

Medical Image Imputation from Image Collections

Adrian V. Dalca, Katherine L. Bouman, William T. Freeman, Natalia S. Rost, Mert R. Sabuncu, Polina Golland
for the Alzheimers Disease Neuroimaging Initiative*

Abstract—We present an algorithm for creating high resolution anatomically plausible images consistent with acquired clinical brain MRI scans with large inter-slice spacing. Although large data sets of clinical images contain a wealth of information, time constraints during acquisition result in sparse scans that fail to capture much of the anatomy. These characteristics often render computational analysis impractical as many image analysis algorithms tend to fail when applied to such images. Highly specialized algorithms that explicitly handle sparse slice spacing do not generalize well across problem domains. In contrast, we aim to enable application of existing algorithms that were originally developed for high resolution research scans to significantly undersampled scans. We introduce a generative model that captures fine-scale anatomical structure across subjects in clinical image collections and derive an algorithm for filling in the missing data in scans with large inter-slice spacing. Our experimental results demonstrate that the resulting method outperforms state-of-the-art upsampling super-resolution techniques, and promises to facilitate subsequent analysis not previously possible with scans of this quality. Our implementation is freely available at <https://github.com/adalca/papago>.

Index Terms—Imputation, super-resolution, clinical scans, thick slices, sparse slices, MRI, brain scans

I. INTRODUCTION

Increasingly open image acquisition efforts in clinical practice are driving dramatic increases in the number and size of patient cohorts in clinical archives. Unfortunately, clinical scans are typically of dramatically lower resolution than the research scans that motivate most methodological development. Specifically, while slice *thickness* can vary depending on the clinical study or scan, inter-slice *spacing* is often significantly larger than the in-plane resolution of individual slices. This results in missing voxels that are typically filled via interpolation.

Our work is motivated by a study that includes brain MRI scans of thousands of stroke patients acquired within 48 hours of stroke onset. The study aims to quantify

Adrian V. Dalca is with the Computer Science and Artificial Intelligence Lab, MIT (main contact: adalca@csail.mit.edu) and also Martinos Center for Biomedical Imaging, Massachusetts General Hospital, HMS.

Katherine L. Bouman and Polina Golland are with the Computer Science and Artificial Intelligence Lab, MIT.

William T. Freeman is with the Computer Science and Artificial Intelligence Lab, MIT and Google.

Mert R. Sabuncu is with the the School of Electrical and Computer Engineering, and Meinig School of Biomedical Engineering, Cornell University.

Natalia S. Rost is with the Department of Neurology, Massachusetts General Hospital, HMS.

*Data used in preparation of this article were obtained from the Alzheimers Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

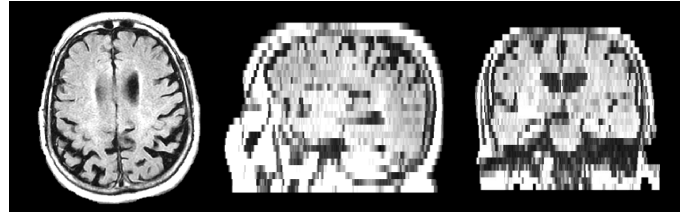


Fig. 1: An example scan from our clinical dataset. The three panels display axial, sagittal and coronal slices, respectively. While axial in-plane resolution can be similar to that of a research scan, slice spacing is significantly larger. We visualize the sagittal and coronal views using nearest neighbor interpolation.

white matter disease burden [23], necessitating skull stripping and deformable registration into a common coordinate frame [27], [31], [32]. The volumes are severely undersampled ($0.85\text{mm} \times 0.85\text{mm} \times 6\text{mm}$) due to constraints of acute stroke care (Fig. 1). Such undersampling is typical of modalities, such as T2-FLAIR, that aim to characterize tissue properties, even in research studies like ADNI [15].

In undersampled scans, the image is no longer smooth, and the anatomical structure may change substantially between consecutive slices (Fig. 1). Since such clinically acquired scans violate underlying assumptions of many algorithms, even basic tasks such as skull stripping and deformable registration present significant challenges, yet are often necessary for downstream analysis [4], [7], [12], [15], [23], [30], [31].

We present a novel method for constructing high resolution anatomically plausible volumetric images consistent with the available slices in sparsely sampled clinical scans. Importantly, our method does not require any high resolution scans or expert annotations for training. It instead imputes the missing structure by learning solely from the available collection of sparsely sampled clinical scans. The restored images represent plausible anatomy. They promise to act as a medium for enabling computational analysis of clinical scans with existing techniques originally developed for high resolution, isotropic research scans. For example, although imputed data should not be used in clinical evaluation, the brain mask obtained through skull stripping of the restored scan can be applied to the original clinical scan to improve subsequent analyses.

A. Prior Work

Many image restoration techniques depend on having enough information in a single image to synthesize data. Traditional interpolation methods, such as linear, cubic or spline [28], assume a functional representation of the image. They treat the low resolution voxels as samples, or observations, and estimate function parameters to infer missing voxel values. Patch-based superresolution algorithms use fine-scale

redundancy within a single scan [10], [11], [19], [20], [22]. The key idea is to fill in the missing details by identifying similar image patches in the same image that might contain relevant detail [19], [22]. This approach depends on having enough repetitive detail in a scan to capture and re-synthesize high frequency information. Unfortunately, clinical images are often characterized by sampling that is too sparse to adequately fit functional representations or provide enough fine-scale information to recover the lost detail. For example, 6mm slice spacing, typical of many clinical scans including our motivating example, is far too high to accurately estimate approximating functions without prior knowledge. In such cases, a single image is unlikely to contain enough fine-scale information to provide anatomically plausible reconstructions in the direction of slice acquisition, as we demonstrate later in the paper.

Alternatively, one can use additional data to synthesize better images. Many superresolution algorithms use multiple scans of the same subject, such as multiple low resolution acquisitions with small shift differences to synthesize a single volume [2], [16], [22]. However, such acquisitions are not commonly available in the clinical setting.

Nonparametric and convolutional neural-network (CNN) based upsampling methods that tackle the problem of super-resolution often rely on an external dataset of high resolution data or cannot handle extreme undersampling present in clinical scans. For example, some methods fill in missing data by matching a low resolution image patch from the input scan with a high resolution image patch from the training dataset [3], [13], [16], [17], [25], [24]. Similarly, CNN-based upsampling methods approximate completion functions, but require high resolution scans for training [8], [21]. A recent approach to improve resolution from a collection of scans with sparse slices jointly upsamples all images using non-local means [26]. However this method has only been demonstrated on slice spacing of roughly three times the in-plane resolution, and in our experience similar non-parametric methods fail to upsample clinical scans with more significant undersampling.

Our work relies on a low dimensional embedding of image patches with missing voxels. Parametric patch methods and low dimensional embeddings have been used to model the common structure of image patches from full resolution images, but are typically not designed to handle missing data. Specifically, priors [33] and Gaussian Mixture Models [35], [36] have been used in both medical and natural images for classification [1] and denoising [9], [36]. The procedures used for training of these models rely on having full resolution patches with no missing data in the training phase.

Unfortunately, high (full) resolution training datasets are not readily available for many image contrasts and scanners, and may not adequately represent pathology or other properties of clinical populations. Acquiring the appropriate high resolution training image data is often infeasible, and here we explicitly focus on the realistic clinical scenario where only sparsely sampled images are available.

B. Method Overview

We take advantage of the fact that local fine scale structure is shared in a population of medical images, and each scan with sparse slices captures some partial aspect of this structure. We borrow ideas from Gaussian Mixture Model (GMM) for image patch priors [36], low dimensional Gaussian embeddings [14], [34], and missing data models [14], [18] to develop a probabilistic generative model for sparse 3D image patches around a particular location using a low-dimensional GMM with partial observations. We derive the EM algorithm for maximum likelihood estimation of the model parameters and discuss related modeling choices. Given a new sparsely sampled scan, the maximum *a posteriori* estimate of the latent structure yields the imputed high resolution image. We evaluate our algorithm using scans from the ADNI cohort, and demonstrate its utility in the context of the motivating stroke study. We investigate the behaviour of our model under different parameter settings, and illustrate an example of potential improvements in the downstream analysis using an example task of skull stripping.

This paper extends the preliminary version of the method presented at the 2017 Conference on Information Processing in Medical Imaging [5]. Here, we improve model inference by removing parameter co-dependency between iterations and providing new parameter initialization. We provide detailed derivations and discuss an alternative related model. Finally, we provide an analysis of important model parameters, present results for more subjects, and illustrate more example reconstructions. The paper is organized as follows. Section II introduces the model and learning algorithm. Section III discusses implementation details. We present experiments and analysis of the algorithm’s behavior in Section IV. We discuss important modeling aspects and related models in Section V. We include an Appendix and Supplementary Material with detailed derivations of the EM algorithm for the proposed models.

II. METHOD

In this section, we construct a generative model for sparse image patches, present the resulting learning algorithm, and describe our image restoration procedure.

Let $\{Y_1, \dots, Y_N\}$ be a collection of scans with large inter-slice spaces, *roughly* aligned into a common atlas space (we use affine transformations in our experiments). For each image Y_i in the collection, only a few slices are observed. We seek to restore an anatomically plausible high resolution volume by imputing the missing voxel values.

We capture local structure using image patches. We assume a constant patch shape, and in our experiments use a 3D 11x11x11 shape. We use y_i to denote a D -length vector that contains voxels of the image patch centered at a certain location in image Y_i . We perform inference at each location independently and stitch the results into the final image as described later in this section. Fig. 2 provides an overview of the method.

