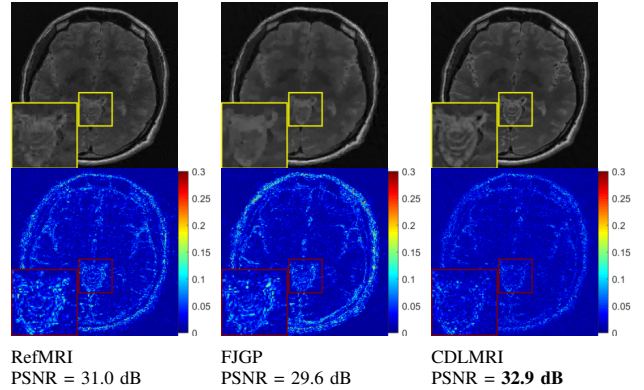


(b) Reconstruction and error map from 3 algorithms

Figure 13: Guided reconstruction for FLAIR, with fully-sampled T2 as reference with 2 fold radial under-sampling, using RefMRI [19], FJGP [24], and the proposed CDLMRI.



(a) Groundtruth Target Slice (left), Mask (middle), Guidance Slice (right)



(b) Reconstruction and error map from 3 algorithms

Figure 14: Guided reconstruction for adjacent MRI slices with 6.67 fold radial under-sampling, using RefMRI [19], FJGP [24], and the proposed CDLMRI.

blurred or over-smoothed regions. See the zoom-in regions. In contrast, the reconstruction from CDLMRI are more visually appealing, with sharper and interpretable fine details. The performance improvement is also demonstrated by the PSNR gains with 1.9 dB and 3.3 dB improvement over RefMRI [19] and FJGP [24], respectively.

B. Joint Reconstruction

1) *Joint Reconstruction for T1- and T2-weighted:* In this experiment, both T1- and T2-weighted contrasts are under-

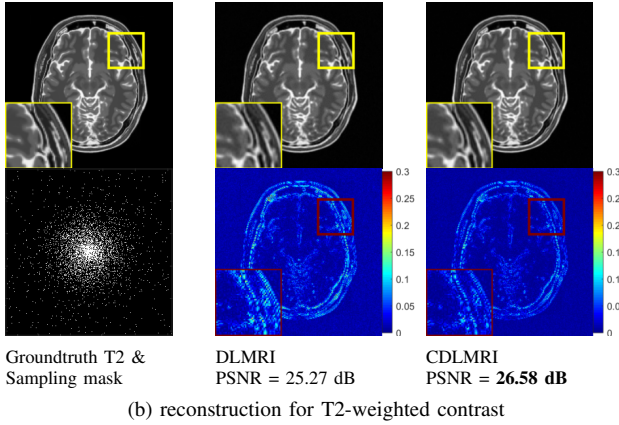
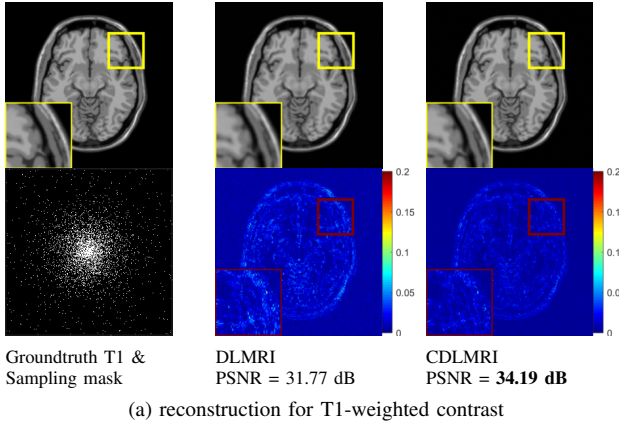


Figure 15: Joint reconstruction for both T1- and T2-weighted MRIs with 20 fold 2D random under-sampling using DLMRI [2] and proposed CDLMRI.

sampled using two different 2D random sampling masks, but with the same sampling ratio. Then, we perform joint reconstruction for both contrasts from their k -space samples. Fig. 15 shows joint reconstruction with 20 fold 2D random under-sampling using DLMRI [2] and CDLMRI. Fig. 16 shows the quantitative performance at 5, 10, 15, and 20 fold under-sampling for both T1-weighted and T2-weighted MRIs. It can be noticed that CDLMRI consistently outperforms DLMRI in terms of PSNR, leading to smaller residual error.

2) *Joint Reconstruction for FLAIR and T2-weighted*: In this experiment, both FLAIR and T2-weighted contrasts are under-sampled using the same radial sampling, and are then jointly reconstructed. Fig. 17 shows the visual performance using 2 fold radial sampling. It can be seen that the proposed CDLMRI outperforms FJGP [24] in terms of both visual and quantitative metrics.

C. Impact of parameters

In this section, we explore the impact of some key parameters in reference-based experiments.

Under-sampling ratio. We evaluate the performance with respect to different under-sampling ratios for 2D random sampling schemes. Fig. 18 shows that CDLMRI outperforms competing methods DLMRI [2] and STVMRI [21] for different undersampling ratios in terms of PSNR. The curve of PSNR with respect to outer iterations also demonstrates stable

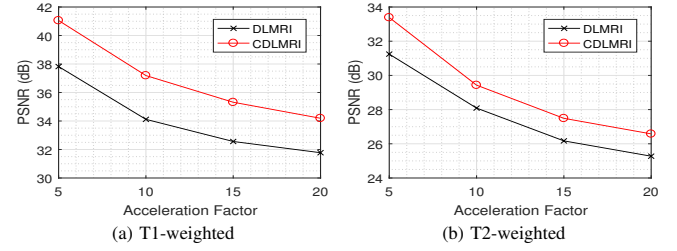


Figure 16: Joint reconstruction for both T1- and T2-weighted MRIs with different 2D random under-sampling folds, using DLMRI [2] and CDLMRI. It can be noticed that CDLMRI consistently outperforms DLMRI.

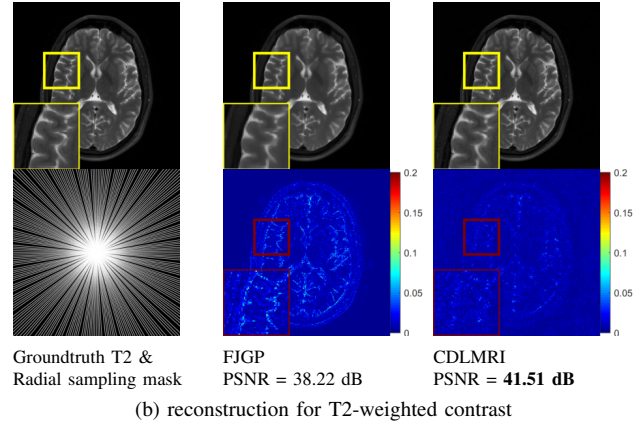
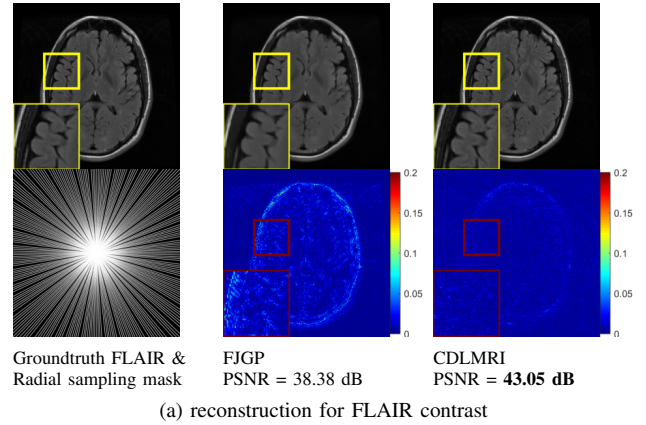


Figure 17: Joint reconstruction for both FLAIR and T2-weighted MRIs with 2 fold radial under-sampling using FJGP [24] and the proposed CDLMRI. Sub-figure (a) shows the groundtruth and reconstruction for FLAIR contrast and sub-figure (b) shows the groundtruth and reconstruction for T2 contrast.

