LETTER

doi:10.1038/nature25988

Image reconstruction by domain-transform manifold learning

Bo Zhu[1,2,3], Jeremiah Z. Liu[4], Stephen F. Cauley[1,2], Bruce R. Rosen[1,2] & Matthew S. Rosen[1,2,3]

Image reconstruction is essential for imaging applications across the physical and life sciences, including optical and radar systems, magnetic resonance imaging, X-ray computed tomography, positron emission tomography, ultrasound imaging and radio astronomy[1–3]. During image acquisition, the sensor encodes an intermediate representation of an object in the sensor domain, which is subsequently reconstructed into an image by an inversion of the encoding function. Image reconstruction is challenging because analytic knowledge of the exact inverse transform may not exist *a priori*, especially in the presence of sensor non-idealities and noise. Thus, the standard reconstruction approach involves approximating the inverse function with multiple *ad hoc* stages in a signal processing chain[4,5], the composition of which depends on the details of each acquisition strategy, and often requires expert parameter tuning to optimize reconstruction performance. Here we present a unified framework for image reconstruction—automated transform by manifold approximation (AUTOMAP)—which recasts image reconstruction as a data-driven supervised learning task that allows a mapping between the sensor and the image domain to emerge from an appropriate corpus of training data. We implement AUTOMAP with a deep neural network and exhibit its flexibility in learning reconstruction transforms for various magnetic resonance imaging acquisition strategies, using the same network architecture and hyperparameters. We further demonstrate that manifold learning during training results in sparse representations of domain transforms along low-dimensional data manifolds, and observe superior immunity to noise and a reduction in reconstruction artefacts compared with conventional handcrafted reconstruction methods. In addition to improving the reconstruction performance of existing acquisition methodologies, we anticipate that AUTOMAP and other learned reconstruction approaches will accelerate the development of new acquisition strategies across imaging modalities.

The paradigm shift from manual to automatic feature extraction in a host of machine learning tasks including speech recognition[6] and image classification[7] has demonstrated the advantage of allowing real-world data to guide efficient representation through a structured training process. This strategy is mirrored in biological organisms for refining visual perception in a process known as perceptual learning[8]. Human visual reconstruction of time-domain neural codes into the percept image is trained through experience during cognitive development into adulthood. This conditioning on prior data has been shown to be critical to robust performance in low signal-to-noise settings[9], which are fundamentally challenging for artificial imaging systems across disciplines and applications. In contemporary medical imaging, faithful reconstruction of noisy image acquisitions is of particular importance as the clinical push for faster scanning increasingly relies on acquisition strategies that result in a reduction of the signal-to-noise ratio, be they undersampled magnetic resonance imaging (MRI), or low-dose X-ray computed tomography imaging.

Inspired by the perceptual learning archetype, we describe here a data-driven unified image reconstruction approach, which we call AUTOMAP, that learns a reconstruction mapping between the sensor-domain data and image-domain output (Fig. 1a). As this mapping is trained, a low-dimensional joint manifold of the data in both domains is implicitly learned (Fig. 1b), capturing a highly expressive representation that is robust to noise and other input perturbations.

We implemented the AUTOMAP unified reconstruction framework with a deep neural network feed-forward architecture composed of fully connected layers followed by a sparse convolutional autoencoder (Fig. 1c). The fully connected layers approximate the between-manifold projection from the sensor domain to the image domain. The convolutional layers extract high-level features from the data and force the image to be represented sparsely in the convolutional-feature space. Our network operates similarly to the denoising autoencoder described previously[10], but rather than finding an efficient representation of the identity to map $f(x) = \phi_x \circ \phi_x^{-1}(x) = x$ over the manifold of inputs $\mathcal{X}$ (where ϕ_x maps the intrinsic coordinate system of $\mathcal{X}$ to Euclidean space near x), AUTOMAP determines both a between-manifold projection g from $\mathcal{X}$ (the manifold of sensor inputs) to $\mathcal{Y}$ (the manifold of output images), and a manifold mapping ϕ_y to project the image from manifold $\mathcal{Y}$ back to the representation in Euclidean space. A composite inverse transformation $f(x) = \phi_y \circ g \circ \phi_x^{-1}(x)$ over the joint manifold $\mathcal{M}_{\mathcal{X},\mathcal{Y}} = \mathcal{X} \times \mathcal{Y}$ (Fig. 1b) is achieved. A full mathematical description of this manifold learning process is detailed in Methods.

In contrast to previous efforts that use neural networks to solve inverse functions[11–13], our approach searches for an inverse that best represents the data in a low-dimensional feature space determined by manifold learning as well as the trained sparse convolutional filters. Furthermore, AUTOMAP solves a generalized reconstruction problem and thus differs from work using neural networks to implement a specific image reconstruction task[14–17]. These previous approaches use known properties of the canonical domain transform to formulate the neural network model, or perform the explicit transform before processing by a neural network used for image-space artefact reduction.

We demonstrate AUTOMAP image reconstruction using MRI as a model system, but we emphasize that our approach is applicable to image reconstruction problems across a broad range of modalities given the mathematical similarities of tomographic spatial encoding functions typically governed by Fredholm integral equations[1]. The plethora of MRI acquisition strategies makes it a particularly appropriate platform to exhibit the flexibility of AUTOMAP reconstruction over a variety of encoding schemes. We first evaluated the performance of AUTOMAP alongside conventional methods in four nontrivial reconstruction tasks: (1) Radon projection imaging and model-based iterative reconstruction[4]; (2) spiral-trajectory k-space (rapid acquisition with non-Cartesian sampling) and conjugate-gradient sensitivity encoding (SENSE) reconstruction employing non-uniform fast Fourier transform (NUFFT) regridding[5]; (3) Poisson-disk undersampled k-space (incoherent sparse acquisition) and compressed sensing

[1]A. A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Boston, Massachusetts, USA. [2]Harvard Medical School, Boston, Massachusetts, USA. [3]Department of Physics, Harvard University, Cambridge, Massachusetts, USA.. [4]Department of Biostatistics, Harvard University, Cambridge, Massachusetts, USA.