A. Generative Model

We treat an image patch as a high dimensional manifestation of a low dimensional representation, with the intuition

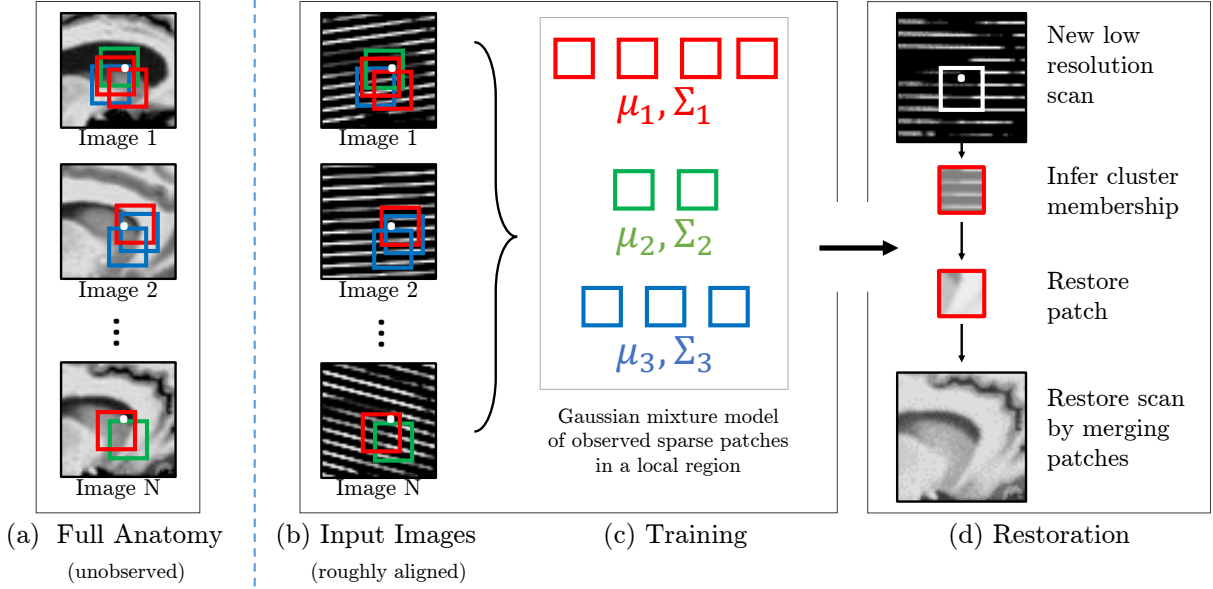


Fig. 2: Image imputation for a subvolume. (a) Full resolution images, shown for illustration only. These are unobserved by the algorithm. (b) Sparse planes acquired in clinical scans. (c) During learning, we train a GMM that captures the low dimensional nature of patch variability in a region around a particular location (white dot). (d) Given a sparsely sampled scan, we infer the most likely cluster for each 3D patch, and restore the missing data using the learned model and the observed voxels. We form the final volume from overlapping restored patches. 2D images are shown for illustration only, the algorithms operate fully in 3D.

that the covariation within image patches has small intrinsic dimensionality relative to the number of voxels in the patch. To capture the anatomical variability across subjects, we employ a Gaussian Mixture Model (GMM) to represent local structure of 3D patches in the vicinity of a particular location across the entire collection. We then explicitly model the observed and missing information. Fig 3 presents the corresponding graphical model.

We model the latent low dimensional patch representation x_i of length $d < D$ as a normal random variable

$$x_i \sim \mathcal{N}(0, I_{d \times d}), \quad (1)$$

where $\mathcal{N}(\mu, \Sigma)$ denotes the multivariate Gaussian distribution with mean μ and covariance Σ . We draw latent cluster assignment k from a categorical distribution defined by a length- K vector π of cluster probabilities, and treat image patch y_i as a high dimensional observation of x_i drawn from a K -component multivariate GMM. Specifically, conditioned on the drawn cluster k ,

$$y_i = \mu_k + W_k x_i + \epsilon_i, \quad \text{where} \quad (2)$$

$$\epsilon_i \sim \mathcal{N}(0, \sigma_k^2 I_{D \times D}), \quad \text{and} \quad \epsilon_i \perp x_i.$$

Vector μ_k is the patch mean of cluster k , matrix W_k shapes the covariance structure of y_i , and σ_k^2 is the variance of image noise. This model implies $\mathbf{E}[y_i | k] = \mu_k$ and $C_k \triangleq \mathbf{E}[(y_i - \mu_k)(y_j - \mu_k)^T | k] = W_k W_k^T + \sigma_k^2 I_{D \times D}$.

Defining $\theta = \{\mu_k, W_k, \sigma_k^2, \pi_k\}_{k=1}^K$, the likelihood of all patches $\mathcal{Y} = \{y_i\}$ at this location under the mixture model is

$$p(\mathcal{Y}; \theta) = \prod_i \sum_k \pi_k \mathcal{N}(y_i; \mu_k, C_k). \quad (3)$$

In our clinical images, only a few slices are known. To model sparse observations, we let \mathcal{O}_i be the set of observed voxels in patch y_i , and $y_i^{\mathcal{O}_i}$ be the corresponding vector of their intensity values:

$$y_i^{\mathcal{O}_i} = \mu_k^{\mathcal{O}_i} + W_k^{\mathcal{O}_i} x_i + \epsilon_i^{\mathcal{O}_i}, \quad (4)$$

where $W_k^{\mathcal{O}_i}$ comprises rows of W_k that correspond to the observed voxel set \mathcal{O}_i . The likelihood of the *observed* data $\mathcal{Y}^\circ = \{y_i^{\mathcal{O}_i}\}$ is therefore

$$p(\mathcal{Y}^\circ; \theta) = \prod_i \sum_k \pi_k \mathcal{N}(y_i^{\mathcal{O}_i}; \mu_k^{\mathcal{O}_i}, C_k^{\mathcal{O}_i \mathcal{O}_i}), \quad (5)$$

where matrix $C_k^{\mathcal{O}_i \mathcal{O}_i}$ extracts the rows and columns of C_k that correspond to the observed voxel subset \mathcal{O}_i .

We do not explicitly model slice thickness, as in many clinical datasets this thickness is unknown or varies by site, scanner or acquisition. Instead, we simply treat the original data as high resolution thin planes and analyze the effects of varying slice thickness on the results in the experimental evaluation of the method.

We also investigated an alternative modeling choice where each missing voxel of patch y_i is modelled as a latent variable. This assumption can optionally be combined with the latent low-dimensional patch representation. We discuss this alternative choice in Section V, and provide parameter updates in the Supplementary Material. Unfortunately, the resulting algorithm is prohibitively slow.

B. Learning

Given a collection of observed patches \mathcal{Y}° , we seek the maximum likelihood estimates of the model parameters $\{\mu_k, W_k, \sigma_k^2\}$ and π under the likelihood (5). We derive the

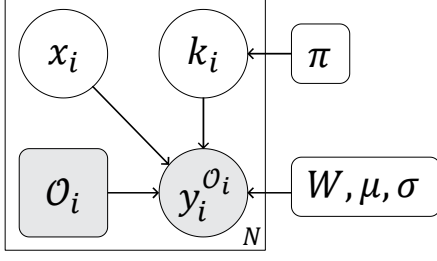


Fig. 3: Graphical representation of our model. Circles indicate random variables and rounded squares represent parameters. Shading represents observed quantities and the plate indicates replication. The observed patch voxels $y_i^{O_i}$ form a subset of patch y_i extracted by the mask O_i and are generated from a multivariate Gaussian distribution conditioned on the latent cluster k_i and the latent patch representation x_i . Parameters μ and W define the mean and the variance of the Gaussian components of the mixture, and σ^2 is the image noise variance.

Expectation Maximization algorithm [6] in Appendix A, and present the update equations and their interpretations below.

The **expectation step** updates the class memberships:

$$\begin{aligned} \gamma_{ik} &\triangleq p(k|y_i^{O_i}; \theta) \\ &= \frac{\pi_k \mathcal{N}(y_i^{O_i}; \mu_k^{O_i}, C_k^{O_i O_i})}{\sum_{k'} \pi_{k'} \mathcal{N}(y_i^{O_i}; \mu_{k'}^{O_i}, C_{k'}^{O_i O_i})}, \end{aligned} \quad (6)$$

and the statistics of the low dimensional representation x_i for each image patch $y_i^{O_i}$ as "explained" by cluster k :

$$\begin{aligned} \hat{x}_{ik} &\triangleq \mathbf{E}[x_i|k] \\ &= ((W_k^{O_i})^T (W_k^{O_i}) + \sigma_k^2 I_{d \times d})^{-1} (W_k^{O_i})^T (y_i^{O_i} - \mu_k^{O_i}), \end{aligned} \quad (7)$$

$$\begin{aligned} S_{ik} &\triangleq \mathbf{E}[x_i x_i^T | k] - \hat{x}_{ik} \hat{x}_{ik}^T \\ &= \sigma_k^2 ((W_k^{O_i})^T (W_k^{O_i}) + \sigma_k^2 I_{d \times d})^{-1}. \end{aligned} \quad (8)$$

We let \mathcal{P}_j be the set of patches in which voxel j is observed, and form the following normalized mean statistics:

$$\delta_{ik} = \frac{\gamma_{ik}}{\sum_{i' \in \mathcal{P}_j} \gamma_{i'k}} \quad (9)$$

$$b_j = \sum_{i \in \mathcal{P}_j} \delta_{ik} \mathbf{E}[x_i | k] = \sum_{i \in \mathcal{P}_j} \delta_{ik} \hat{x}_{ik} \quad (10)$$

$$A_j = \sum_{i \in \mathcal{P}_j} \delta_{ik} \mathbf{E}[x_i x_i^T | k] = \sum_{i \in \mathcal{P}_j} \delta_{ik} (\hat{x}_{ik} \hat{x}_{ik}^T + S_{ik}). \quad (11)$$

The **maximization step** uses the observed voxels to update the model parameters. We let y_i^j be the j^{th} element of vector y_i , and update the cluster mean as a convex combination of observed voxels:

$$\mu_k^j \leftarrow \frac{\sum_{i \in \mathcal{P}_j} \gamma_{ik} (1 - \hat{x}_i^T A_j^{-1} b_j) y_i^j}{\sum_{i' \in \mathcal{P}_j} \gamma_{i'k} (1 - \hat{x}_{i'}^T A_j^{-1} b_j)}. \quad (12)$$

The covariance factors and image noise variance are updated based on the statistics of the low dimensional representation

from (10) and (11):

$$\begin{aligned} W_k^j &\leftarrow \sum_{i \in \mathcal{P}_j} \delta_{ik} (y_i^j - \mu_k^j) \hat{x}_{ik}^T A_j^{-1}, \\ \sigma_k^2 &\leftarrow \frac{\sum_j \sum_{i \in \mathcal{P}_j} \gamma_{ik} [(y_i^j - \mu_k^j - W_k^j \hat{x}_{ik})^2 + W_k^j S_{ik} (W_k^j)^T]}{\sum_j \sum_{i' \in \mathcal{P}_j} \gamma_{i'k}}. \end{aligned} \quad (13)$$

where W_k^j is the j^{th} row of matrix W_k . Finally, we update the cluster proportions:

$$\pi_k = \frac{1}{N} \sum_i \gamma_{ik}. \quad (15)$$

Intuitively, learning our model with sparse data is possible because each image patch provides a slightly different subset of voxel observations that contribute to the parameter estimation (Fig. 2). In our experiments, all subject scans have the same acquisition direction. Despite different affine transformations to the atlas space for each subject, some voxel *pairs* are still never observed in the same patch, resulting in missing entries of the covariance matrix. Using a low-rank approximation for the covariance matrix regularized the estimates.