convergence of CDLMRI and the advantage of adjusting the patch overlapping stride logically. It is noticed that in the first 40 outer iterations (overlap stride $r = 1$), the PSNR increases steadily as the smallest stride gives the most overlapping patches and thus results in the strongest average effect, which effectively removes noise. In the last 10 outer iterations (overlap stride $r = 6$), the fewer overlapping patches reduce over-smoothness effects and preserve shape edges better. Thus the image quality is enhanced further, shown by the increase in the PSNR, especially for smaller sampling ratio, such as 5

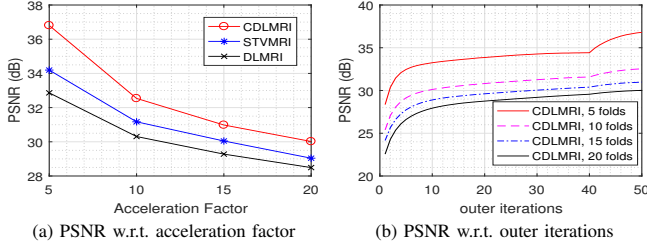


Figure 18: (a) Impact of different 2D random under-sampling folds for DLMRI [2], STVMRI [21] and CDLMRI. (b) the convergence of PSNR with respect to outer iterations for CDLMRI. The first 40 outer iterations correspond to patch overlap stride = 1 and the last 10 outer iterations correspond to patch overlap stride = 6 for CDLMRI.

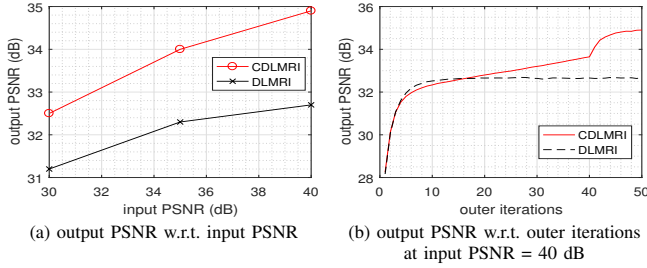


Figure 19: Impact of different noise levels. (a) PSNR with respect to noise at 5 fold 2D random under-sampling for DLMRI [2] and CDLMRI. (b) the convergence of PSNR with respect to outer iterations at input PSNR = 40 dB. The first 40 outer iterations correspond to patch overlap stride = 1 and the last 10 outer iterations correspond to patch overlap stride = 6 for CDLMRI.

fold.

A few factors account for the better performance of CDLMRI over competing methods. One critical factor is the improved dictionary learning algorithm and sparse coding strategy. The use of a more sophisticated model also plays an important role in the performance, as it allows to capture complex dependencies, including structural and gray-scale similarity between multi-contrast MRIs and in turn takes advantage of these powerful priors for guided or joint reconstruction. Changing the stride parameter was applied in CDLMRI experiments to fine-tune the reconstruction only in the fine-tune stage, i.e. the last 10 iterations, instead of the whole process, since experimental results showed that using a larger stride jeopardizes the denoising and de-blurring effect in the earlier stage. In addition, it is noticed that changing the stride only brings benefits for smaller undersampling factors, such as 5 fold. For heavy undersampling situations, such as 10, 15, 20 fold, changing the stride gives negligible benefits as demonstrated in Fig. 18 (b).

Different noise levels. In this experiment, we add white Gaussian noise with zero-mean and standard deviation $\sigma = [9, 5, 2.8]$ to the k -space of the target T1-weighted contrast and reduce the corresponding input PSNR to [30 dB, 35 dB, 40 dB], respectively. We still use a fully-sampled T2-weighted contrast as guidance and a 5 fold 2D random under-sampling mask as in previous experiments. Fig. 19 shows the

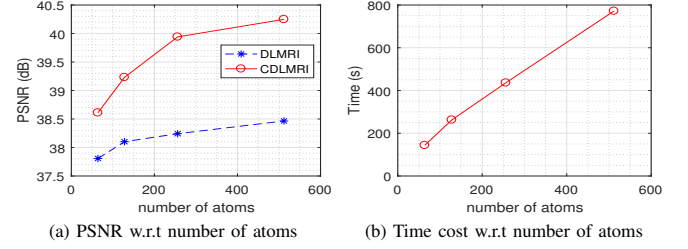


Figure 20: Impact of dictionary size on CDLMRI. It shows that the PSNR improves with the increase of the dictionary size from 64, to 128, 256 and 512. On the other hand, the computation time increases almost linearly with the dictionary size as well.

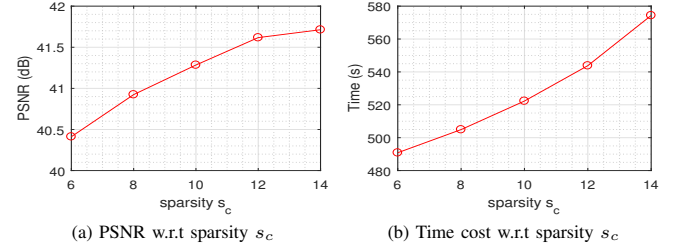


Figure 21: Impact of sparsity constraints.

reconstruction performance of CDLMRI and DLMRI [2]. It can be seen that the proposed CDLMRI is more robust to noise than DLMRI.

Dictionary Size. We explore the effect of dictionary size, i.e., the number of atoms and evaluate the corresponding performance of the proposed approach. Intuitively, more atoms tend to capture more features. Thus, a larger dictionary may yield a more accurate sparse approximation to the image of interest. On the other hand, a large dictionary size increases the complexity of the non-convex problem, thus requiring more computation. Fig. 20 shows the comparison of CDLMRI with DLMRI [2] at various number of atoms, such as $K = 64$, $K = 128$, $K = 256$, and $K = 512$ in terms of PSNR and computation time. In specific, each dictionary Ψ_c , Ψ , Φ_c , Φ of CDLMRI have K atoms and the dictionary of DLMRI has also the same number K of atoms. There is no fine-tune stage that uses a different stride for CDLMRI, and the same number of outer iterations are applied to both DLMRI and CDLMRI. It is noticed that CDLMRI gives better performance than DLMRI quantitatively in terms of PSNR. In particular, the PSNR improvement of DLMRI using more atoms is marginal while the proposed CDLMRI brings a notable improvement. On the other hand, the performance improvement is at the expense of linearly increased computation burden. But, owing to improved training strategy (e.g. minibatch training, stochastic dictionary updating, early stopping for coupled sparse coding, decreasing error thresholds), the increment rate of computational complexity using CDLMRI can be to some extent controlled. It shows that a dictionary size of 512 atoms for CDLMRI yields decent results while allowing affordable computational complexity. In contrast, a dictionary size of 128 or 64 atoms seems a reasonable choice for DLMRI to balance well between reconstruction quality and complexity.

Sparsity Constraints. A larger sparsity constraint can lead to better approximation of the data. On the other hand, larger sparsity also requires more iterations to find these non-zeros via OMP. In this experiment, we exploit the impact of common sparsity constraint s_c ranging from 6 to 14, with distinct sparsity constraints set as $s_1 = s_2 = \text{ceil}(0.2s_c)$. As shown in Fig. 21, the PSNR of the reconstruction, as well as the computational time, increase along with the sparsity constraint. When the sparsity constraint goes beyond a certain level, e.g., 10, the retrieved extra non-zeros coefficients are trivial and contribute very little to the PSNR.

Parameter Setting. The proposed algorithm requires some parameters to be set, such as dictionary size, sparsity constraints, patch size, and iterations. We choose appropriate values for a few critical parameters via investigating their impact and sensitivity. We found that more dictionary atoms tend to provide better reconstruction quality at the price of increased computation time. However, when the number of atoms is large enough, the gains in reconstruction quality are negligible compared to the additional computational cost they entail. Regarding the sparsity constraints, a larger value encourages data fidelity and leads to a less sparse solution. In contrast, a smaller value encourages a sparser solution and therefore improve denoising and de-aliasing. Since the patch size is equal to the dimension of a dictionary atom, larger patch sizes imply larger dictionaries and, thus, a larger quantity of training data. In addition, a larger patch usually contains more complex patterns, thus requiring more atoms for representation. As a consequence, the complexity of training and sparse coding increases. On the other hand, a larger patch captures more overlapping areas among adjacent patches, and therefore strengthens the averaging effect and promotes the denoising and de-aliasing effect. The inner iterations for dictionary learning should be set large enough in order to ensure convergence, especially for the final iterations. However, more inner iterations increase the computational burden. We found that 50 is a good choice for inner iterations. A similar rule of thumb also applies to the number of outer iterations which is set to between 40 and 60. The algorithm is not sensitive to the remaining parameters (λ , variance threshold, etc.) and can therefore be set to default values.