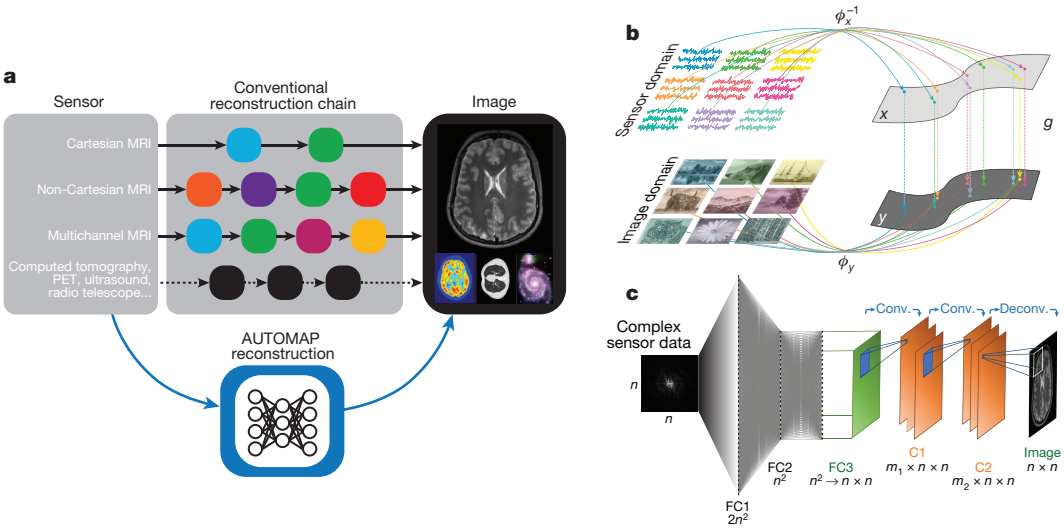


Figure 1 | Schematic representations of AUTOMAP image reconstruction. a, Conventional image reconstruction is implemented with sequential modular reconstruction chains composed of handcrafted signal processing stages that may include discrete transforms (for example, Fourier, Hilbert or Radon), data interpolation techniques, nonlinear optimization, and various filtering mechanisms. AUTOMAP replaces this approach with a unified image reconstruction framework that learns the reconstruction relationship between sensor and image domain without expert knowledge. **b**, A mapping between sensor domain and image domain is determined via supervised learning of sensor (top) and image (bottom) domain pairs. The training process implicitly learns a low-dimensional joint manifold $\mathcal{X} \times \mathcal{Y}$ over which the reconstruction function $f(x) = \phi_y \circ g \circ \phi_x^{-1}(x)$ is conditioned. **c**, AUTOMAP is implemented with a deep neural network architecture composed of fully connected layers (FC1 to FC3) with hyperbolic tangent activations, followed by convolutional layers with rectifier nonlinearity activations that form a convolutional autoencoder. Our network contains m_1 and m_2 convolutional feature maps at C1 and C2 respectively. The convolution and deconvolution operations are labelled 'conv.' and 'deconv.', respectively. The dimensionality of the input to the network is $n \times n$. See Methods for model architecture details.

reconstruction with wavelet sparsifying transform[18]; and (4) misaligned *k*-space (a commonplace sampling inaccuracy due to hardware limitations or physiologic effects) and the conventional inverse fast Fourier transform. Evaluation of the AUTOMAP network was performed on brain magnetic resonance images selected from the Human Connectome Project (HCP)[19], which were transformed to the sensor domain according to the four encoding schemes (see Methods for data preparation details) and with varying levels of additive white Gaussian noise introduced so that we could observe reconstruction performance in noisy conditions.

All reconstruction tasks employed the same network architecture and hyperparameters—only the training data differed at the network input and output. To demonstrate AUTOMAP's generalizability in training dataset scope, all reconstruction tasks except the undersampled encoding were trained from datasets derived entirely from photographs of natural scenes from ImageNet[20] as schematically portrayed in Fig. 1b; for these acquisitions, the network was not exposed to any MRI or other medical images until the test phase (see Methods for data preparation and training details).

The results shown in Fig. 2 demonstrate the ability of AUTOMAP to reconstruct sensor-domain data across varying encoding acquisition strategies. We emphasize here that the reconstruction transforms emerged strictly from training on data samples, without higher-level knowledge (for example, mathematical transforms or domain representations) introduced at any stage. To learn a new reconstruction for a particular encoding acquisition, one simply needs to generate a training dataset with the encoding forward model. The ability of AUTOMAP to represent a variety of sophisticated transform functions with a single network architecture is grounded in the inherent universal approximation properties of nonlinear multilayer perceptron systems[21].

Furthermore, AUTOMAP reconstructions exhibit superior noise immunity compared to those from conventional methods, as quantified by image signal-to-noise ratio and root-mean-squared error (RMSE) metrics (Fig. 2). Visual inspection of reconstructed images and error maps in Fig. 2 reveals that noise and reconstruction artefacts are diminished in AUTOMAP reconstructions compared to conventional reconstructions: streaking artefacts and white noise amplification for iterative inverse-Radon[22], noise amplification due to iterative reconstruction with NUFFT regridding of noisy samples[23], structured artefacts from noisy undersampled compressed sensing reconstruction[24], and Nyquist N/2 ghosting from misaligned sampling trajectories[25]. Additive Gaussian noise was not injected during training; the noise immunity we observe was not trained explicitly, or imposed by predictive noise modelling, but rather emerged as a result of the manifold learning process extracting robust features of the data, leading to improvement in signal-to-noise ratio during reconstruction. This emphasis on modelling features of the signal rather than the noise characteristics to achieve high performance in low-signal-to-noise-ratio regimes is consistent with the neural mechanisms underlying human visual perceptual learning[26].

We next examined the hidden-layer activity of our AUTOMAP network during the feed-forward reconstruction process. We trained AUTOMAP using training data derived from either ImageNet, HCP brain images, or random-valued Gaussian noise without any real-world image structure. Each trained network was then used to reconstruct the fully sampled Cartesian *k*-space of a single brain image (Extended Data Fig. 2). The activation values of the hidden-layer FC2 (Fig. 1c) are plotted in Fig. 3a–c. As the training moves from general (Fig. 3a) to specific (Fig. 3c), we observe the hidden-layer activity exhibiting greater sparsity, indicating successful extraction of robust features[27], consistent with the noise immunity observed in our experiments. We note that fully connected hidden-layer sparsity was not explicitly imposed (that is, not enforced by a penalty in the loss function), but emerged naturally through the training process. A normalized histogram of the hidden-layer activations is shown in Fig. 3d. A representative set of the convolutional kernels applied to feature maps in layer C2 (Fig. 1c) is shown in Fig. 3h. Processing by the convolutional layers is similar to that of compressed sensing, except that instead of assuming an explicit sparsifying transform (for example, wavelet), AUTOMAP simultaneously learns a sparse convolutional domain and its sparse representations through a joint optimization (see Methods for details).