Upon convergence of the EM updates, we compute the cluster covariance $C_k = W_k W_k^T + \sigma_k^2 I_{D \times D}$ for each k .

C. Imputation

To restore an individual patch y_i , we compute the *maximum-a-posteriori* (MAP) estimate of the image patch:

$$\begin{aligned} \hat{y}_i &= \arg \max_{y_i} p(y_i | y_i^{O_i}; \theta) \\ &= \arg \max_{y_i} \sum_k p(k | y_i^{O_i}) \int_{x_i} p(y_i | x_i) p(x_i | k, y_i^{O_i}) dx_i \\ &= \arg \max_{y_i} \sum_k \gamma_{ik} \int_{x_i} p(y_i | x_i) p(x_i | k, y_i^{O_i}) dx_i \\ &= \arg \max_{y_i} \sum_k \gamma_{ik} \mathcal{N}(y_i; \mu_k + W_k \hat{x}_{ik}, \Sigma_{ik}), \end{aligned}$$

where $\Sigma_{ik} = \sigma_k^2 I_{D \times D} + W_k (S_{ik} + \hat{x}_{ik} \hat{x}_{ik}^T) W_k^T$. Due to the high-dimensional nature of the data, most cluster membership estimates are very close to 0 or 1. We therefore first estimate the most likely cluster \hat{k} for patch y_i by selecting the cluster with the highest membership γ_{ik} . We estimate the low dimensional representation $\hat{x}_{i\hat{k}}$ given the observed voxels $y_i^{O_i}$ using (7), which yields the high resolution imputed patch:

$$\hat{y}_i = \mu_{\hat{k}} + W_{\hat{k}} \hat{x}_{i\hat{k}}. \quad (16)$$

By restoring the scans using this MAP solution, we perform *conditional mean imputation* (c.f. 17, Sec.4.2.2), and demonstrate the reconstructions in our experiments. In addition, our model enables imputation of each patch by sampling the posterior $p(\cdot | y_i^{O_i}; \theta) \approx \mathcal{N}(\cdot; \mu_{\hat{k}} + W_{\hat{k}} \hat{x}_{i\hat{k}}, \Sigma_{i\hat{k}})$, providing a better estimation of the residual noise. Depending on the desired downstream application, sampling-based imputation may be desired.

We average overlapping restored patches using standard techniques [18] to form the restored volume.

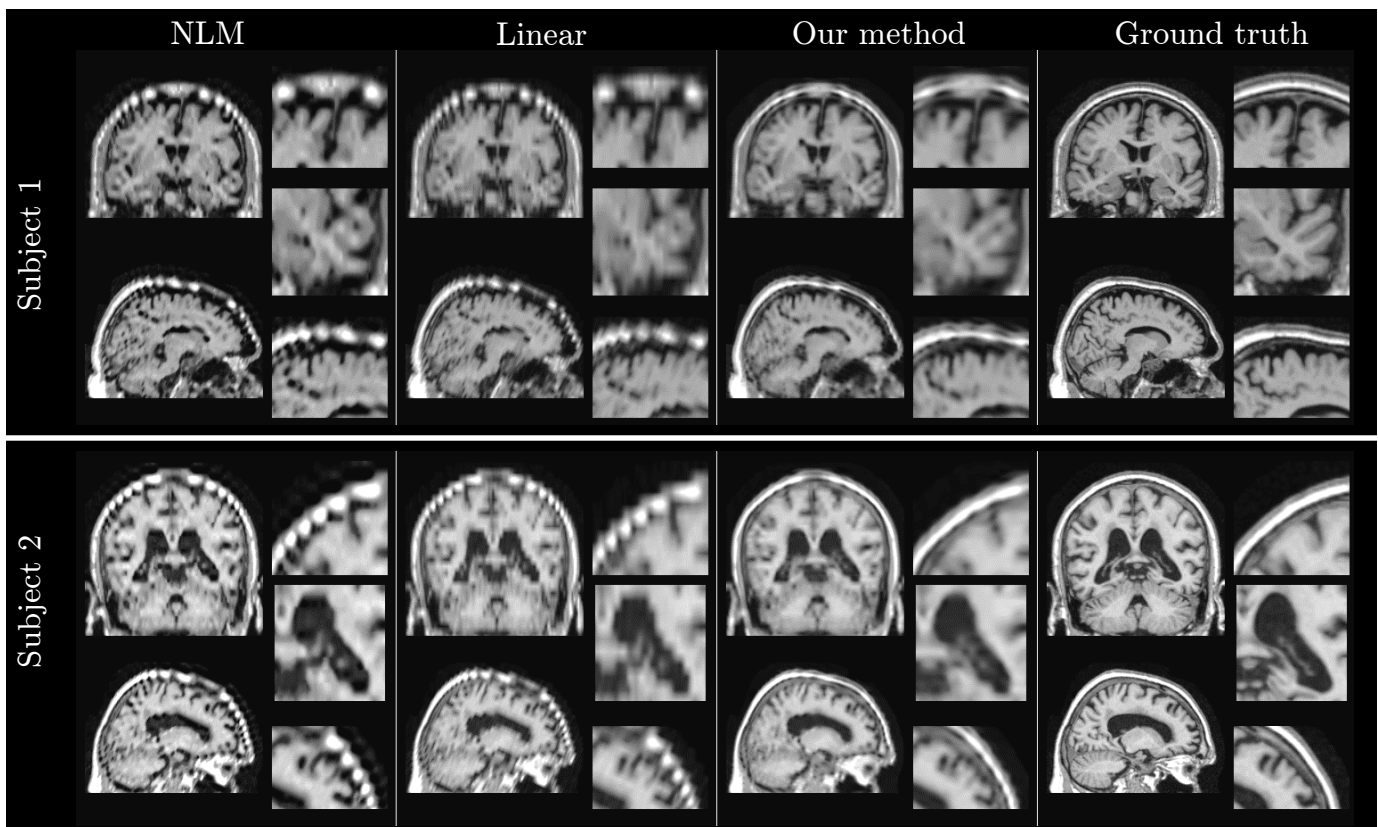


Fig. 4: Representative restorations in the ADNI dataset. Reconstruction by NLM, linear interpolation, and our method, and the original high resolution images for two representative subjects in the study. Our method reconstructs more anatomically plausible substructures as can be especially seen in the close-up panels of the skull, ventricles, and temporal lobe. Additional examples are available in the Supplementary Materials.

III. IMPLEMENTATION

We work in the atlas space, and approximate voxels as either observed or missing in this space by thresholding interpolation weights. To limit interpolation effects due to affine alignment on the results, we set a higher threshold for regions with high image gradients than in regions with low gradients. Parameter estimation could be implemented to include transformation of the model parameters into the subject-specific space in order to optimally use the observed voxels, but this leads to computationally prohibitive updates.

We stack together the affinely registered sparse images from the entire collection. We learn a single set of mixture model parameters within overlapping subvolumes of $21 \times 21 \times 21$ voxels in the isotropically sampled common atlas space. Subvolumes are centered 11 voxels apart in each direction. We use a cubic patch of size $11 \times 11 \times 11$ voxels, and instead of selecting just one patch from each volume at a given location, we collect all overlapping patches within the subvolume centered at that location. This aggregation provides more data for each model, which is crucial when working with severely undersampled volumes. Moreover, including nearby voxels offers robustness in the face of image misalignment. Given the learned parameters at each location, we restore all overlapping patches within a subvolume.

While learning is performed in the common atlas space, we restore each volume in its original image space to limit the effects of interpolation. Specifically, we apply the inverse of the estimated subject-specific affine transformation to the

cluster statistics prior to performing subject-specific inference.

Our implementation is freely available at <https://github.com/adalca/papago>.

IV. EXPERIMENTS

We demonstrate the proposed imputation algorithm on two datasets and evaluate the results both visually and quantitatively. We also include an example of how imputation can aid in a skull stripping task.

A. Data: ADNI dataset

We evaluate our algorithm using 826 T1-weighted brain MR images from ADNI [15]¹. We downsample the isotropic 1mm^3 images to slice separation of 6mm ($1\text{mm} \times 1\text{mm}$ in-plane) in the axial direction to be of comparable quality with the clinical dataset. We use these low resolution images as input. All downsampled scans are affinely registered to a T1 atlas. The original images serve as the ground truth for quantitative comparisons. After learning model parameters using the data set, we evaluate the quality of the resulting imputations.

¹Data used in the preparation of this article were obtained from the Alzheimers Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimers disease (AD).