The parameters ν_1 and ν_2 balance fidelity between the model and the measurements. Intuitively, if the measurement noise is weak, one should trust more the k-space samples, and therefore use a large ν to enforce greater fidelity to the measurements. In contrast, if the measurement noise is strong, one should reduce confidence on the k-space samples and thus use a small ν . We thus set $\nu_1 = \lambda/\sigma_1$ and $\nu_2 = \lambda/\sigma_2$, where σ_1 and σ_2 denote the standard deviation of the measurement noise, and λ is a positive constant. In particular, when there is no noise, setting $\nu_1 = \nu_2 \rightarrow \infty$ transforms Equation (15) into $\tilde{\mathbf{y}}_{[p]}^{(1)} = (\mathbf{F}\mathbf{F}_{u1}^H\mathbf{y}^{(1)})_{[p]}, (\forall [p] \in \Omega^{(1)})$, which implies plugging the original k-space samples in the sampling positions. In noisy scenarios, Equation (15) is as before: $\tilde{\mathbf{y}}_{[p]}^{(1)} = \frac{1}{1+\tilde{\nu}_1} (\mathbf{F}\hat{\mathbf{x}}^{(1)} + \tilde{\nu}_1\mathbf{F}\mathbf{F}_{u1}^H\mathbf{y}^{(1)})_{[p]}, (\forall [p] \in \Omega^{(1)})$, which implies plugging the weighted sum of the original k-space samples and the Fourier transform of the denoised image $\hat{\mathbf{x}}^{(1)}$

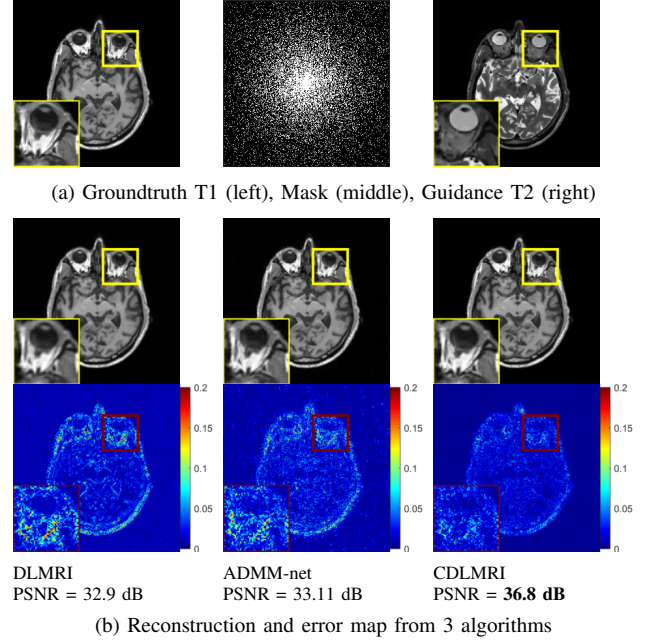


Figure 22: Guided reconstruction for T1, with fully-sampled T2 as reference with 5 fold 2D random under-sampling. Sub-figure (a) shows the groundtruth T1-weighted contrast, sampling mask and guidance T2-weighted contrast. Sub-figure (b) shows the reconstructed images and the corresponding residual error for DLMRI [2], ADMM-net [14], and the proposed CDLMRI.

in the sampling positions.

D. Deep learning based approaches

Some deep learning based single-contrast MRI reconstruction approaches have been presented in [14]–[18]. Deep-learning based multi-contrast MRI reconstruction approaches are also emerging in the literature [28], [29].

Sun et al. [14] propose a deep ADMM-net which is defined over a data flow graph that is derived from the iterative procedures in the Alternating Direction Method of Multipliers (ADMM) algorithm for optimizing a CS-based MRI model. In this way, all the parameters (e.g., transforms, shrinkage functions, penalty parameters, etc.) in the deep architecture can be discriminatively learned from training pairs of under-sampled data in k-space and reconstructed image using fully sampled data by backpropagation over the data flow graph.

Hyun et al. [15] present a Unet based MRI reconstruction approach to learn a function which maps aliasing MRI (from the inverse Fourier transform of zero-filled subsampled k-space data) to the groundtruth MR image. Then, they enforce k-space consistency to obtain updated k-space data, and finally perform IFFT on the updated k-space data to get a better MRI image. Their approach exploits the assumption that the MR images exist in a much lower dimensional manifold embedded in the high dimensional space $\mathcal{C}^{N \times N}$.

Schlemper et al. [16] propose a deep cascade of convolutional neural networks (CNN) where each CNN expresses the dictionary learning reconstruction step in DLMRI [16], thus can be seen as unfolding the optimization process of DLMRI.

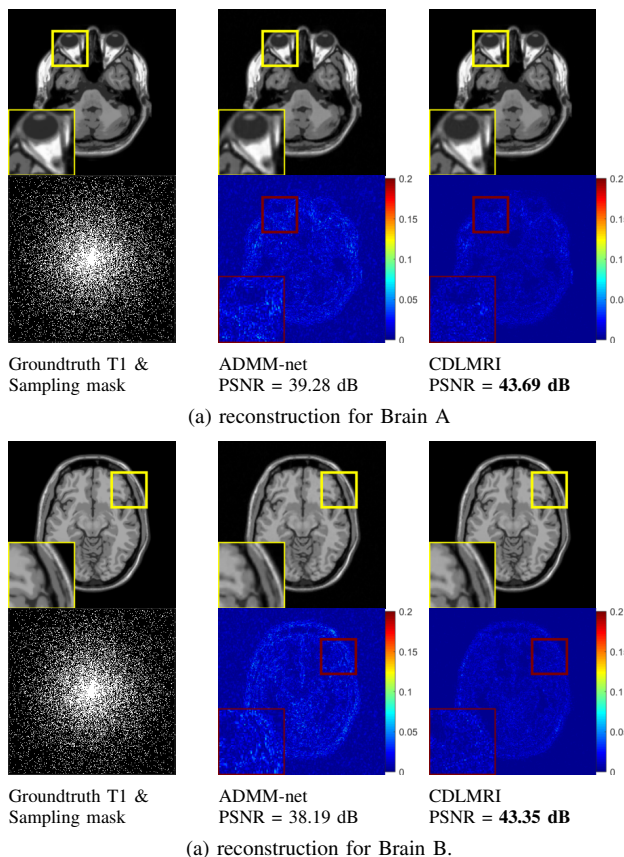


Figure 23: Comparing the proposed CDLMRI with single-contrast MRI reconstruction approach ADMM-net [14] on more contrasts. We use 5 fold 2D random under-sampling and corresponding fully-sampled T2 as reference.

They aim at learning a function which maps aliasing MRI (from the inverse Fourier transform of zero-filled subsampled k-space data) to the groundtruth MR image, similar to Hyun et al. [15]’s Unet MRI reconstruction. But they introduce a cascaded structure with each unit containing a reconstruction operation followed by data consistency operation.

Yang et al. [17] propose a conditional Generative Adversarial Networks-based model (DAGAN for de-aliasing GAN) to reconstruct CS-MRI. In the DAGAN architecture, they designed an improved U-Net architecture with skip connections for the generator network. In addition, they have designed a refinement learning method to stabilise the U-Net based generator, which provides an end-to-end network to reduce aliasing artefacts. To better preserve texture and edges in the reconstruction, they propose a content loss with 4 parts: image domain pixel-wise mean square error (MSE), frequency domain MSE, VGG perceptual loss, and adversarial loss.

We compare our approach with the state-of-the-art method ADMM-net [14]; other approaches are not included in our comparison due to no training or testing code available. We repeat the guided reconstruction experiment for T1 with T2 as reference. We examine two under-sampling patterns: 5 fold 2D random under-sampling and 4 fold Cartesian 1D random under-sampling, the same as in previous experiments.