We then studied the weight parameters of each trained network using a *t*-distributed stochastic neighbour embedding (t-SNE) analysis[28] (Fig. 3e–g), which embeds a high-dimensional dataset into a low-dimensional space for visualization. Here we visualize the spatial

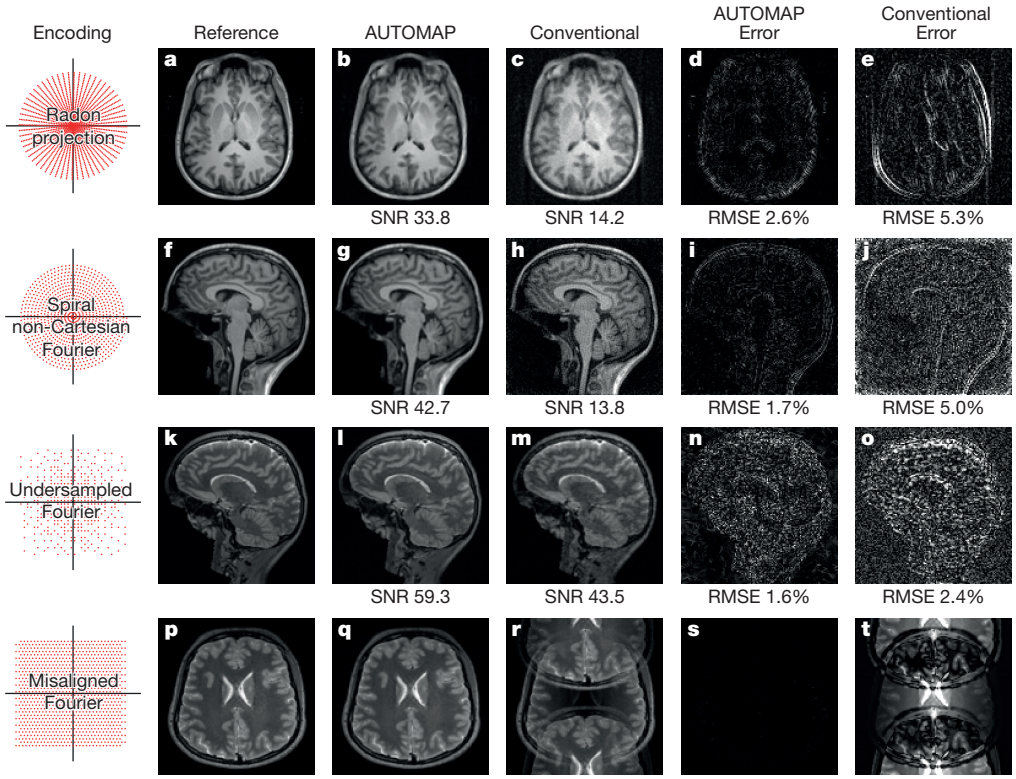


Figure 2 | Reconstruction performance of AUTOMAP compared with conventional techniques. Reference brain images were encoded into sensor-domain sampling strategies with varying levels of additive white Gaussian noise and reconstructed with both AUTOMAP and conventional approaches. **a**–**e**, Radon projection encoding compared with model-based iterative reconstruction. **f**–**j**, Spiral *k*-space encoding compared with conjugate-gradient SENSE reconstruction with NUFFT regridding. **k**–**o**, Poisson-disk undersampled (40%) Cartesian *k*-space encoding compared with compressed sensing reconstruction using the wavelet sparsifying transform. **p**–**t**, Misaligned Cartesian *k*-space, compared with conventional inverse fast Fourier transform. Image magnitude signal-to-noise ratios (SNRs) and error maps (with root-mean-squared error (RMSE) calculations) with respect to reference ground truth images are also shown. For each encoding experiment, both error maps are windowed to the same level. The same network architecture was used for all AUTOMAP reconstructions.

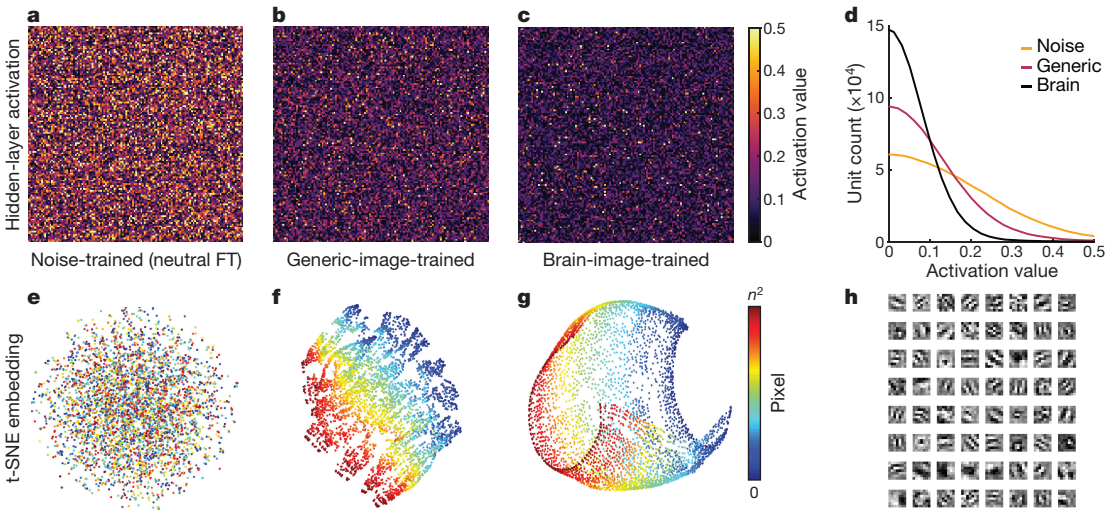


Figure 3 | Analysis of AUTOMAP neural networks. AUTOMAP was trained on three separate datasets for a Cartesian *k*-space encoding: generic natural images, brain images, and random-valued noise without any real-world image structure (see Methods for training details). **a**–**c**, Activation values of the fully connected hidden layer (FC2 in Fig. 1c) for each trained network while reconstructing the same *k*-space of a brain image. The noise-trained network generates high-amplitude and widely distributed hidden layer activation values (**a**), while the networks trained on generic images (**b**) and brain images (**c**) exhibit greater sparsity, indicating efficient processing of input data due to successful feature extraction when trained on relevant data. **d**, Histogram of FC2 activation values for the three networks, accumulated over 100 brain-image *k*-space reconstructions. **e**–**g**, Three-dimensional t-SNE embedding of network weights from FC2 to FC3 for the differently trained networks (see Methods for t-SNE analysis details). The t-SNE of the noise-trained network, agnostic to real-world image structure, exhibits disorganized structure (**e**), in contrast to **f** and **g**, which reflect the local spatial correlation that exists in real-world images. The domain-specific training of the brain-trained network shows the highest similarity between weights to two-dimensional neighbours for all pixel locations (**g**). **h**, Representative sparse convolutional kernels of the final convolutional stage (C2–Image) learned from training on brain images.