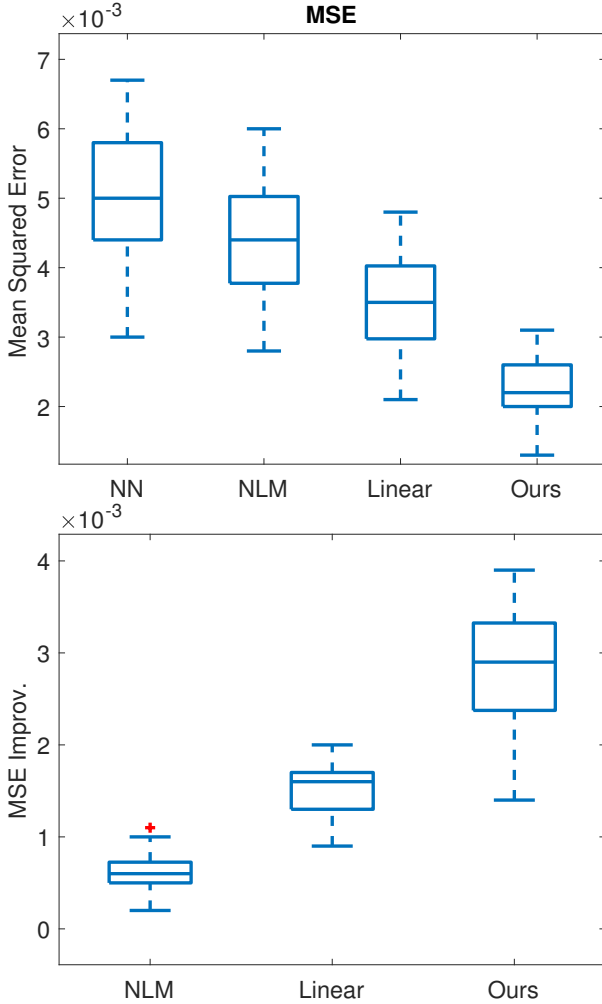


Fig. 5: Reconstruction accuracy statistics. Accuracy for different image restoration methods (top), and improvement over nearest neighbor interpolation using MSE (bottom). All statistics were computed over 50 scans randomly chosen from the ADNI dataset. Image intensities are scaled to a $[0, 1]$ range.

B. Evaluation

We compare our algorithm to three upsampling methods: nearest neighbour (NN) interpolation, non-local means (NLM) upsampling, and linear interpolation [19]. We compare the reconstructed images to the original isotropic volumes both visually and quantitatively. We use the mean squared error,

$$\text{MSE}(Z, Z_o) = \frac{1}{N} \sum \|Z - Z_o\|^2, \quad (17)$$

of the reconstructed image Z relative to the original high resolution scan Z_o . We also compute the related peak signal to noise ratio,

$$\text{PSNR} = \log_{10} \frac{\max(Z_o)}{\text{MSE}(Z, Z_o)}. \quad (18)$$

Both metrics are commonly used in measuring the quality of reconstruction of compressed or noisy signals.

C. Results

Fig. 4 illustrates representative restored images for subjects in the ADNI dataset. Our method produces more plausible

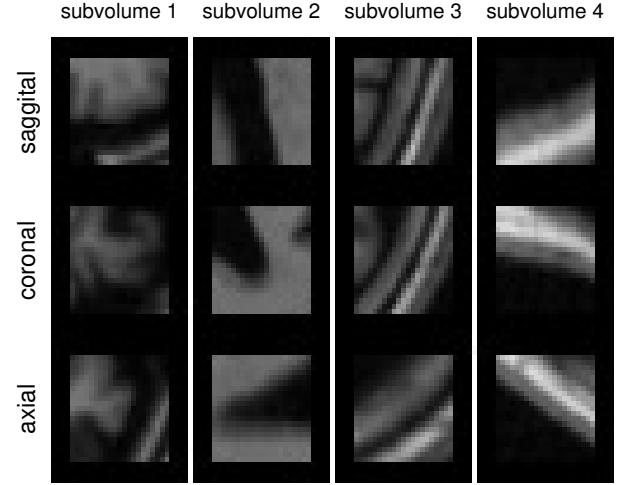


Fig. 6: Regions used for hyper-parameter analysis. Representative example of four subvolumes used in analyses, shown in sagittal, coronal and axial views.

structure. The method restores anatomical structures that are almost entirely missing in the other reconstructions, such as the dura or the sulci of the temporal lobe by learning about these structures from the image collection. We provide additional example results in the Supplementary Materials.

Fig. 5 reports the error statistics in the ADNI data. Due to high variability of MSE among subject scans, we report improvements of each method over the nearest neighbor interpolation baseline in the same scan. Our algorithm offers significant improvement compared to nearest neighbor, NLM, and linear interpolation ($p \leq 10^{-5}, 10^{-42}, 10^{-27}$, respectively). Our method performs significantly better on all subjects. The improvement in MSE is observed in every single scan. Similarly, our method performs consistently better using the PSNR metric (not shown), with mean improvements of up to 1.4 ± 0.44 compared to the next best restored scan.

D. Parameter Setting

We analyze the performance of our algorithm while varying the values of the parameters, and the sparsity patterns of the observed voxels. For these experiments, we use four distinct subvolumes that encompass diverse anatomy from ADNI data, as illustrated in Fig. 6. We start with isotropic data and use different observation masks as described in each experiment.

Hyper-parameters. We evaluate the sensitivity of our method under different hyper parameters: the number of clusters, $k \in [1, 2, 5, 10, 15]$ and the number of dimensions of the low dimensional embedding $d \in [10, 20, 30, 40, 50]$. While different regions give optimal results with different settings, overall our algorithm produces comparable results for the middle range of these parameters. We run all of our experiments with $k = 5$ and $d = 30$.

Sparsity patterns. First, we evaluate how our algorithm performs under three different mask patterns, all of which allow for the same number of observed voxels. Specifically, we (i) use the true sparsely observed planes as in the first experiment; (ii) simulate random rotations of the observation planes mimicking acquisitions in different directions; and (iii)

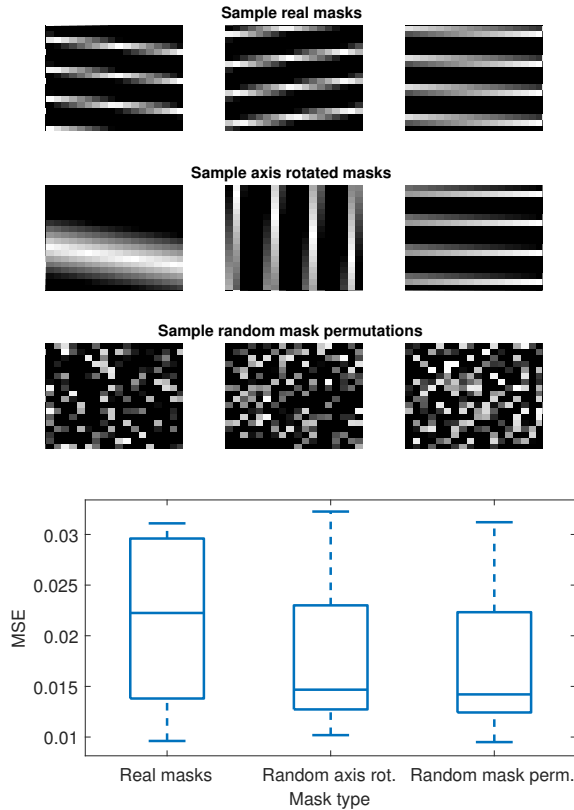


Fig. 7: Mask Analysis. Top: example masks for each simulation, shown in the sagittal plane. The first experiment reflects the limited variability of axial-only acquisitions, whereas the second and third experiments represent increasingly more varied patterns of observed voxels. Bottom: imputation errors. More varied masks leads to improved reconstructions.

simulate random mask patterns. The latter setup is useful for denoising or similar tasks, and is instructive of the performance of our algorithm. Fig. 7 demonstrates that our algorithm performs better under acquisition with different directions, and similarly under truly random observations as more entries of the cluster covariance matrices are directly observed. This demonstrates a promising application of this model to other settings where different patterns of image voxels are observed.

Slice Thickness. We also investigate the effects of slice thickness on the results. The model treats the original data as high resolution planes. Here, we simulate varying slice thickness by blurring isotropic data in the direction perpendicular to the slice acquisition direction. We then use the sampling masks of the scans used in the main experiments to identify observed, albeit blurred, voxels. Fig. 8 shows that although the algorithm performs worse with larger slice thickness, it provides plausible imputation results. For example, results show minimal noticeable differences, even for a blur kernel of $\sigma = 1\text{mm}$, simulating a slice with significant signal contribution from 4mm away. Our method, which treats observed slices as thin, is nevertheless robust to slice thickness variations.

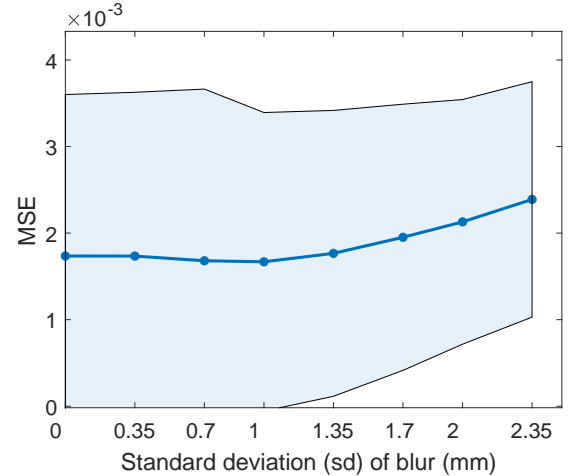
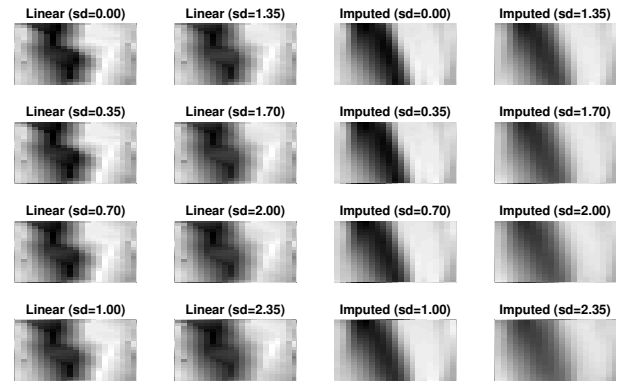


Fig. 8: Slice thickness simulation. Top: sagittal close-up of a region where axial slices were blurred in the direction perpendicular to the acquisition direction; followed by respective imputed results. Bottom: performance of our algorithm under different slice thickness simulations are shown MSE (solid line) and standard deviation interval (shaded region).