As shown in Fig. 22, ADMM-net outperforms DLMRI with 5 fold 2D random under-sampling pattern. In particular,

ADMM-net produces better high-frequency components than DLMRI, but introduces noticeable artifacts in smoother areas, such as the background, which makes the reconstruction less appealing than ours. Identical trends are shown in Fig. 23. Our CDLMRI outperforms ADMM-net over 4 dB in average. We attribute the better performance of our method to its ability to leverage additional data modalities, which deep learning based approaches ADMM-net cannot.

E. Comparison with more state-of-the-art methods

We compared our method with state-of-the-art MRI reconstruction methods, including PBDW [34] which exploits patch based directional wavelets for MRI reconstruction, PANO [23] which forms a general patch-based nonlocal operator to leverage sparse representation of similar patches for reconstruction, FDLCP [36] which exploits a fast dictionary learning method to perform MRI reconstruction from classified patches, and pFISTA [35] which develops a projected Fast Iterative Soft-Thresholding Algorithm for MRI reconstruction. Extensive experiments are conducted on highly sub-sampled MRI images with under-sampling factor 20 fold (5%) and 40 fold (2.5%). The performance of various methods is shown in Table I and Fig. 24. They demonstrate that CDLMRI outperforms the competing methods with a large gain both quantitatively and qualitatively. From the reconstructed images, it is noticed that our reconstruction leads to much clearer details (for example, see the brain sulcus and gyrus), and exhibits minimum artifacts, blurring and noise.

F. Application in Magnetic Resonance Fingerprinting

Quantitative Magnetic Resonance Imaging (QMRI) requires successive multi-contrast MR images to perform curve fitting or dictionary matching in order to reconstruct clinically relevant tissue parameters, such as T1, T2 relaxation times. It shows promising benefits in monitoring pathological tissue changes. Magnetic Resonance Fingerprinting (MRF) [47]–[54] emerged as a promising Quantitative Magnetic Resonance Imaging (QMRI) approach, with the capability of providing multiple tissue’s intrinsic spin parameters simultaneously, such as the spin-lattice magnetic relaxation time (T1) and the spin-spin magnetic relaxation time (T2). Based on the fact that the response from each tissue with respect to a given pseudo-random pulse sequence is unique, MRF exploits pseudo-randomized acquisition parameters to create unique temporal signal signatures, analogous to a "fingerprint", for different tissues. A dictionary matching operation is then performed to map an inquiry temporal signature to the best matching entry in a precomputed dictionary, leading to multiple tissue parameters directly. Multi-contrast MRI reconstruction can be applied to MRF to restore a series of MRI contrasts which will be used as temporal signal signatures to find tissue parameters via dictionary matching.

In this section, we evaluate our CDLMRI approach in MRF tasks, and compare the performance with other state-of-the-art MRF methods [47]–[49].⁷ The pseudo-random excitation

⁷The experiment procedures involving human subjects were approved by the Institutional Review Board of Tel-Aviv Sourasky Medical Center, Israel. We thank the authors of [49] for sharing the code and data.

Table I: Comparing with state-of-the-art methods under 20 fold (top table) and 40 fold (bottom table) under-sampling.

| 20 fold under- sampling | PBDW [34] PSNR SNR | | PANO [23] PSNR SNR | | FDLCP [36] PSNR SNR | | pFISTA [35] PSNR SNR | | CDLMRI PSNR SNR | |
|-------------------------------|-----------------------|-------|-----------------------|-------|------------------------|--------------|-------------------------|-------|--------------------|--------------|
| BrainWebA | 30.74 | 19.28 | 32.95 | 21.49 | 32.52 | 21.06 | 34.38 | 22.91 | 35.86 | 24.40 |
| BrainWebB | 29.49 | 19.51 | 32.74 | 22.76 | 24.41 | 14.43 | 33.59 | 23.61 | 35.61 | 25.63 |
| BrainWebC | 31.74 | 20.24 | 34.63 | 23.13 | 33.89 | 22.39 | 35.93 | 24.43 | 37.91 | 26.41 |
| patientA | 27.11 | 15.97 | 28.87 | 17.73 | 28.98 | 17.84 | 29.82 | 18.68 | 30.28 | 19.15 |
| patientB | 22.76 | 12.79 | 24.58 | 14.62 | 24.15 | 14.19 | 24.71 | 14.75 | 26.04 | 16.08 |
| patientC | 33.78 | 20.03 | 35.77 | 22.03 | 38.27 | 24.53 | 36.14 | 22.39 | 37.01 | 23.26 |
| Average | 29.27 | 17.97 | 31.59 | 20.29 | 30.37 | 19.07 | 32.43 | 21.13 | 33.79 | 22.49 |
| 40 fold under- sampling | PBDW [34] PSNR SNR | | PANO [23] PSNR SNR | | FDLCP [36] PSNR SNR | | pFISTA [35] PSNR SNR | | CDLMRI PSNR SNR | |
| BrainWebA | 27.04 | 15.58 | 29.41 | 17.95 | 28.97 | 17.51 | 30.49 | 19.03 | 32.90 | 21.43 |
| BrainWebB | 25.99 | 16.01 | 28.70 | 18.72 | 23.92 | 13.94 | 29.53 | 19.55 | 32.97 | 22.99 |
| BrainWebC | 27.94 | 16.44 | 30.81 | 19.30 | 32.95 | 21.45 | 31.23 | 19.72 | 35.59 | 24.09 |
| patientA | 24.26 | 13.13 | 26.16 | 15.03 | 25.70 | 14.56 | 26.74 | 15.60 | 28.07 | 16.93 |
| patientB | 20.98 | 11.01 | 22.54 | 12.58 | 21.85 | 11.89 | 22.60 | 12.63 | 23.85 | 13.89 |
| patientC | 29.53 | 15.78 | 31.79 | 18.04 | 34.51 | 20.76 | 32.46 | 18.71 | 33.49 | 19.75 |
| Average | 25.96 | 14.66 | 28.24 | 16.94 | 27.98 | 16.69 | 28.84 | 17.54 | 31.15 | 19.85 |

pulse sequence used in our experiments is FISP pulse sequence that is designed with varying parameters such as repetition time (TR), time of echo (TE), and radio frequency flip angle (FA) over time, as shown in Fig. 25. In our experiment, the echo time was constant of 2ms. The repetition time TR was randomly varied in the range of 11.5 - 14.5 ms with a Perlin noise pattern. All the flip angles (FA) constituted a sinusoidal variation in the range of 0 - 70 degrees to ensure smoothly varying transient state of the magnetization. The RF pulse sequence has been used in previous publications in the field of MRF [49], [50], [54]. Given pseudo-random pulse sequence, we synthesize an MRF dictionary \mathbf{D} and a lookup-table \mathbf{LUT} using Bloch equations. Each entry in the dictionary represents a unique temporal signature associated with a specific tissue and its quantitative parameters, such as the T1 and T2 relaxation times, stored in the corresponding lookup-table. A few examples are shown in Fig. 26. Thus, once the best matching (i.e. most correlated) entry is found, it directly leads to multiple tissue parameters simultaneously via a lookup-table operation.

For the range of tissue parameters, T1 relaxation times are set to cover a range of [1, 5000] ms and T2 relaxation times to cover a range of [1, 2000] ms with an increment of 10 ms for both. Such parameter ranges cover the relaxation time values that can be commonly found in a brain scan [55]. All the valid combinations of T1 and T2 values are stacked together, generating a lookup-table \mathbf{LUT} of dimension 80100×2 . Given the lookup-table and RF pulse sequences, dictionary entries are synthesized by solving the Bloch equations using the extended phase graph formalism, leading to a dictionary of dimension 80100×200 where 200 denotes the number of image frames.