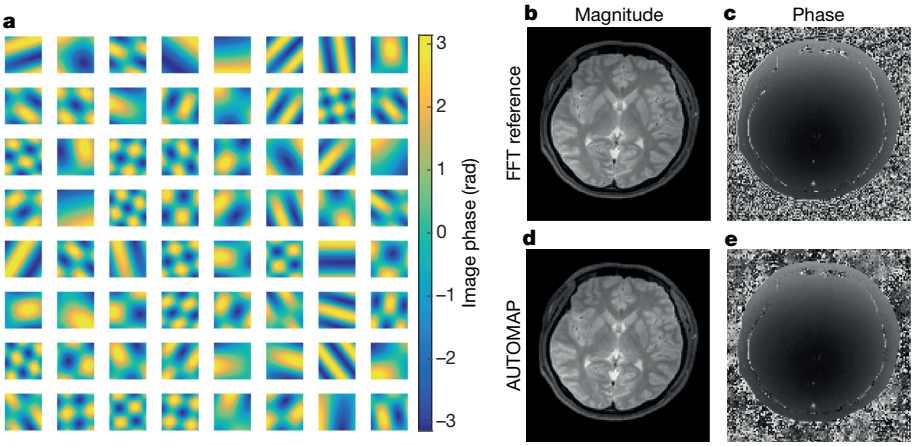


Figure 4 | Learning reconstruction of phase for *in vivo* data. The inclusion of synthetic phase to the training dataset enables AUTOMAP to properly reconstruct both the magnitude and phase. **a**, Synthetic *k*-space data for training was generated from HCP magnitude images by phase modulation with two-dimensional sinusoids of varying spatial frequencies. **b**–**e**, After training, the magnitude (**b**) and phase (**c**) of T2-weighted raw *k*-space acquired from a human subject are properly reconstructed by AUTOMAP (**d**, **e**). (FFT, fast Fourier transform.)

relationship of trained network weights, particularly from FC2 to the pre-convolutional FC3 image layer (see Methods for details). Figure 3d shows that the t-SNE embedding of the noise-trained network weights is highly unorganized and unbiased with respect to pixel location; this is unsurprising because the network learns a 'pure' or 'neutral' Fourier transform that does not recognize the local spatial correlation that exists in real-world images[29]. The generic-image-trained model captures this local spatial correlation, so the weights of neighbouring pixels are more similar than the weights of pixels further apart, as shown by the t-SNE embedding (Fig. 3f). This feature is most clearly exhibited by the t-SNE of the brain-trained network weights (Fig. 3g), which are organized into a two-dimensional sheet within the three-dimensional embedding space, demonstrating extremely high similarity between the weights of two-dimensional neighbours for all pixel locations.

We then demonstrated AUTOMAP's ability to learn reconstruction of image phase from complex-valued sensor data by including phase-modulated data in the network training. A phase-modulated training set was created by generating synthetic phase patterns (examples shown in Fig. 4a) to modulate the magnitude-only training images collected from the HCP database before being encoded in *k*-space. Using the same *k*-space data as input, we trained separate AUTOMAP networks to reconstruct magnitude and phase with their respective target training images, and validated this by reconstructing *in vivo* *k*-space raw data taken from a human subject on a 3-tesla (3T) MRI scanner (Fig. 4d, e); see Methods for acquisition details. This reconstruction can also be performed on one larger network with concatenated magnitude and phase data. This phase-modulation training allows public medical image databases and private PACS (picture archiving and communication system) repositories to be used for training, despite the typical absence of phase data in these datasets. Furthermore, our results show that parameterized influences on the input signal can be simulated for training and subsequently disentangled by AUTOMAP, which may be useful in sophisticated reconstruction problems such as automated motion compensation (see Methods and Extended Data Fig. 4 for more detailed discussion).

A potential concern in the reconstruction of real-world experimentally acquired sensor data is whether the AUTOMAP network would overfit to the ideal sampling parameters used during training and as a result be overly sensitive to sampling deviation during actual acquisitions. We quantified the effect of divergence from the nominal sampling trajectory with Monte Carlo simulations of varying amplitudes of trajectory error (Fig. 5a) from a spiral acquisition, and measure the resulting reconstruction RMSE from the ground truth reference. To examine a broad range of potential errors with realistic trajectories, we measured the actual trajectory during a spiral acquisition and computed the difference vectors of samples between the actual and ideal designed trajectories. This was used to scale the vector magnitudes to generate offset sampling trajectories. We tested errors from zero deviation (perfect match) to four times the measured deviation, and found that AUTOMAP's reconstruction error smoothly increased as a function of trajectory error, similar to a conventional NUFFT reconstruction's error curve (Fig. 5b), demonstrating reasonable robustness to trajectory deviation. Although the AUTOMAP error curve was slightly steeper, AUTOMAP still achieved better reconstruction accuracy than did NUFFT, out to very large trajectory errors, more than 3.5 times larger than the measured experimental deviation (Fig. 5b) from a commercial 3T MRI scanner.

These simulation results are consistent with the reconstruction performance on real scanner data acquired from human subjects. Figure 5c, d shows AUTOMAP and NUFFT reconstructions of a 10-interleave spiral magnetic resonance acquisition, in which both methods assume the nominal trajectory that deviates from the actual experimental scan trajectory. Although there is no ground truth with which to calculate reconstruction error, image signal-to-noise ratio was measured to be higher in the AUTOMAP output (21.6 versus 17.6). Figure 5e, f displays windowed versions of Fig. 5c, d, revealing coherent object-dependent and ringing artefacts in the NUFFT reconstruction (Fig. 5f); these are much reduced in the AUTOMAP reconstruction, primarily exhibiting standard Gaussian white noise (Fig. 5e).

Finally, we demonstrate reconstruction of multichannel magnetic resonance data acquired on a clinical 3T scanner with 15.5 times undersampling (acceleration factor $R = 4 \times 4$ uniform with low frequency region) retrospective undersampling (Fig. 5g–k). In comparison to a conventional SENSE reconstruction (Fig. 5i), AUTOMAP (Fig. 5h) demonstrates reduced noise and reconstruction artefacts, which can clearly be observed in the error maps (Fig. 5j, k) and quantified by a reduction in RMSE from 10.8% to 6.72%). Further acquisition and reconstruction details can be found in Methods.

At its core, AUTOMAP is a conceptual approach for trained image reconstruction with manifold learning; the specific neural network implementation presented here is not the only possible implementation, but a first demonstration that can be extended and improved upon in many directions. In Methods we discuss the application of AUTOMAP to other reconstruction problems and ways to address practical implementation challenges. As an example of applicability beyond MRI, human ^{18}F-fluorodeoxyglucose (FDG) positron emission tomography (PET) data are reconstructed with AUTOMAP in Extended Data Fig. 5.