E. Skull Stripping

We also illustrate how imputed data might facilitate downstream image analysis. Specifically, the first step in many analysis pipelines is brain extraction – isolating the brain from the rest of the anatomy. Typical algorithms assume that the brain consists of a single connected component separated from the skull and dura by cerebral spinal fluid [29]. Thus, they often fail on sparsely sampled scans that no longer include a clear contrast between these regions. Fig. 9 provides an example where the brain extraction fails on the original subject scan but succeeds on our reconstructed image.

F. Clinical Dataset

We also demonstrate our algorithm on a clinical set of 766 T2-FLAIR brain MR scans in a stroke patient cohort. These scans are severely anisotropic ($0.85 \times 0.85\text{mm}$ in-plane, slice separation of 6mm). All subjects are affinely registered to an atlas and the intensity is normalized.

Fig. 10 illustrates representative restoration improvements in T2-FLAIR scans from a clinical population. Our method produces more plausible structure, as can be especially seen in the close-up panels focusing on anatomical details. We provide additional example results in the Supplementary Materials.

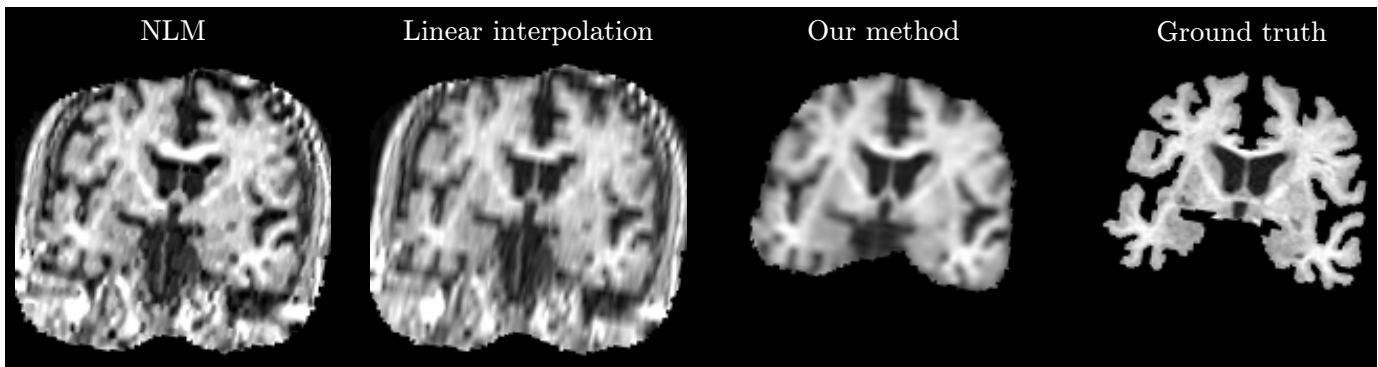


Fig. 9: **Skull Stripping Example.** Example of a skull stripping failure for linear and NLM interpolation. Skull stripping dramatically improves when applied to the imputed image for this example.

V. DISCUSSION

Modeling Choices. We explicitly model and estimate a latent low-dimensional embedding for each patch. The likelihood model (5) does not include the latent patch representation x_i , leading to observed patch covariance $C_k^{\mathcal{O}_i, \mathcal{O}_i} = W_k^{\mathcal{O}_i} (W_k^{\mathcal{O}_i})^T + \sigma_k^2 I$. Since the set of observed voxels \mathcal{O}_i varies across subjects, the resulting Expectation Maximization algorithm [6] becomes intractable if we marginalize the latent representation out before estimation. Introducing the latent structure simplifies the optimization problem.

We investigated an alternative modeling choice that instead treats each missing voxel as a latent variable. In particular we consider the missing values of patch y_i as latent variables, which can be optionally combined with the latent low-dimensional patch representation. These assumptions lead to an Expectation Conditional Maximization (ECM) [14], [18], a variant of the Generalized Expectation Maximization where parameter updates depend on the previous parameter estimates. The resulting algorithm estimates the expected missing voxel mean and covariance directly, and then updates the cluster parameters (see Supplementary Materials for a complete derivation). The most notable difference between this formulation and simpler algorithms that iteratively fill in missing voxels and then estimate GMM model parameters is in the estimation of the expected data covariance, which captures the covariance of the missing and observed data (c.f. [18], Ch.8). We found that compared to the method presented in Section II, this variant often got stuck in local minima, had difficulty moving away from the initial missing voxel estimates, and was an order of magnitude slower than the presented method. We provide both implementations in our code.

Restoration. Our restoration method assumes that the observed voxels are noisy manifestations of the low dimensional patch representation, and reconstructs the entire patch, including the observed voxels, leading to smoother images. This formulation assumes the original observed voxels are noisy observations of the *true* data. Depending on the downstream analyses, the original voxels could be kept in the reconstruction. In addition, we also investigated an alternative reconstruction method of filling in the missing voxels given the observed voxels as *noiseless* ground truth (not shown). This formulation leads to sharper but noisier results. The

two restoration methods therefore yield images with different characteristics. This tradeoff is a function of the noise in the original acquisition: higher noise in the clinical acquisition leads to noisier reconstructions using the alternative method, whereas in the ADNI dataset the two methods perform similarly. In addition, imputation can be achieved by sampling the posterior distribution rather than using conditional mean estimation, enabling a better estimate of the residual noise for downstream analysis.

Usability. Our model assumes that whether a voxel is observed is independent of the intensity of that voxel. Although the voxels missing in the sparsely sampled images clearly form a spatial pattern, we assume there is no correlation with the actual intensity of the voxels. The model can therefore be learned from data with varying sparseness patterns, including restoring data in all acquisition directions simultaneously.

The proposed method can be used for general image imputation using datasets of varying resolution. For example, although acquiring a large high resolution dataset for a clinical study is often infeasible, our algorithm will naturally make use of any additional image data available. Even a small number of acquisitions in different directions or higher resolution than the study scans promise to improve the accuracy of the resulting reconstruction.

The presented model depends on the image collection containing similar anatomical structures roughly aligned, such as affinely aligned brain or cardiac MR scans. Smaller datasets that contain vastly different scans, such as traumatic brain injuries or tumors, may not contain enough consistency to enable the model to learn meaningful covariation. However, a wide range of clinical datasets contain the anatomical consistency required, and can benefit from the proposed method.

Initialization. We experimented with several initialization schemes, and provide them in our implementation. A natural initialization is to first learn a simple GMM from the linearly interpolated volumes, and use the resulting parameter values as initializations for our method. This leads to results that improve on the linear interpolation but still maintain somewhat blocky effects caused by interpolation. More agnostic initializations, such as random parameter values, lead to more realistic anatomy but noisier final estimates. Different methods perform well in different regions of the brain. The experimental results are initialized by first learning a simple GMM

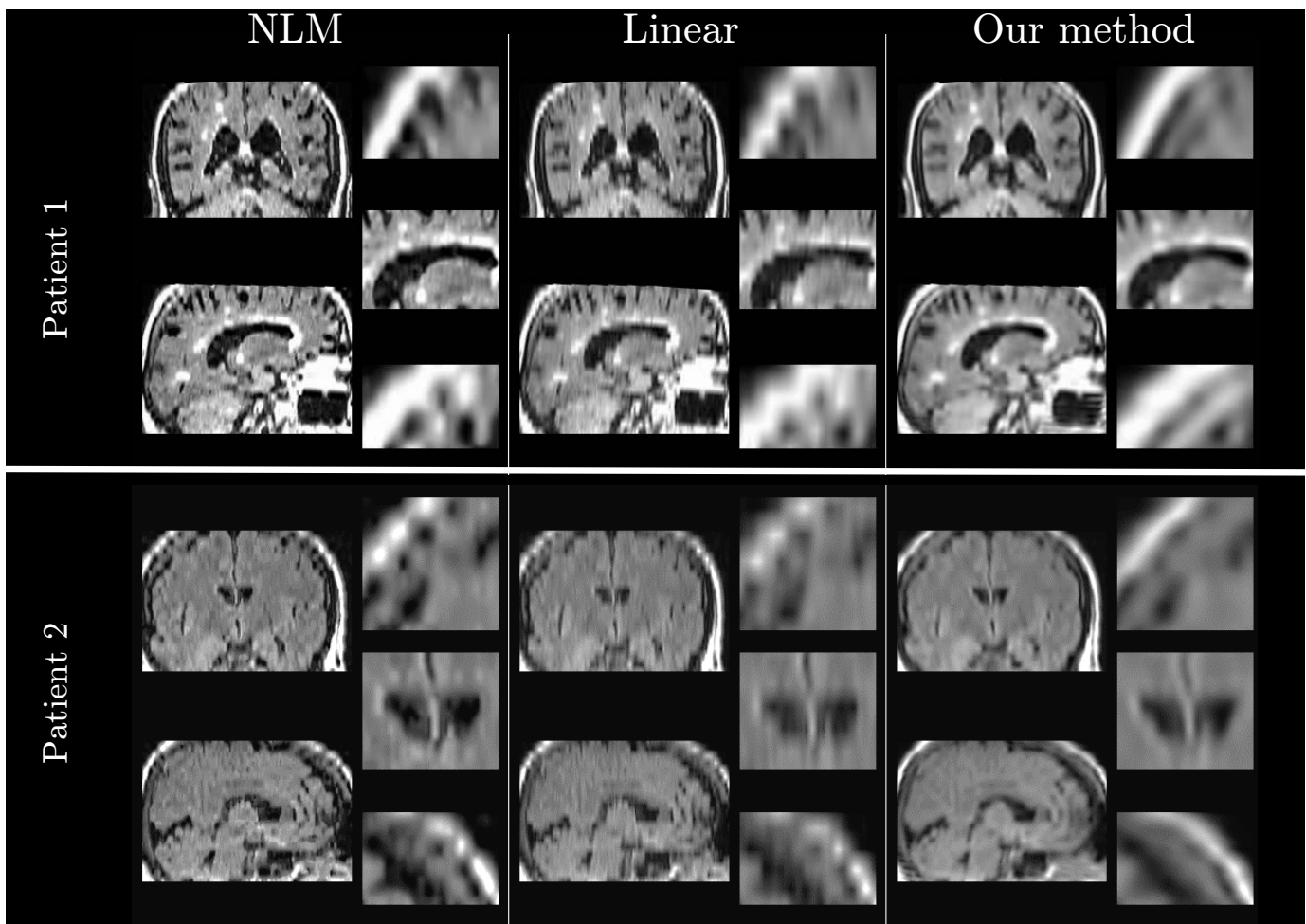


Fig. 10: Representative restorations in the clinical dataset. Reconstruction using NLM, linear interpolation and our method for two representative subjects. Our method reconstructs more plausible substructures, as can be especially seen in the close-up panels of the skull and the periventricular region. Additional examples are available in the Supplementary Materials.

from the linearly interpolated volumes, and using the resulting means with diagonal covariances as an initial setting of the parameters. We start with a low dimensional representation to be of dimension 1, and grow it with every iteration up to the desired dimension. We found that this approach outperforms all other strategies.