The developed CDLMRI method is applied to MR fingerprinting. First, we use CDLMRI to perform the signature restoration, that is, restore a series of MRI contrasts from

their k-space samples. In particular, every two adjacent MRI contrasts are jointly reconstructed using CDLMRI until all the 200 contrasts are recovered from their under-sampled k-space data. Then, the restored MRI contrasts are used to recover quantitative tissue parameters, including T1 and T2 relaxation times via dictionary matching. In the experiments, the anatomical dataset is from work [49], which is constructed from brain scans that were acquired with GE Signa 3T HDXT scanner from a healthy subject. During the k-space under-sampling, the under-sampling ratio is set to be 9%. That is, 9% k-space data is acquired by a series of 2D variable density Gaussian sampling patterns, shown in Fig. 27, leading to a k-space measurement matrix \mathbf{Y} of size 2458×200 . The same k-space measurements \mathbf{Y} and dictionary \mathbf{D} are also used by competing methods [47]–[49] for comparison. The overall performance is evaluated in terms of SNR, PSNR, RMSE, shown in Table II. It shows that our CDLMRI outperforms or is comparable to competing state-of-the-art methods. The reason is that the learned coupled dictionaries are more adaptive to the data modalities than predefined transform matrices, and they also capture the correlations and dependencies between different modalities effectively. Such characteristics promote the sparsity of the joint representation of the data and contribute to better denoising and de-aliasing. Therefore, the restored MRI contrasts exhibit higher quality, which benefits the dictionary matching operation and thus contribute to better parameter restoration performance.

APPENDIX C. ADDITIONAL DISCUSSION

In this appendix, we provide additional discussion related to our technique.

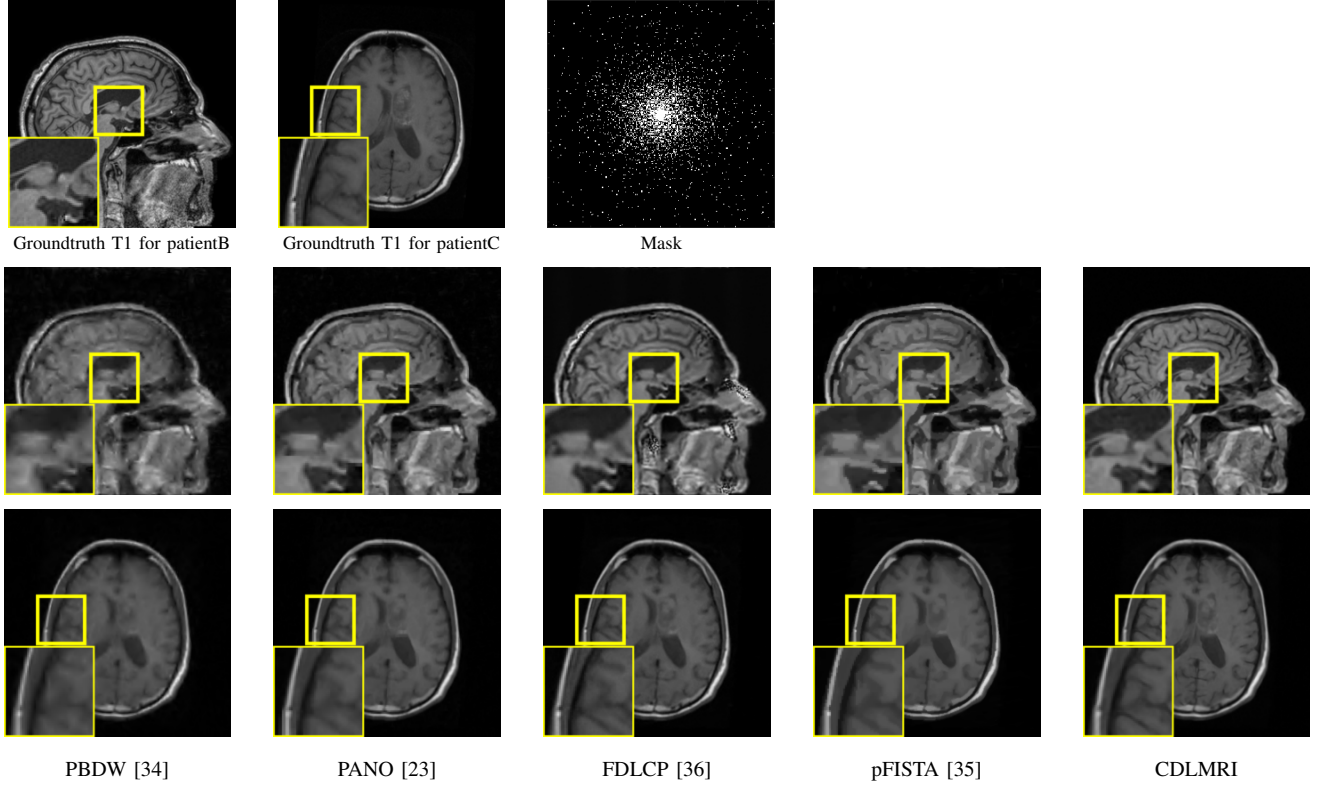


Figure 24: Visual comparison with state-of-the-art methods under 20 fold under-sampling on patientB and patientC.

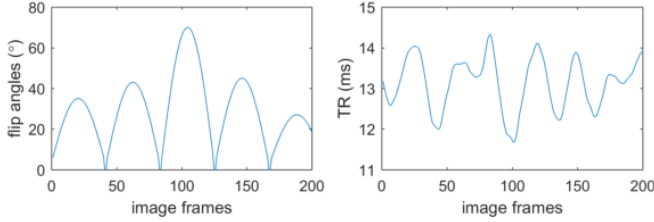


Figure 25: FISP pulse sequence parameters. Left: All the flip angles (FA) constituted a sinusoidal variation in the range of 0 - 70 degrees to ensure smoothly varying transient state of the magnetization. Right: The repetition time (TR) was randomly varied in the range of 11.5 - 14.5 ms with a Perlin noise pattern.

Table II: Quantitative performance evaluation in terms of PSNR, SNR, RMSE and correlation coefficient, with k-space under-sampling ratio 9% and 200 image frames.

| | SNR (dB) | | PSNR (dB) | | RMSE | | CorrCoef | |
|-----------|----------|-------|-----------|-------|--------|-------|----------|------|
| | T1 | T2 | T1 | T2 | T1 | T2 | T1 | T2 |
| Ma [47] | 14.39 | 8.12 | 28.83 | 32.35 | 162.80 | 60.33 | 0.98 | 0.92 |
| BLIP [48] | 11.21 | 10.90 | 25.64 | 35.12 | 234.95 | 43.83 | 0.95 | 0.95 |
| FLOR [49] | 14.11 | 10.00 | 28.55 | 34.23 | 168.18 | 48.59 | 0.97 | 0.94 |
| Ours | 16.01 | 9.60 | 30.45 | 33.82 | 135.06 | 50.92 | 0.98 | 0.93 |

A. Intuitive explanation of the proposed model and its variants

The proposed data model may have other variants for taking advantage of the common information among multi-contrast images. One of them is based on an assumption that they share a common dictionary instead of common sparse representations. In particular, if different contrasts are assumed

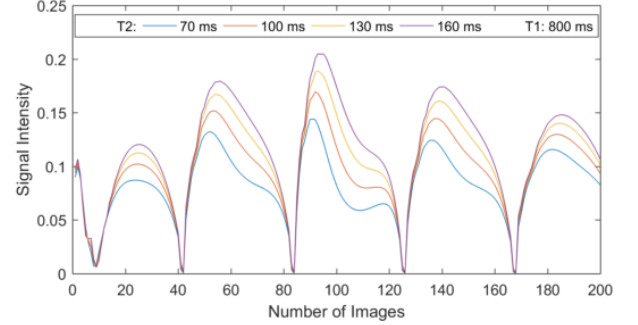


Figure 26: Synthetic MRF temporal signatures with 200 image frames. Temporal signatures corresponding to parameter values $\{(T1, T2)\}$ ms = $\{(800, 70), (800, 100), (800, 130), (800, 160)\}$ ms.