AUTOMAP provides a new paradigm for image reconstruction that learns a reconstruction function for arbitrary acquisition strategies,

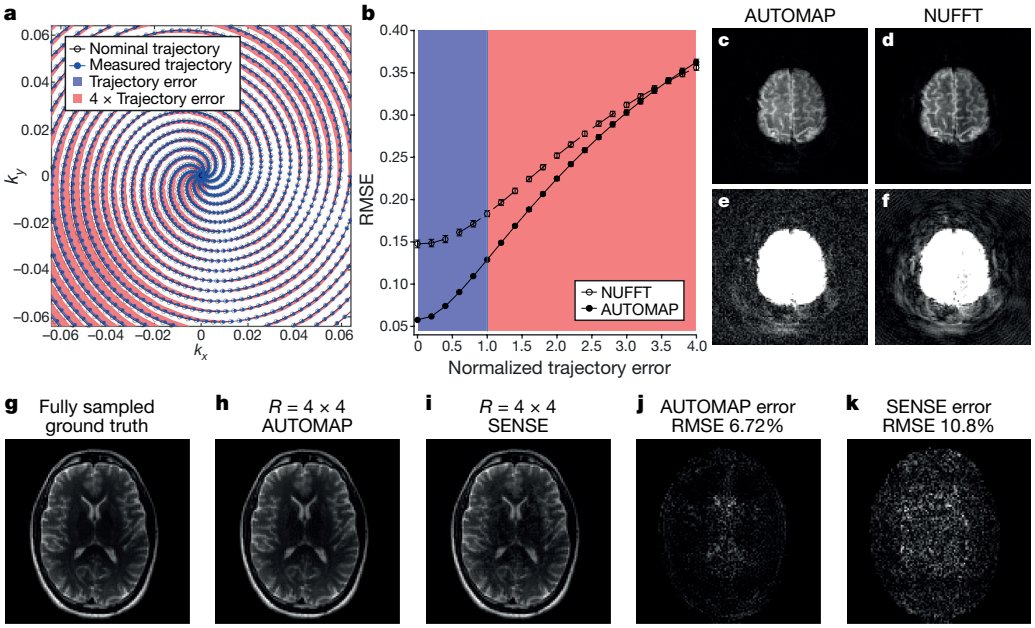


Figure 5 | Performance of AUTOMAP in real-world acquisitions. **a**, Plot of the family of *k*-space trajectory deviations used in Monte Carlo reconstruction analysis as described in the text. **b**, RMSE of AUTOMAP and NUFFT Monte Carlo reconstructions as a function of normalized trajectory error, with unity error corresponding to the measured experimental trajectory deviation in a 3T clinical scanner. Error bars indicate standard error of the mean. **c**, **d**, AUTOMAP and NUFFT reconstructions of a real-world 3T magnetic resonance spiral acquisition; the mean image signal-to-noise ratio is improved in the AUTOMAP reconstruction (21.6 versus 17.6). **e**, **f**, The presence of reconstruction artefacts is also reduced, as revealed in the windowed images. **g**–**i**, AUTOMAP and SENSE reconstructions of a real-world $R = 4 \times 4$ undersampled 32-channel 3T magnetic resonance acquisition are displayed and compared with the fully sampled reference. **j**, **k**, Error maps are windowed to the same intensity level. The RMSE reconstruction error is reduced with AUTOMAP compared to SENSE (6.72% versus 10.8%).

conditioned upon low-dimensional features of real-world data to improve artefact reduction and reconstruction accuracy for noisy and undersampled acquisitions. We anticipate that the noise robustness attainable with our approach will improve imaging quality and speed for a broad range of applications exhibiting low signal-to-noise ratio, including low-dose X-ray computed tomography[30], low-light charge-coupled devices[31], large-baseline radio astronomy[32] and rapid volumetric optical coherence tomography[33]. Finally, we also anticipate that the AUTOMAP paradigm will enable the development of new classes of acquisition strategies across imaging modalities as the reconstruction of arbitrary encoding schemes can be learned without domain expert knowledge.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 12 April 2017; accepted 23 January 2018.

1. Grangeat, P. *Tomography* (John Wiley & Sons, 2013).
2. Gull, S. F. & Daniell, G. J. Image reconstruction from incomplete and noisy data. *Nature* **272,** 686–690 (1978).
3. Zeng, G. L. *Medical Image Reconstruction* (Springer, 2010).
4. Yu, Z., Thibault, J.-B., Bouman, C. A., Sauer, K. D. & Hsieh, J. Fast model-based X-ray CT reconstruction using spatially nonhomogeneous ICD optimization. *IEEE Trans. Image Process.* **20,** 161–175 (2011).
5. Pruessmann, K. P., Weiger, M., Börnert, P. & Boesiger, P. Advances in sensitivity encoding with arbitrary k-space trajectories. *Magn. Reson. Med.* **46,** 638–651 (2001).
6. Hinton, G. *et al.* Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. *IEEE Signal Process. Mag.* **29,** 82–97 (2012).
7. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* 1097–1105 (2012).
8. Gilbert, C. D., Sigman, M. & Crist, R. E. The neural basis of perceptual learning. *Neuron* **31,** 681–697 (2001).
9. Lu, Z.-L., Hua, T., Huang, C.-B., Zhou, Y. & Dosher, B. A. Visual perceptual learning. *Neurobiol. Learn. Mem.* **95,** 145–151 (2011).
10. Vincent, P., Larochelle, H., Bengio, Y. & Manzagol, P.-A. Extracting and composing robust features with denoising autoencoders. In *Proc. 25th Int. Conf. on 'Machine Learning'* 1096–1103, http://www.cs.toronto.edu/~larocheh/publications/icml-2008-denoising-autoencoders.pdf (2008).
11. Ogawa, T., Kosugi, Y. & Kanada, H. Neural network based solution to inverse problems. In *IEEE World Congr. on 'Computational Intelligence'* Vol. 3, 2471–2476, http://ieeexplore.ieee.org/document/687250/ (1998).
12. Schiller, H. & Doerffer, R. Neural network for emulation of an inverse model operational derivation of Case II water properties from MERIS data. *Int. J. Remote Sens.* **20,** 1735–1746 (1999).
13. Hoole, S. R. H. Artificial neural networks in the solution of inverse electromagnetic field problems. *IEEE Trans. Magn.* **29,** 1931–1934 (1993).
14. Floyd, C. E. An artificial neural network for SPECT image reconstruction. *IEEE Trans. Med. Imaging* **10,** 485–487 (1991).
15. Pelt, D. M. & Batenburg, K. J. Fast tomographic reconstruction from limited data using artificial neural networks. *IEEE Trans. Image Process.* **22,** 5238–5251 (2013).
16. Jin, K. H., McCann, M. T., Froustey, E. & Unser, M. Deep convolutional neural network for inverse problems in imaging. *IEEE Trans. Image Process.* **26,** 4509–4522 (2017).
17. Hammernik, K. *et al.* Learning a variational network for reconstruction of accelerated MRI data. *Magn. Reson. Med.* **79,** 3055–3071 (2017).
18. Lustig, M., Donoho, D. & Pauly, J. M. Sparse MRI: the application of compressed sensing for rapid MR imaging. *Magn. Reson. Med.* **58,** 1182–1195 (2007).
19. Fan, Q. *et al.* MGH–USC Human Connectome Project datasets with ultra-high *b*-value diffusion MRI. *Neuroimage* **124,** 1108–1114 (2016).
20. Deng, J. *et al.* ImageNet: a large-scale hierarchical image database. In *IEEE Conf. on 'Computer Vision and Pattern Recognition'* 248–255, http://www.image-net.org/papers/imagenet_cvpr09.pdf (2009).
21. Hornik, K., Stinchcombe, M. & White, H. Multilayer feedforward networks are universal approximators. *Neural Netw.* **2,** 359–366 (1989).
22. Di Carli, M. F. & Lipton, M. J. *Cardiac PET and PET/CT Imaging* (Springer, 2007).
23. Yang, Z. & Jacob, M. Mean square optimal NUFFT approximation for efficient non-Cartesian MRI reconstruction. *J. Magn. Reson.* **242,** 126–135 (2014).
24. Virtue, P. & Lustig, M. On the empirical effect of Gaussian noise in undersampled MRI reconstruction. Preprint at https://arxiv.org/abs/1610.00410 (2016).
25. Brown, R. W., Cheng, Y. C. N., Haacke, E. M., Thompson, M. R. & Venkatesan, R. *Magnetic Resonance Imaging: Physical Principles and Sequence Design* 2nd edn (Wiley, 2014).
26. Gold, J., Bennett, P. J. & Sekuler, A. B. Signal but not noise changes with perceptual learning. *Nature* **402,** 176–178 (1999).
27. Wright, J. *et al.* Sparse representation for computer vision and pattern recognition. *Proc. IEEE* **98,** 1031–1044 (2010).
28. Maaten, L. V. D. & Hinton, G. Visualizing Data using t-SNE. *J. Mach. Learn. Res.* **9,** 2579–2605 (2008).
29. Getis, A. in *Handbook of Applied Spatial Analysis* (eds Fisher, M. M. & Getis, A.) 255–278 (Springer, 2010).