VI. CONCLUSIONS

We propose an image imputation method that employs a large collection of low-resolution images to infer fine-scale anatomy of a particular subject. We introduce a model that captures anatomical similarity across subjects in large clinical image collections, and imputes, or fills in, the missing data in low resolution scans. The method produces anatomically plausible volumetric images consistent with sparsely sampled input scans.

Our approach does not require high resolution scans or expert annotations for training. We demonstrate that our algorithm is robust to many data variations, including varying slice thickness. The resulting method enables the use of untapped clinical data for large scale scientific studies and promises to facilitate novel clinical analyses.

ACKNOWLEDGMENT

We acknowledge the following funding sources: NIH NINDS R01NS086905, NIH NICHD U01HD087211, NIH NIBIB NAC P41EB015902, NIH R41AG052246-01, 1K25EB013649-01, 1R21AG050122-01, NSF IIS 1447473, Wistron Corporation, and SIP.

Data collection and sharing for this project was funded by the Alzheimer’s Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from several agencies listed at <http://adni.loni.usc.edu/about/>.

APPENDIX A
EXPECTATION MAXIMIZATION UPDATES

Following (5), the complete likelihood of our model is:

$$p(\mathcal{Y}^\circ, \mathcal{X}; \theta) = \prod_i \sum_k \pi_k \mathcal{N}(y_i^{\circ_i}, x_i; \mu_k^{\circ_i}, C_k^{\circ_i \circ_i}), \quad (19)$$

where $\mathcal{X} = \{x_i\}$. The expectation of this probability is then

$$\begin{aligned} Q(\theta|\theta) &= \mathbf{E}_{\mathcal{X}|\mathcal{Y}^\circ, \theta} [\log p(\mathcal{Y}^\circ, \mathcal{X}; \theta)] \\ &= \sum_{i,k} \mathbf{E} [k(-\frac{d}{2} \log 2\pi - \frac{1}{2} \log |\sigma_k^2 I| \\ &\quad - \frac{1}{2\sigma_k^2} (y_i^{\circ_i} - W_k^{\circ_i} x_i - \mu_k^{\circ_i})^T (y_i^{\circ_i} - W_k^{\circ_i} x_i - \mu_k^{\circ_i}) \\ &\quad - \frac{d}{2} \log 2\pi - \frac{1}{2} \log |I| - \frac{1}{2} x_i x_i^T)]. \end{aligned} \quad (20)$$

Computing this expectation requires evaluating $\mathbf{E}[k]$, $\mathbf{E}[x_i|k]$, and $\mathbf{E}[x_i x_i^T|k]$, which is trivially done to obtain the expectation step updates (6) -(8) .

For the maximization step, we optimize (20) with respect to the model parameters.

$$\begin{aligned} \frac{\partial Q}{\partial \mu_k} &\propto \sum_{i \in \mathcal{P}_j} \frac{\partial}{\partial \mu_k} \mathbf{E} [k(y_i^{\circ_i} - W_k^{\circ_i} x_i - \mu_k^{\circ_i})^T (y_i^{\circ_i} - W_k^{\circ_i} x_i - \mu_k^{\circ_i})] \\ &\propto \sum_{i \in \mathcal{P}_j} \mathbf{E} [k(y_i^{\circ_i} - W_k^{\circ_i} x_i - \mu_k^{\circ_i})] \\ &= \sum_{i \in \mathcal{P}_j} \gamma_{ik} (y_i^{\circ_i} - W_k^{\circ_i} \hat{x}_{ik} - \mu_k^{\circ_i}) = 0, \end{aligned}$$

$$\mu_k^j = \frac{1}{\sum_{i'} \gamma_{i'k}} \sum_{i \in \mathcal{P}_j} \gamma_{ik} (y_i^j - W_k^j \hat{x}_{ik}) \quad (21)$$

$$\begin{aligned} \frac{\partial Q}{\partial W_k} &\propto \sum_{i \in \mathcal{P}_j} \frac{\partial \mathbf{E} [k(y_i^{\circ_i} - W_k^{\circ_i} x_i - \mu_k^{\circ_i})^T (y_i^{\circ_i} - W_k^{\circ_i} x_i - \mu_k^{\circ_i})]}{\partial W_k} \\ &\propto \sum_{i \in \mathcal{P}_j} \mathbf{E} [k(W_k^{\circ_i} x_i - (y_i^{\circ_i} - \mu_k^{\circ_i})) x_i^T] \\ &= \sum_{i \in \mathcal{P}_j} \gamma_{ik} W_k^{\circ_i} (\hat{x}_{ik} \hat{x}_{ik}^T + S_{ik}) - (y_i^{\circ_i} - (\mu_k^{\circ_i})) \hat{x}_{ik}^T = 0 \\ W_k^j &= \left[\sum_{i \in \mathcal{P}_j} \gamma_{ik} (\hat{x}_{ik} \hat{x}_{ik}^T + S_{ik}) \right]^{-1} \sum_{i \in \mathcal{P}_j} \gamma_{ik} (y_i^j - \mu_k^j) \hat{x}_{ik}^T \\ &= \sum_{i \in \mathcal{P}_j} \delta_{ik} (y_i^j - \mu_k^j) \hat{x}_{ik}^T A_j^{-1} \end{aligned} \quad (22)$$

where δ and A are defined in (9) and (11), respectively. By combining (22) and (21), we obtain

$$\begin{aligned} \mu_k^j \sum_i \gamma_{ik} &= \sum_{i \in \mathcal{P}_j} \gamma_{ik} y_i^j - W_k^j \sum_{i \in \mathcal{P}_j} \gamma_{ik} \hat{x}_{ik} \\ \mu_k^j &= \sum_{i \in \mathcal{P}_j} \delta_{ik} y_i^j - W_k^j b_j \\ \mu_k^j &= \sum_{i \in \mathcal{P}_j} \delta_{ik} y_i^j - \sum_{i \in \mathcal{P}_j} \delta_{ik} (y_i^j - \mu_k^j) \hat{x}_{ik}^T A_j^{-1} b_j \end{aligned}$$

$$\begin{aligned} \mu_k^j (1 - \sum_{i \in \mathcal{P}_j} \delta_{ik} \hat{x}_{ik}^T A_j^{-1} b_j) &= \sum_{i \in \mathcal{P}_j} \delta_{ik} y_i^j (1 - \hat{x}_{ik}^T A_j^{-1} b_j) \\ \mu_k^j &= \frac{\sum_{i \in \mathcal{P}_j} \delta_{ik} y_i^j (1 - \hat{x}_{ik}^T A_j^{-1} b_j)}{\sum_{i \in \mathcal{P}_j} \delta_{ik} (1 - \hat{x}_{ik}^T A_j^{-1} b_j)}. \end{aligned} \quad (23)$$

We therefore update μ_k^j via (23), followed by W_k^j using (22). Finally,

$$\begin{aligned} \frac{\partial Q}{\partial \sigma_k^2} &\propto \sum_{i \in \mathcal{P}_j} -\frac{\partial}{\partial \sigma_k^2} \frac{1}{\sigma_k^2} \\ &\quad \mathbf{E} [k(y_i^{\circ_i} - W_k^{\circ_i} x_i - \mu_k^{\circ_i})^T (y_i^{\circ_i} - W_k^{\circ_i} x_i - \mu_k^{\circ_i})] \\ &\quad - \frac{\partial}{\partial \sigma_k^2} N \log \sigma_k^2 \\ &= \sum_{i \in \mathcal{P}_j} \frac{1}{\sigma_k^4} \mathbf{E} [k(y_i^{\circ_i} - W_k^{\circ_i} x_i - \mu_k^{\circ_i})^T (y_i^{\circ_i} - W_k^{\circ_i} x_i - \mu_k^{\circ_i})] \\ &\quad - \frac{N}{\sigma_k^2} = 0 \\ \sigma_k^2 &= \sum_{i \in \mathcal{P}_j} \mathbf{E} [k(y_i^{\circ_i} - W_k^{\circ_i} x_i - \mu_k^{\circ_i})^T (y_i^{\circ_i} - W_k^{\circ_i} x_i - \mu_k^{\circ_i})] \\ \sigma_k^2 &= \frac{\sum_j \sum_{i \in \mathcal{P}_j} \gamma_{ik} [(y_i^j - \mu_k^j - W_k^j \hat{x}_{ik})^2 + W_k^j S_{ik} (W_k^j)^T]}{\sum_j \sum_{i \in \mathcal{P}_j} \gamma_{ik}}. \end{aligned} \quad (24)$$