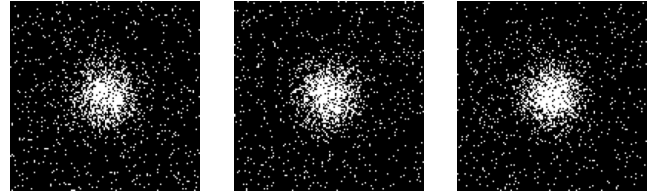


Figure 27: A series of Gaussian patterns used for k-space under-sampling ratio 9%.

to share a common dictionary, it means that they share a set

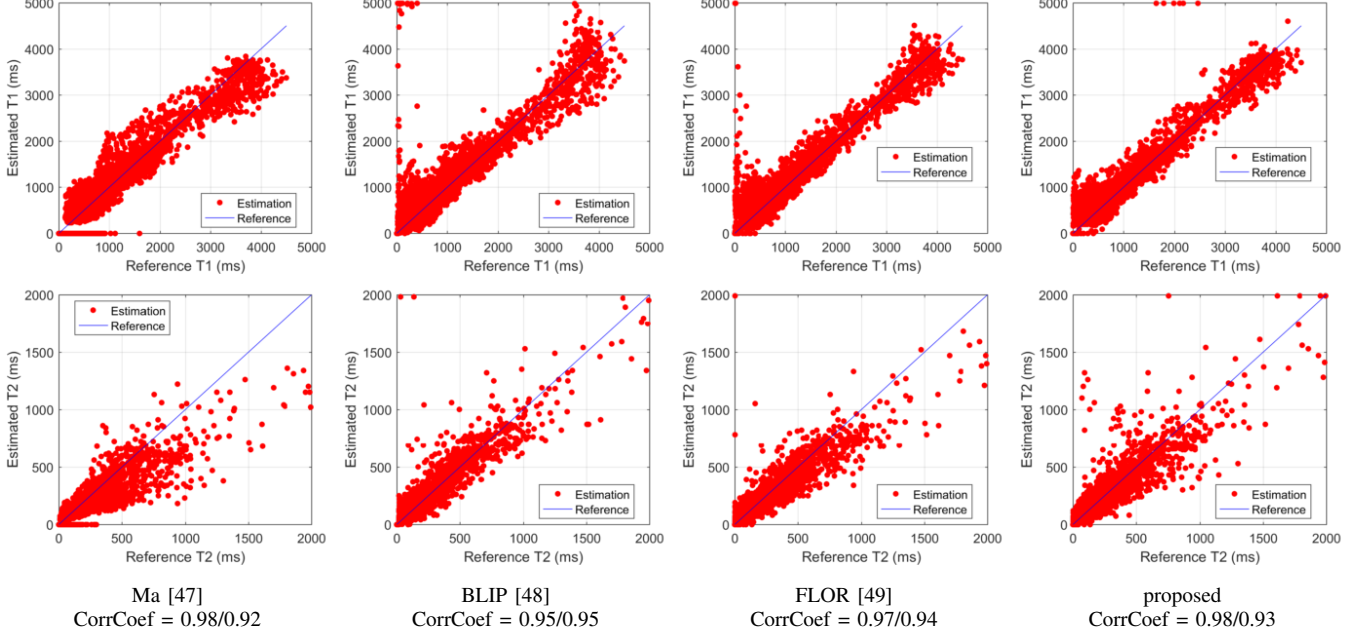


Figure 28: Visual comparison of Ma [47], BLIP [48], FLOR [49] and our approach for MR fingerprinting task using anatomical dataset with k-space under-sampling factor 9% and 200 MRI contrasts.

of identical patterns, leading to the data model given by:

$$\mathbf{x}_{ij}^{(1)} = \mathbf{D} \mathbf{z}_{ij} + \Psi \mathbf{u}_{ij}, \quad (25)$$

$$\mathbf{x}_{ij}^{(2)} = \mathbf{D} \mathbf{w}_{ij} + \Phi \mathbf{v}_{ij}. \quad (26)$$

To make our point clearer, let us ignore the terms $\Psi \mathbf{u}_{ij}$ and $\Phi \mathbf{v}_{ij}$ in (25) and (26), and focus on the main difference resulting from the common dictionary terms $\mathbf{D} \mathbf{z}_{ij}$ and $\mathbf{D} \mathbf{w}_{ij}$. Then equations (25) and (26) can be reformulated as:

$$[\mathbf{x}_{ij}^{(1)} \mathbf{x}_{ij}^{(2)}] = \mathbf{D} [\mathbf{z}_{ij} \mathbf{w}_{ij}]. \quad (27)$$

In comparison, our suggested model given by

$$\mathbf{x}_{ij}^{(1)} = \Psi_c \mathbf{z}_{ij} + \Psi \mathbf{u}_{ij}, \quad (28)$$

$$\mathbf{x}_{ij}^{(2)} = \Phi_c \mathbf{z}_{ij} + \Phi \mathbf{v}_{ij}, \quad (29)$$

leads – under the same simplification – to

$$\begin{bmatrix} \mathbf{x}_{ij}^{(1)} \\ \mathbf{x}_{ij}^{(2)} \end{bmatrix} = \begin{bmatrix} \Psi_c \\ \Phi_c \end{bmatrix} \mathbf{z}_{ij}. \quad (30)$$

It is noticed that model (27) reduces to a model similar to classic dictionary learning. Since $\mathbf{x}_{ij}^{(1)}$ and $\mathbf{x}_{ij}^{(2)}$ are two disjoint parts of the whole training dataset, the contribution from the two modalities is loosely coupled during the dictionary learning. In contrast, model (30) couples $\mathbf{x}_{ij}^{(1)}$ and $\mathbf{x}_{ij}^{(2)}$ together and uses each pair of different contrasts as one training sample, so the contribution from the two modalities are intimately related during coupled dictionary learning. The comparison of model (30) and (27) provides intuitive explanation why our model (28) - (29) are more suitable for capturing dependencies of different modalities than model (25) - (26).

In addition, a set of experiments has been conducted to illustrate the difference of the two models under the same

parameter setting. It is noticed that model (28) - (29) lead to higher PSNR than model (25) - (26) during the reconstruction, as shown in Table III. This demonstrates that our model (28) - (29) exhibit better capacity of capturing the complex dependencies than model (25) - (26) via coupling different modalities more intimately.

B. Incorporation of more image modalities

More image modalities may be incorporated in our scheme in a few ways. One of them is to concatenate all the modalities together and use the same sparse representation for all of them, formulated as

$$\mathbf{x}_{ij}^{(1)} = \Psi_c \mathbf{z}_{ij} + \Psi \mathbf{u}_{ij}, \quad (31)$$

$$\mathbf{x}_{ij}^{(2)} = \Phi_c \mathbf{z}_{ij} + \Phi \mathbf{v}_{ij}, \quad (32)$$

$$\mathbf{x}_{ij}^{(3)} = \Theta_c \mathbf{z}_{ij} + \Theta \mathbf{w}_{ij}. \quad (33)$$

Here, the superscripts $^{(1)}, ^{(2)}, ^{(3)}$ represent 3 different modalities, such as T1-weighted, T2-weighted, and Proton Density (PD)-weighted. This approach assumes that the three modalities share the same sparse representation \mathbf{z}_{ij} with respect to their own dictionaries Ψ_c , Φ_c and Θ_c .

Another way to incorporate more images is to combine them in a pair by pair manner, for example, using $\{\mathbf{x}_{ij}^{(1)}, \mathbf{x}_{ij}^{(2)}\}$ as a pair, $\{\mathbf{x}_{ij}^{(2)}, \mathbf{x}_{ij}^{(3)}\}$ as another pair and $\{\mathbf{x}_{ij}^{(1)}, \mathbf{x}_{ij}^{(3)}\}$ as the last pair. In this way, each pair can be processed using the current model and scheme. Assuming the first case that $\mathbf{x}_{ij}^{(3)}$ serves as the guidance, one may perform guided reconstruction for $\mathbf{x}_{ij}^{(1)}$ and $\mathbf{x}_{ij}^{(2)}$ individually. Assuming the second case that both $\mathbf{x}_{ij}^{(2)}$ and $\mathbf{x}_{ij}^{(3)}$ serve as the guidance, one may reconstruct $\mathbf{x}_{ij}^{(1)}$ guided by $\mathbf{x}_{ij}^{(2)}$ and in the meanwhile also reconstruct another