30. Kubo, T. *et al.* Radiation dose reduction in chest CT: a review. *Am. J. Roentgenol.* **190,** 335–343 (2008).
31. Daigle, O., Djazovski, O., Laurin, D., Doyon, R. & Artigau, É. Characterization results of EMCCDs for extreme low-light imaging. In *Proc. SPIE on 'High Energy, Optical, and Infrared Detectors for Astronomy V'* Vol. 8453, 845303, https://doi.org/10.1117/12.926385 (2012).
32. Girard, J. N. *et al.* Sparse representations and convex optimization as tools for LOFAR radio interferometric imaging. *J. Instrum.* **10,** C08013 (2015).
33. Lebed, E., Sarunic, M. V., Beg, M. F. & Mackenzie, P. J. Rapid volumetric OCT image acquisition using compressive sampling. *Opt. Exp.* **18,** 21003–21012 (2010).

Acknowledgements We acknowledge M. Michalski and the computational resources and assistance provided by the Massachusetts General Hospital (MGH) and the Brigham and Women's Hospital (BWH) Center for Clinical Data Science (CCDS). The CCDS is supported by MGH, BWH, the MGH Department of Radiology, the BWH Department of Radiology, and through industry partnership with NVIDIA. We also acknowledge the Center for Machine Learning at Martinos. We also thank J. Stockmann, J. Polimeni, D. E. J. Waddington and R. L. Walsworth for their comments on this manuscript, and B. Bilgic and C. Liao for their assistance in human subject data acquisition. We acknowledge C. Catana for providing raw PET data and for filtered back projection and OSEM reconstructions. We also thank M. Haskell for providing the MRI motion encoding model. B.Z. was supported by the National Institutes of Health/National Institute of Biomedical Imaging and Bioengineering F32 Fellowship (EB022390). Data were provided in part by the HCP, MGH-USC Consortium (Principal Investigators: Bruce R. Rosen, Arthur W. Toga and Van Wedeen; U01MH093765), which was funded by the NIH Blueprint Initiative for Neuroscience Research grant; the National Institutes of Health grant P41EB015896; and the Instrumentation Grants S10RR023043, 1S10RR023401, 1S10RR019307.

Author Contributions B.Z., J.Z.L., S.F.C., B.R.R. and M.S.R. conceptualized the problem and contributed to experimental design. B.Z. developed, implemented and tested the technical framework. J.Z.L. and B.Z. constructed the theoretical description. B.Z., J.Z.L., S.F.C., B.R.R. and M.S.R. wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to M.S.R. (mrosen@cfa.harvard.edu).

METHODS

Image dataset acquisition and pre-processing. The training dataset of generic images was assembled from ImageNet[20]. 10,000 images from the 'animal', 'plant' and 'scene' categories were each cropped to the central 256 × 256 pixels and subsequently subsampled to 128 × 128. Y-channel luminance was extracted from the RGB colour images to form greyscale intensity images. Each image was then rotated in 90° increments to augment the dataset. The mean intensity of each image was subtracted and the entire dataset was normalized to a constant value defined by the maximum intensity of the dataset.

The training dataset of de-identified brain images was assembled from the MGH-USC HCP[19] public database, which were acquired with the T1-weighted three-dimensional MRI acquisition protocol MPRAGE with repetition time TR = 2,530 ms, echo time TE = 1.15 ms, inversion time TI = 1,100 ms, flip angle FA = 7.0° and bandwidth BW = 651 Hz Px^{-1} on a Siemens Skyra 3T MRI platform (Siemens Medical Solutions, Erlangen, Germany). The de-identification process used by the HCP protocol masked the face and ear regions to protect the subjects' privacy.

Axial, sagittal and coronal T1-weighted slices from 131 subjects were used to generate a 50,000-image dataset. For each image, the central 256 × 256 pixels were cropped and subsampled to 128 × 128. To promote translation invariance in the training, each image was symmetrically tiled to create a larger 256 × 256 image containing four reflections of the original, and cropped to a random 128 × 128 section. The same data normalization process described above for the ImageNet dataset was used.

The test data used in the first evaluation experiment were taken from another subject outside those used for training. The test data also included the T2-weighted MRI acquisition protocol SPACE with TR = 3,200 ms, TE = 561 ms, FOV = 224 mm × 244 mm and BW = 744 Hz Px^{-1} on the same Siemens Skyra 3T MRI platform.

Sensor-domain encoding of image data. Sensor-domain representations for each image were encoded according to the reconstruction task. For the Radon transform experiment (Fig. 2a–c), we used the discrete Radon transform with 180 projection angles and 185 parallel rays. The spiral *k*-space experiment (Fig. 2d–f) used nonuniform fast Fourier transform (NUFFT)[34] to encode a ten-interleave spiral trajectory[35] with variable density factor $\alpha = 1$ and undersampling factor $R = 1/1.2$ based on the pre-subsampled 256 × 256 images (the MATLAB code used for this trajectory encoding is available at http://bigwww.epfl.ch/algorithms/mri-reconstruction/). The undersampled Cartesian *k*-space experiment (Fig. 2g–i) used a Poisson-disk sampling pattern with 40% undersampling of the Fourier-transformed *k*-space generated with the Berkeley Advanced Reconstruction Toolbox (BART)[36]. The misaligned Cartesian *k*-space experiment (Fig. 2j–l) used Fourier-transformed *k*-space where an empirically observed phase shift from a typical echo planar imaging acquisition was applied to every odd readout line.