REFERENCES

- [1] Komal Kumar Bhatia, Akhila Rao, Anthony N Price, Robin Wolz, Joseph V Hajnal, and Daniel Rueckert. Hierarchical manifold learning for regional image analysis. *TMI*, 33(2):444–461, 2014.
- [2] Eyal Carmi, Siuyan Liu, Noga Alon, Amos Fiat, and Daniel Fiat. Resolution enhancement in MRI. *Magnetic resonance imaging*, 24(2):133–154, 2006.
- [3] Pierrick Coupé, José V Manjón, Vladimir Fonov, Jens Pruessner, Montserrat Robles, and Louis D Collins. Patch-based segmentation using expert priors: Application to hippocampus and ventricle segmentation. *NeuroImage*, 54(2):940–954, 2011.
- [4] Adrian V Dalca, Andreea Bobu, Natalia S Rost, and Polina Golland. Patch-based discrete registration of clinical brain images. In *International Workshop on Patch-based Techniques in Medical Imaging*, pages 60–67. Springer, 2016.
- [5] Adrian V Dalca, Katherine L Bouman, William T Freeman, Natalia S Rost, Mert R Sabuncu, and Polina Golland. Population based image imputation. In *Information Processing in Medical Imaging*. Springer, 2017.
- [6] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological)*, 39(1):1–38, 1977.
- [7] Adriana Di Martino, Chao-Gan Yan, Qingyang Li, Erin Denio, Francisco X Castellanos, Kaat Alaerts, et al. The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. *Molecular psychiatry*, 19(6):659–667, 2014.
- [8] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *European Conference on Computer Vision*, pages 184–199. Springer, 2014.
- [9] Michael Elad and Michal Aharon. Image denoising via sparse and redundant representations over learned dictionaries. *IEEE TMI*, 15(12):3736–3745, 2006.
- [10] William T Freeman, Thouis R Jones, and Egon C Pasztor. Example-based super-resolution. *IEEE Computer graphics and Applications*, 22(2):56–65, 2002.
- [11] Daniel Glasner, Shai Bagon, and Michal Irani. Super-resolution from a single image. In *Computer Vision, International Conference on*, pages 349–356. IEEE, 2009.
- [12] Derek LG Hill, Philipp G Batchelor, Mark Holden, and David J Hawkes. Medical image registration. *Physics in medicine and biology*, 46(3):R1, 2001.

- [13] Juan E Iglesias, Ender Konukoglu, Darko Zikic, Ben Glocker, Koen Van Leemput, and Bruce Fischl. Is synthesizing mri contrast useful for inter-modality analysis? *MICCAI*, LNCS 8149:631–638, 2013.
- [14] Alexander Ilin and Tapani Raiko. Practical approaches to principal component analysis in the presence of missing values. *J Mach Learn Res*, 11:1957–2000, 2010.
- [15] Clifford R Jack, Matt A Bernstein, Nick C Fox, Paul Thompson, Gene Alexander, Danielle Harvey, et al. The Alzheimer’s disease neuroimaging initiative (ADNI): MRI methods. *Journal of Magnetic Resonance Imaging*, 27(4):685–691, 2008.
- [16] Amod Jog, Aaron Carass, and Jerry L Prince. Improving magnetic resonance resolution with supervised learning. In *ISBI*, pages 987–990. IEEE, 2014.
- [17] Ender Konukoglu, Andre van der Kouwe, Mert R Sabuncu, and Bruce Fischl. Example-based restoration of high-resolution magnetic resonance image acquisitions. *MICCAI*, LNCS 8149:131–138, 2013.
- [18] Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*. Wiley, 2014.
- [19] José V Manjón, Pierrick Coupé, Antonio Buades, Vladimir Fonov, D Louis Collins, and Montserrat Robles. Non-local MRI upsampling. *Medical image analysis*, 14(6):784–792, 2010.
- [20] José V Manjón, Pierrick Coupé, Antonio Buades, D Louis Collins, and Montserrat Robles. New methods for MRI denoising based on sparseness and self-similarity. *Med. I.A.*, 16(1):18–27, 2012.
- [21] Ozan Oktay, Wenjia Bai, Matthew Lee, Ricardo Guerrero, Konstantinos Kamnitsas, Jose Caballero, Antonio de Marvao, Stuart Cook, Declan O’Regan, and Daniel Rueckert. Multi-input cardiac image super-resolution using convolutional neural networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 246–254. Springer, 2016.
- [22] Esben Plenge, Dirk HJ Poot, Wiro J Niessen, and Erik Meijering. Super-resolution reconstruction using cross-scale self-similarity in multi-slice MRI. *MICCAI: Medical Image Computing and Computer-Assisted Intervention*, LNCS 8151:123–130, 2013.
- [23] Natalia S Rost, Kaitlin Fitzpatrick, Alessandro Biffi, Allison Kanakis, William Devan, Christopher D Anderson, Lynelle Cortellini, Karen L Furie, and Jonathan Rosand. White matter hyperintensity burden and susceptibility to cerebral ischemia. *Stroke*, 41(12):2807–2811, 2010.
- [24] Francois Rousseau, Piotr A Habas, and Colin Studholme. A supervised patch-based approach for human brain labeling. *IEEE Tran. Med. Imag.*, 30(10):1852–1862, 2011.
- [25] François Rousseau, Alzheimers Disease Neuroimaging Initiative, et al. A non-local approach for image super-resolution using intermodality priors. *Medical image analysis*, 14(4):594–605, 2010.
- [26] François Rousseau, Kio Kim, and Colin Studholme. A groupwise super-resolution approach: application to brain MRI. In *ISBI*, pages 860–863. IEEE, 2010.
- [27] Paul Schmidt, Christian Gaser, Milan Arsic, Dorothea Buck, Annette Förschler, Achim Berthele, Muna Hoshi, Rüdiger Ilg, Volker J Schmid, Claus Zimmer, et al. An automated tool for detection of flair-hyperintense white-matter lesions in multiple sclerosis. *Neuroimage*, 59(4):3774–3783, 2012.
- [28] Isaac Jacob Schoenberg. *Cardinal spline interpolation*. 12. SIAM, 1973.
- [29] Florent Ségonne, Anders M Dale, Evelina Busa, Maureen Glessner, David Salat, Horst K Hahn, and Bruce Fischl. A hybrid approach to the skull stripping problem in MRI. *Neuroimage*, 22(3):1060–1075, 2004.
- [30] Wenzhe Shi, Jose Caballero, Christian Ledig, Xiahai Zhuang, Wenjia Bai, et al. Cardiac image super-resolution with global correspondence using multi-atlas patchmatch. *MICCAI: Medical Image Computing and Computer-Assisted Intervention*, LNCS 8151:9–16, 2013.
- [31] Ramesh Sridharan, Adrian V Dalca, Kaitlin M Fitzpatrick, Lisa Cloonan, Allison Kanakis, Ona Wu, et al. Quantification and analysis of large multimodal clinical image studies: Application to stroke. *MICCAI - MBIA Workshop*, LNCS 8159:18–30, 2013.
- [32] Koen Van Leemput, Frederik Maes, Dirk Vandermeulen, Alan Colchester, and Paul Suetens. Automated segmentation of multiple sclerosis lesions by model outlier detection. *IEEE transactions on medical imaging*, 20(8):677–688, 2001.
- [33] Jianchao Yang, John Wright, Thomas S Huang, and Yi Ma. Image super-resolution via sparse representation. *IEEE Transactions on Image Processing*, 19(11):2861–2873, 2010.
- [34] Ruoqiao Zhang, Charles A Bouman, Jean-Baptiste Thibault, and Ken D Sauer. Gaussian mixture markov random field for image denoising and reconstruction. In *IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pages 1089–1092. IEEE, 2013.
- [35] Daniel Zoran and Yair Weiss. From learning models of natural image patches to whole image restoration. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 479–486. IEEE, 2011.
- [36] Daniel Zoran and Yair Weiss. Natural images, gaussian mixtures and dead leaves. In *Advances in Neural Information Processing Systems*, pages 1736–1744, 2012.

SUPPLEMENTARY MATERIAL

DERIVATION OF ALTERNATIVE MODEL

In this section, we explore the parameter estimation for an alternative model. Specifically, letting \mathcal{M}_i be the set of missing voxels of patch y_i , we treat $y_i^{\mathcal{M}_i}$ as latent variables, instead of explicitly modeling a low-dimensional representation x . We show the maximum likelihood updates of the model parameters under the likelihood (5). We employ the Expectation Conditional Maximization (ECM) [14], [18] variant of the Generalized Expectation Maximization, where parameter updates depend on the previous parameter estimates.

The complete data likelihood is

$$p(\mathcal{Y}; \theta) = \prod_i \sum_k \pi_k \mathcal{N}(y_i^{\circ_i}, y_i^{\mathcal{M}_i}; \mu_k^{\circ_i}, \Sigma_k^{\circ_i \circ_i}). \quad (25)$$

The **expectation step** updates the statistics of the missing data, computed based on covariates of the known and unknown voxels:

$$\begin{aligned} \gamma_{ik} &\equiv \mathbf{E}[k_i] \\ &= \frac{\pi_k \mathcal{N}(y_i^{\circ_i}; \mu_k^{\circ_i}, \Sigma_k^{\circ_i})}{\sum_k \pi_k \mathcal{N}(y_i^{\circ_i}; \mu_k^{\circ_i}, \Sigma_k^{\circ_i})} \end{aligned} \quad (26)$$

$$\begin{aligned} \hat{y}_{ij} &\equiv \mathbf{E}[y_{ij}] \\ &= \begin{cases} y_{ij} & \text{if } y_{ij} \\ & \text{is observed} \\ \mu_{ij} + \Sigma_i^{j \circ_i} (\Sigma_i^{\circ_i \circ_i})^{-1} (y_i^{\circ_i} - \mu^{\circ_i}) & \text{otherwise} \end{cases} \end{aligned} \quad (27)$$

$$\begin{aligned} \hat{s}_{ijl} &\equiv \mathbf{E}[y_{ij} y_{il}] - \mathbf{E}[y_{ij}] \mathbf{E}[y_{il}] \\ &= \begin{cases} 0 & \text{if } y_{ij} \text{ or } y_{il} \\ & \text{is observed} \\ \Sigma_i^{jl} - (\Sigma_i^{\circ_i j})^T (\Sigma_i^{\circ_i \circ_i})^{-1} \Sigma_i^{\circ_i l} & \text{otherwise} \end{cases} \end{aligned} \quad (28)$$

where the correction in \hat{s}_{ijl} can be interpreted as the uncertainty in the covariance estimation due to the missing values.