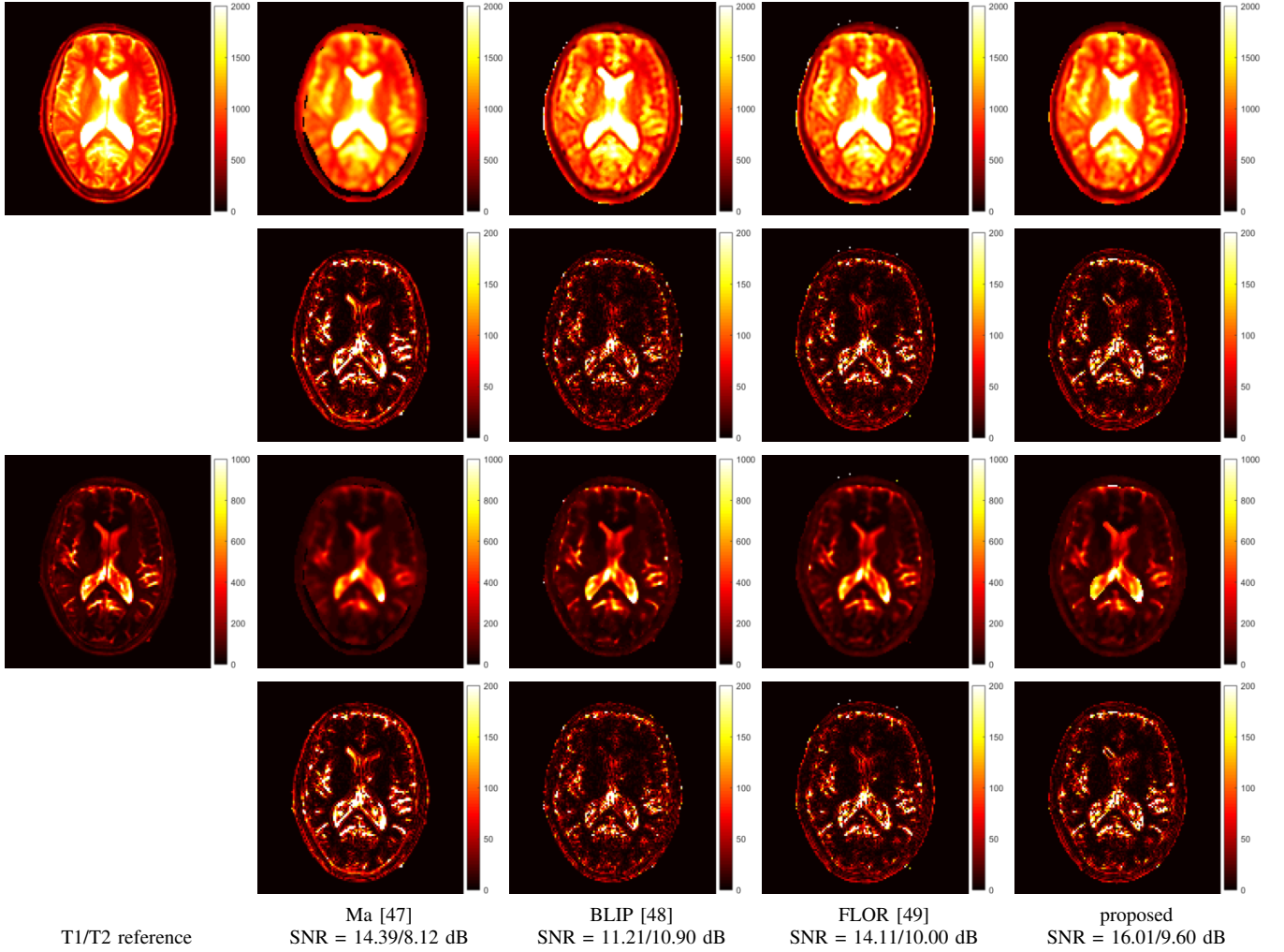


Figure 29: Visual comparison of Ma [47], BLIP [48], FLOR [49] and our approach for MR fingerprinting task using anatomical dataset with k-space under-sampling factor 9% and 200 MRI contrasts. Top two rows are the T1 maps and corresponding residuals while bottom two rows are the T2 maps and residuals. Two SNR values correspond to reconstruction performance for T1 (left) and T2 (right), respectively.

$\mathbf{x}_{ij}^{(1)}$ guided by $\mathbf{x}_{ij}^{(3)}$. Then, two reconstruction versions of $\mathbf{x}_{ij}^{(1)}$ can be fused by a weighted average operation.

C. Time and memory consumption

As we mentioned in the introduction section, the benefits of our algorithm are at the expense of increased computational complexity and memory requirements with respect to methods that have no learning stage [19], [21], [23], [24], [34]. The running time and memory consumption of CDLMRI and other competing methods is shown in Table IV. The running time results show that our computational complexity is comparable with learning-based approach [36], and is smaller than the CS-MRI based method pFISTA [35]. Although the complexity of the proposed CDLMRI algorithm is $O(\delta n K(2s_c + s_1 + s_2)LNT)$ which is on the same order as that of DLMRI [2] $O(\delta n K s LNT)$ in theory, our practical computational complexity is smaller than DLMRI [2] owing to advanced training strategy, including minibatch training, stochastic dictionary updating, early stopping for coupled sparse coding, and decreasing error thresholds.

In our present experiments, the image size is set to 256×256 which is a typical size for practical MRI scenarios. When the patch size is set to 8×8 and the stride is set to 1 pixel (i.e. the minimum value), an image of 256×256 produces $62,001 = (256 - 8 + 1)^2$ patches which occupy 29.6M memory for double-precision floating-point format. In general, given stride s and patch size $n \times n$, an image of size $N \times N$ produces $(\lfloor (N - n)/s \rfloor + 1)^2$ patches. Setting the number of atoms in each dictionary as $K = 512$, every dictionary has a dimension of 64×512 which occupies around 0.24 MB memory for double-precision floating-point format. When using the entire dataset, each sparse code matrix (\mathbf{Z} , \mathbf{U} , and \mathbf{V}) has dimensions of $512 \times 62,001$, and occupies around 2 KB - 0.31 MB memory using MATLAB sparse form, depending on the sparsity. It is observed that the memory cost is dominated by image patches, instead of sparse codes and dictionaries. In addition, we always use mini-batches instead of the whole dataset during training and sparse coding in order to further reduce memory requirements. The total memory cost of all the variables is around 63.3 MB, which is slightly larger than the

Table III: Comparison of the proposed model with its variant.

| MRI image | 5 fold undersampling | | | 20 fold undersampling | | |
|-----------|----------------------|-------------------|-------|-----------------------|-------------------|-------|
| | model (25) - (26) | model (28) - (29) | gains | model (25) - (26) | model (28) - (29) | gains |
| | PSNR | PSNR | | PSNR | PSNR | |
| BrainWebA | 42.54 | 43.69 | 1.14 | 34.16 | 34.77 | 0.61 |
| BrainWebB | 42.14 | 43.37 | 1.23 | 33.39 | 34.71 | 1.33 |
| BrainWebC | 44.10 | 45.45 | 1.35 | 35.24 | 37.08 | 1.84 |
| Average | 42.93 | 44.17 | 1.24 | 34.26 | 35.52 | 1.26 |

Table IV: Average time costs (seconds) and memory consumption (MB) of reconstructing an image of size 256×256 using various methods.

| | RefMRI [19] | FJGP [24] | DLMRI [2] | STVMRI [21] | PBDW [34] | PANO [23] | FDLCP [36] | pFISTA [35] | CDLMRI |
|--------|-------------|-----------|-----------|-------------|-----------|-----------|------------|-------------|--------|
| time | 13.9 | 55.7 | 1263.5 | 12.3 | 44.5 | 11.8 | 122.7 | 504.6 | 268.6 |
| memory | 18.8 | 30.2 | 36.1 | 8.3 | 14.2 | 10.1 | 22.9 | 15.3 | 63.3 |

sum of all image patches, sparse codes and dictionaries due to extra auxiliary variables in use, such as indexing variables, intermediate variables, etc. Auxiliary variables can be cleared after use to release the memory.

A large image may result in a large amount of patches and consume a considerable amount of memory. In order to alleviate the memory burden, we rely on a few tricks in dictionary training and image reconstruction from implementation perspectives.

Reducing the number of training samples. During the training stage, we randomly select around 1/3 of patches for coupled dictionary learning. Of those, we keep the patches whose variance is above a given threshold, and discard the remaining ones that contain negligible information for learning. These tricks reduce the number of patches to around 20,000, which take only 9.6M memory, and accelerate the training without compromising the testing performance. By randomly selecting patches as we do, we bypass the need of extracting all the possible image patches. We only select patches with variance above a threshold and discard those with small variance. The variance threshold has been manually decided prior to the experiment by statistically examining the variance of other MRI datasets. So there is no need to extract all patches from the image of interest to compute the variance threshold. Even though the procedure of discarding patches whose variance is smaller than a given threshold is not informed by any theory, we experimentally found that the specific value of the variance threshold is not critical to the overall performance once it is in a reasonable range. In our experiments, we set the variance threshold to be 0.1.