Modulation with synthesized phase. Phase-modulated training data was used in the raw sensor data experiments of Figs 4 and 5 to train networks to accurately process phase-modulated sensor data. Synthesized phase maps were created by generating two-dimensional sinusoids with varying spatial frequencies independent along each image axis, and rotated by a random angle with respect to the image axes. The intensities of the sinusoids represented phase values, and were normalized to be between 0 and 2π. Each magnitude image in the training dataset was then modulated with a randomly generated phase map to form the complex-valued target image, which was then encoded by the appropriate forward encoding model to produce the corresponding sensor-domain input.

Model architecture. The input to the neural network consists of a vector of sensor-domain-sampled data produced by the preprocessing steps detailed above. Because the input layer is fully connected to the first hidden layer, for each reconstruction task the sensor-domain data (typically represented in two dimensions for images) can be vectorized in any order without any effect on the training. Since the neural network computational framework used here (Tensorflow[37]) operates on real-valued inputs and parameters, complex data must be separated into real and imaginary components concatenated in the input vector. Thus, an $n \times n$ complex-valued *k*-space matrix, for example, is reshaped to a $2n^2 \times 1$ real-valued vector (for our experiments, $n = 128$). As schematically illustrated in Fig. 1c, the input layer FC1 is fully connected to an $n^2 \times 1$-dimensional hidden layer FC2 and activated by the hyperbolic tangent function. This first hidden layer is fully connected to another $n^2 \times 1$-dimensional hidden layer FC3 with hyperbolic tangent activation, and is reshaped to an $n \times n$ matrix in preparation for convolutional processing. The first convolutional layer C1 convolves 64 filters of 5 × 5 with stride 1 followed by a rectifier nonlinearity[38]. The second convolutional layer C2 again convolves 64 filters of 5 × 5 with stride 1 followed by a rectifier nonlinearity. The final output layer deconvolves the C2 layer with 64 filters of 7 × 7 with stride 1. The output layer represents the reconstructed magnitude image, except for the phase-modulation experiment, where the network was trained separately to reconstruct the real and imaginary components of the image.

Training details. The same network architecture and hyperparameters were used for our experiments. For each sensor encoding reconstruction task, a different network was trained from the corresponding sensor-domain encodings and target images applied to the inputs and outputs, respectively, of the neural network (details of training data and network architecture described above). One per cent multiplicative noise was applied to the input to promote manifold learning during training by forcing the network to learn robust representations from corrupted inputs[10]. We note that the specific noise distribution of this corruption process did not serve to model the additive Gaussian noise that was applied during evaluation. The RMSProp algorithm (see http://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf) was used with minibatches of size 100, learning rate 0.00002, momentum 0.0 and decay 0.9. The loss function minimized during training was a simple squared loss between the network output and target image intensity values, with an additional L1-norm penalty ($\lambda = 0.0001$) applied to the feature map activations in the final hidden layer C2 to promote sparse convolutional representations. The convolutional layers are inspired by Winner-Take-All autoencoders[39] that jointly optimize the sparse convolutional codes as well as the deconvolutional kernel 'dictionaries' upon which the final image is built (Fig. 3h). Note that this imposed sparsity on the convolutional layers is separate from the fully connected hidden-layer activation sparsity that emerged without an applied sparsifying penalty (Fig. 3a–c), and occurs even without imposed convolutional sparsity. Each network was trained for 100 epochs (duration typically 7–8 h) on the Tensorflow[37] deep learning framework using two NVIDIA Tesla P100 graphics processing units (GPUs) with 16-GB memory capacity each, specifically employing either a conventional server platform with two P100 GPUs or the NVIDIA DGX-1 using two GPUs per experiment. Example plots of training convergence are shown in Extended Data Fig. 3. The validation error tracking the training error without upward divergence demonstrates a stable training regime with good bias-variance tradeoff, indicating that model complexity is well matched to the reconstruction problem.

Evaluation on simulated sensor data. The performance of AUTOMAP-trained networks for the four acquisition strategies was evaluated by reconstructing the four sensor-domain encodings of T1- and T2-weighted MRI brain images of a human subject from the HCP database as described above. For the Radon transform, spiral *k*-space, and misaligned *k*-space experiments, the network was trained using ImageNet data; for the undersampled *k*-space experiment, the network was trained with data from the HCP brain image dataset using only T1-weighted images from other subjects in the HCP database.

We reconstructed the same set of sensor-domain inputs with conventional reconstruction techniques for each acquisition strategy: For Radon projection imaging, model-based iterative reconstruction[4] was used with generalized Huber function parameters $\delta = 0.05$ and $T = 4.0$ and run until the average magnitude of voxel updates was less than 1%, implemented with OpenMBIR software available at http://engineering.purdue.edu/~bouman/OpenMBIR/. Spiral-trajectory *k*-space was reconstructed with a single-coil implementation of conjugate gradient-SENSE using NUFFT regridding[5,34] with kernel size $J_d = 8$ samples, run over 30 conjugate gradient iterations, with MATLAB code available at http://bigwww.epfl.ch/algorithms/mri-reconstruction/. The Poisson-disk undersampled *k*-space was reconstructed with compressed sensing[18] using the wavelet sparsifying transform with the L1 penalty parameter $\lambda = 0.01$, using BART[36] with code available at https://mrirecon.github.io/bart/. Misaligned *k*-space (commonplace sampling inaccuracy due to hardware limitations or physiologic effects) was reconstructed with the native MATLAB implementation of the two-dimensional inverse fast Fourier transform.

To probe the noise sensitivity of the reconstructions, varying levels of additive white Gaussian noise (AWGN) were introduced to the sensor-domain signals: 25 dB AWGN signal-to-noise ratio (SNR) for the spiral experiment, 30 dB AWGN SNR for the undersampled Cartesian experiment and 40 dB AWGN SNR for the Radon projection experiment; these SNR measures indicate the power level of the additive noise relative to the power of the signals. The influence of this additive noise on the resultant image SNR of the reconstructions was measured with a standard Monte Carlo SNR map calculation[40] $\mathrm{SNR}(\boldsymbol{r}) = \bar{\boldsymbol{x}}_n(\boldsymbol{r})/\boldsymbol{\sigma}_n(\boldsymbol{r})$, where $\bar{\boldsymbol{x}}_n(\boldsymbol{r})$ and $\boldsymbol{\sigma}_n(\boldsymbol{r})$ are the mean and standard deviation, respectively, over Monte Carlo instances $n = 1, 2, \ldots, 100$, of the image magnitude at voxel $\boldsymbol{r}$. The representative SNR for an image was computed by taking the mean of the SNR map over the region of interest of voxels in the brain. We did not noise-corrupt the misaligned *k*-space because the sampling trajectory already represented a perturbed input. More extreme cases of noise corruption and its effects on reconstruction are shown in Extended Data Fig. 1.

Evaluation on raw MRI scanner data. Cartesian *k*-space test data (of Fig. 4) were acquired from a healthy volunteer on a 3T Siemens Trio MRI scanner with a spin-

echo imaging sequence with TR = 3,110 ms, TE = 23.0 ms, matrix size = 208 × 256 and slice thickness 3 mm. Data from the 12-channel receiver head coil was coil-compressed to one channel with singular value decomposition (SVD) and the central 128 × 128 *k*-space samples formed the input for the 128 × 128 matrix reconstruction task. As described above, the AUTOMAP network was trained on phase-modulated HCP brain data with Cartesian Fourier encoding; the acquired raw scanner data was then input to the trained network for reconstruction.