Given estimates for the missing data, the **maximization step** leads to familiar Gaussian Mixture Model parameters updates:

$$\mu_k = \frac{1}{\gamma_{ik}} \sum_i \gamma_{ik} \hat{y}_{ik} \quad (29)$$

$$\Sigma_k = \frac{1}{\gamma_{ik}} \sum_i \gamma_{ik} [(\hat{y}_{ik} - \mu_k)(\hat{y}_{ik} - \mu_k)^T + S_i^T]. \quad (30)$$

$$\pi_k = \frac{1}{N} \sum_i \gamma_{ik} \quad (31)$$

where $[S_i]_{jl} = \hat{s}_{ijl}$.

In addition to the latent missing voxels, we can still model each patch as coming from a low dimensional representation. We form $C_k = W_k W_k^T + \sigma_k^2 I$ as in (3), leading to the complete data likelihood:

$$p(\mathcal{Y}; \theta) = \prod_i \sum_k \pi_k \mathcal{N}(y_i^{\circ_i}, y_i^{\mathcal{M}_i}; \mu_k^{\circ_i}, C_k^{\circ_i \circ_i}). \quad (32)$$

The **expectation steps** are then unchanged from (26)-(28) with C_k replacing Σ_k . The **maximization steps** are unchanged from (29)-(31), with Σ_k now the *empirical* covariance in (30). We let $U \Lambda V^T = \text{SVD}(\Sigma_k)$ be the singular value decomposition of Σ_k , leading to the low dimensional updates

$$\sigma_k^2 \leftarrow \frac{1}{d-q} \sum_{j=d+1}^d \Lambda(j, j) \quad (33)$$

$$W_k \leftarrow U (\Lambda - \sigma_k^2 I)^{1/2}. \quad (34)$$

Finally, we let $C_k = W_k W_k^T + \sigma_k^2 I$.

Unfortunately, both learning procedures involve estimating all of the missing voxel covariances, leading to a large and unstable optimization.

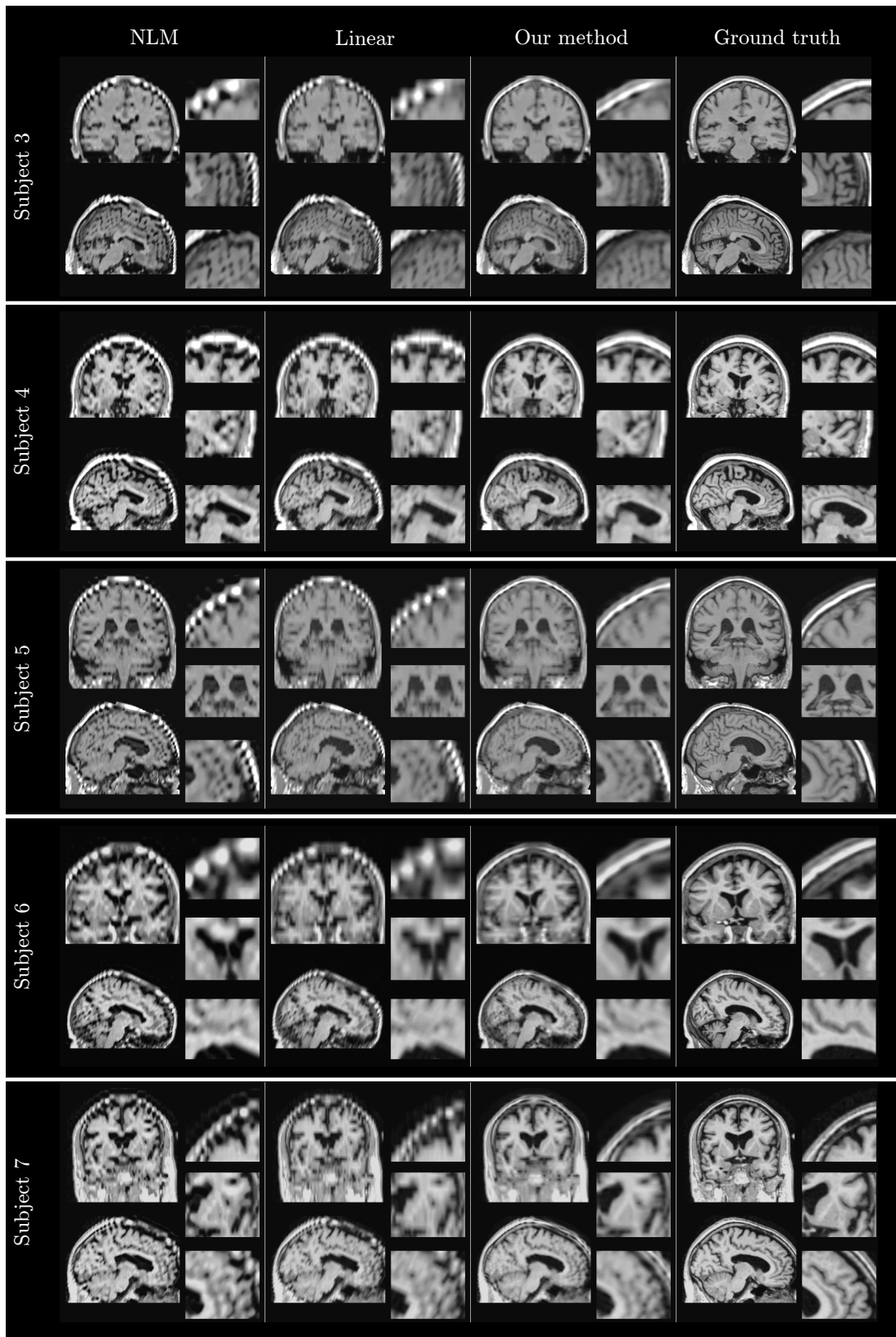


Fig. 11: Additional restorations in the ADNI dataset. Reconstruction by NLM, linear interpolation, and our method, and the original high resolution images.

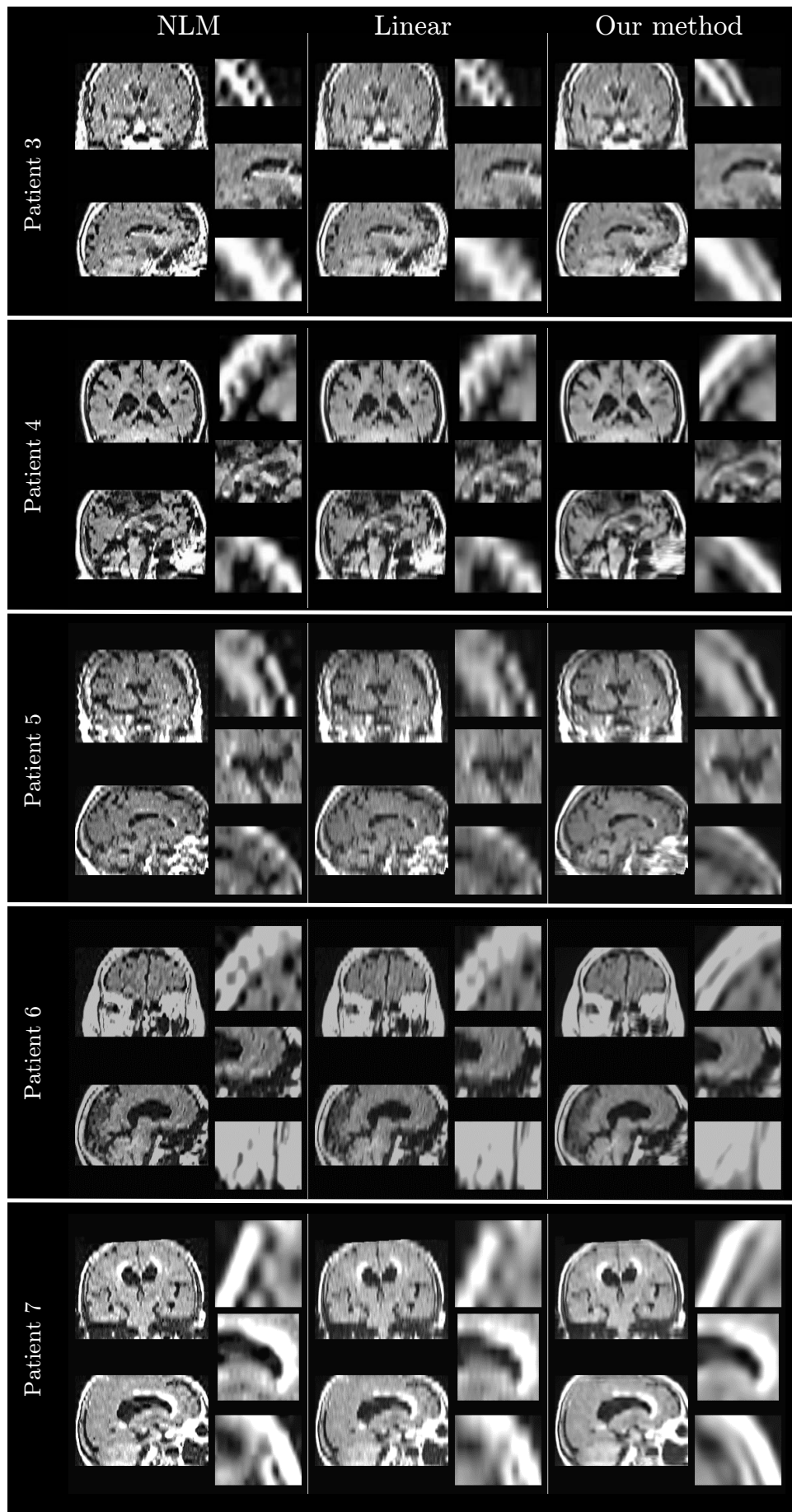


Fig. 12: Additional restorations in the clinical dataset. Reconstruction using NLM, linear interpolation and our method.