Reconstructing image batches. Even though the sparse coding based image reconstruction requires all the image patches to be reconstructed in Stage 2, we do not need to achieve this all at once. Instead, a moderate number of patches are extracted at a time as a batch for sparse coding. Then, this batch of denoised patches is inserted into the reconstructed image and then release memory for processing the subsequent batch. In other words, in Stage 2, we sequentially implement the following steps repeatedly until the whole image is reconstructed: (1) extract a batch of image patches from the noisy, blurring target image, (2) perform sparse coding on the

batch, (3) insert them into the denoised image and then release memory. In our experiments, each batch contains 10000 image patches. With these tricks, the memory consumption issue is alleviated, which makes the extension to larger-dimensional imaging problems feasible.

D. Difference with CDLSR [38]

The proposed data model for CDLMRI admits a resemblance to that in CDLSR [38] to some extent, as they both exploit the assumption that different data modalities may share common components represented by identical sparse codes. However, we highlight that there exists a variety of differences between our CDLMRI approach and CDLSR not only in the data model, but also including the task, goal, training strategy, and optimization method.

- Difference in the task and goal. The task in reference [38] is to address a super-resolution problem. The goal focuses on learning a mapping from a low-resolution image modality to a high-resolution image modality in learned sparse-transform domains. CDLMRI solves a reconstruction problem involving denoising and de-aliasing. Accordingly, the goal is to remove artefacts such noise and aliasing resulting from inverse Fourier transforms of sub-sampled k-space measurements.
- Difference in the data model. In reference [38], the data model is used to couple high-resolution (HR), low-resolution (LR) target images with high-resolution guidance images together. In contrast, the data model here involves only low-resolution (LR) target images with high-resolution guidance images, and they are different MRI contrasts sampled in k-space.
- Difference in the training strategy and optimization scheme. In reference [38], the optimization algorithm is adapted from the K-means Singular Value Decomposition (K-SVD). The algorithm in CDLMRI is based on adapted Block Coordinate Descent and its online training version. In [38], the sampling matrix is assumed to be a simple under-sampling matrix in the pixel domain, thus the overall scheme is not iterative. However, the sampling matrix in CDLMRI is a Gaussian random 2D under-sampling pattern, Cartesian 1D under-sampling pattern,

or radial under-sampling pattern applied in the Fourier domain. Thus, we need to consider the k-space consistency problem.

E. Inverse Crime

M. Guerquin-Kern et. al [42] pointed out that the inverse-crime issues happen in the situation where the same discrete model is exactly used for both simulation and reconstruction, leading to artificially good results. In the context of MRI, if algorithms are developed based on simulations of rasterized images, it might not account for the full continuous-domain reality, because it neglects the aliasing that is inherent to spatial discretization. Therefore, resolution-independent simulations are more realistic since they formulate the simulation analytically in the continuous domain, thus removed the bias.

Since the data in the experiments is from rasterized images, there might exist inverse crime [42] which may lead to artificially good results due to the neglect of the aliasing that is inherent to spatial discretization. This issue influences both our approach and the competing approaches. However, since similar papers as well as the competing methods published in TMI also use this strategy for data generation, we believe that we align with the conventional approaches in this domain and the comparison is still fair. In addition, the "forward model" (also termed predictor) in our experiment is a linear model due to sub-sampled Fourier transform, while the "inverse model" (also termed estimator) is a non-linear model due to greedy optimization. The predictor and estimator are very different in our approach, which implies that the "forward model" is not connected with "inverse model". This characteristic helps alleviating the inverse crime issue.

F. Nonlinear Effects

Nonlinear effects, such as off-resonance effects, usually result from field inhomogeneity, eddy current effect, radiofrequency pulse frequency offset, chemical shift effect, etc. To make it simple, we use the same setting as in other literature and did not consider the possible effect of these factors. However, in practical situations, the distortion due to these nonlinear effects can be reduced by adopting some strategies [43]–[46], for example, using higher performance hardware, parallel imaging, Dixon techniques, slice-by-slice shimming, specific sequences that are insensitive to off resonance effects, phase correction based on a inferred field map derived from multi-echo gradient-echo images, and other post-processing corrections.

G. Multi-coil Imaging

Multi-coil MR imaging, a.k.a. multi-channel parallel MR imaging, is a type of important MR imaging technique which has found widespread application in various aspects of MRI. By combining multiple small surface coils into large arrays and using multichannel receiver array coil, it enables MR imaging to achieve higher signal-to-noise ratio and larger fields of view. However, the techniques proposed in our paper are limited to single-coil MRI, and it is not clear yet how to

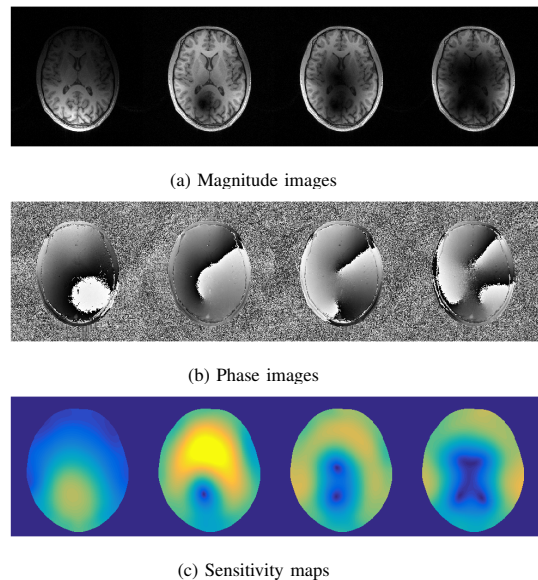


Figure 30: Multi-channel MR magnitude, phase images and sensitivity maps.

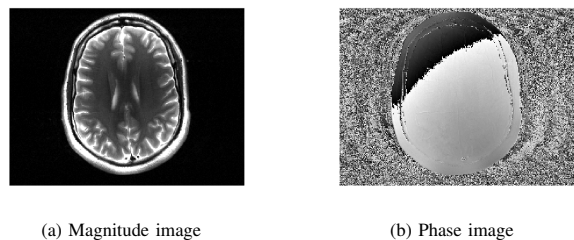


Figure 31: MR magnitude and phase image.

extend them to the multi-coil setting. We conjecture that, to extend the proposed method to multi-coil MR imaging case, a possible direct way is to integrate the multi-coil data together to be standard single-coil form, and then apply the proposed method directly. However, considering special characteristics in multi-coil MR imaging, shown in Fig. 30, it may require to deal with additional issues, for example, how to cope with substantial variation of coil-sensitivity, how to perform complex calibration effectively, how to incorporate specific knowledge about individual coil sensitivities into the coupled dictionary learning algorithm, how to combine the under-sampled data from each receiver coil into an unaliased reconstructed image with the full FOV, etc. These operations would entail significant algorithm development representing an entirely different contribution. So, we would like to leave possible generalizations of this algorithm to the multi-coil setting as future work. Indeed, we highlight in the conclusion that this could represent a relevant direction for future research.

H. Phase Images Reconstruction

Phase images can provide important information for depiction of flow, characterization of susceptibility-induced distortions, MR fingerprinting, etc. However, our study focused exclusively on magnitude images in view of the fact that all methods we compare against also focus on the magnitude images. These include DLMRI [2], STVMRI [21], RefMRI [19],

FJGP [24], PBDW [34], PANO [23], FDLCP [36], and pFISTA [35]. In addition, it has not been investigated whether our approach can be applicable to multi-contrast phase image reconstruction. One of the hallmarks of our approach relates to the fact that different modalities share structure similarity or other common attributes that one can leverage to improve reconstruction. The existence of such common structure is evident across the multi-contrast magnitude images e.g. with T1w and T2w MRIs, but it is less evident in the phase images, as depicted in Fig. 31. Considering that phase images usually contain very less structural information and show very few similar attributes, it is not clear whether there exists meaningful common information for exploitation, and our approach may not lead to significant results for phase image reconstruction.