Spiral *k*-space data (Fig. 5) were acquired from a healthy volunteer on a Siemens 3T Prisma MRI scanner using a 32-channel head coil. A constant spiral trajectory was designed to cover a field of view of 256 × 256 mm^2 with 2 × 2 mm^2 in-plane resolution and 5 mm slice thickness. This was achieved using ten spiral interleaves, each having an 8-ms readout duration which included a 1-ms rewinder, with slew rate 133 mT m^{-1} ms^{-1} and maximum gradient strength 24 mT m^{-1}; TE = 35 ms, TR = 200 ms, and flip angle of 20°. Data from the multichannel receiver head coil was SVD coil-compressed to one channel. A calibration acquisition measurement was made to measure the actual sampling trajectory[41]. The AUTOMAP network was trained on phase-modulated HCP brain data with non-Cartesian encoding with the designed spiral trajectory; acquired raw scanner data was then input to the trained network for reconstruction, and compared with NUFFT regridding reconstruction[34] with kernel size $J_d = 6$ samples, using code available from http://web.eecs.umich.edu/~fessler/irt/irt/.

Multichannel T2-weighted data (of Fig. 5) were acquired on a 3T Siemens Trio with the standard Siemens 32-channel head array coil. A turbo spin echo sequence with 224 × 224 mm^2 field of view was acquired across 35 slices with a 30% distance factor. The imaging parameters are as follows: TR = 6.1 s, TE = 98 ms, flip angle 150°, and a resolution of 0.5 × 0.5 × 3.0 mm^3 with a matrix size of 448 × 448. The fully sampled uncombined complex *k*-space data were retrospectively undersampled to a 112 × 112 matrix, corresponding to 2 mm in-plane resolution. The channel data were then mixed down to 16 modes using the standard global SVD-based compression. Iterative SENSE reconstruction[42] was performed using the GMRES solver[43] with a stopping criterion of 1×10^{-4} relative error to generate the ground truth reconstruction. Sensor-domain data were then undersampled by 15.5 times with an $R = 4 \times 4$ coherent undersampling pattern and 5 × 5 low-frequency region, and reconstructed with AUTOMAP and SENSE using the SVD coil sensitivity profiles. The AUTOMAP network was trained on HCP brain images, which were modulated by the SVD coil sensitivity profiles to produce the multichannel training data. Each channel was Fourier-transformed and correspondingly undersampled with the same $R = 15.5$ pattern, and channels were concatenated at the network input.

Evaluation on raw PET scanner data. PET data were acquired from a healthy volunteer using the Biograph mMR scanner (Siemens Healthineers, Erlangen, Germany). The emission data corresponding to the 45–60-min interval post-administration of about 5 mCi of ^{18}F-FDG were used in this work. A PET volume was reconstructed using filtered back projection (Extended Data Fig. 5b) and the standard ordinary Poisson ordered subsets expectation maximization (OP-OSEM) algorithm[44] (Extended Data Fig. 5c), accounting for variable detector efficiency and photon attenuation and scatter using software provided by the manufacturer. The head attenuation map was generated from the magnetic resonance data using software developed in-house[45]. Spatial smoothing was performed after image reconstruction using a 5-mm full-width at half-maximum (FWHM) Gaussian kernel. A set of attenuation-corrected two-dimensional sinograms corresponding to the direct planes was also generated from the three-dimensional sinograms using the single slice rebinning algorithm, and was used as input to the AUTOMAP reconstruction network, which was trained on T1-weighted brain images from the HCP database, encoded with discrete Radon transform and Poisson sampling using native MATLAB functions. Although the absence of a ground truth image makes it difficult to evaluate the differences between the reconstruction techniques (Extended Data Fig. 5b–d), this experiment demonstrates the ability of AUTOMAP to reconstruct PET data acquired on a human scanner with results comparable to clinically used reconstruction methods.

t-SNE analysis. Relationship of trained network weights were visualized with t-SNE[28]. We employed a standard Cartesian Fourier *k*-space encoding for the networks. To reduce computational load, lower-resolution reconstruction networks were trained using 64 × 64 images from either ImageNet, brain images, or random-valued noise without any real-world image structure. In the visualization, each point corresponds to a single pixel in FC3, represented by an n^2-dimensional vector of weights directed to it from the FC2 layer. The label for each point is a scalar pixel location in the image space (from 0 to n^2) that also defines its colour in the visualization; similar colours correspond to similar pixel location. The t-SNE algorithm was implemented with perplexity 64 over 200 iterations with MATLAB code available at https://lvdmaaten.github.io/tsne/.

Description of AUTOMAP manifold learning. Our learning task is twofold. Given $\tilde{x}$, the noisy observation of sensor-domain data x, we first want to learn the stochastic projection operator onto $\mathcal{X}$: $p(\tilde{x}) = P(x|\tilde{x})$. After obtaining x, our second and more important task is to reconstruct $f(x)$ by producing a reconstruction mapping $\hat{f} : \mathbb{R}^{n^2} \to \mathbb{R}^{n^2}$ that minimizes the reconstruction error $L(\hat{f}(x), f(x))$.

We first describe the reconstruction process by considering the idealized scenario in which the input sensor data are noiseless. We denote the data $\{y_i, x_i\}_{i=1}^n$, where for the ith observation x_i indicates a $n \times n$ set of input parameters, and y_i indicates the $n \times n$ real, underlying images. We assume that (1) there exists a unknown smooth and homeomorphic function $f: \mathbb{R}^{n^2} \to \mathbb{R}^{n^2}$, such that $y = f(x)$, and (2) $\{x_i\}_{i=1}^n, \{y_i\}_{i=1}^n$ lie on unknown smooth manifolds $\mathcal{X}$ and $\mathcal{Y}$, respectively. Both manifolds are embedded in the ambient space $\mathbb{R}^{n^2}$, such that $\dim(\mathcal{X}) < n^2$ and $\dim(\mathcal{Y}) < n^2$.

The above two assumptions combine to define a joint manifold $\mathcal{M}_{\mathcal{X},\mathcal{Y}} = \mathcal{X} \times \mathcal{Y}$ within which the entire dataset $(x_i,\ y_i)_{i=1}^n$ lies, which can be written as:

$$\mathcal{M}_{\mathcal{X},\mathcal{Y}} = \{(x,\ f(x)) \in \mathbb{R}^{n^2} \times \mathbb{R}^{n^2} | x \in \mathcal{X},\ f(x) \in \mathcal{Y}\}$$

We note that as $(x, f(x))$ is described using the regular Euclidean coordinate system, we may equivalently describe this point using the intrinsic coordinate system of $\mathcal{M}_{\mathcal{X},\mathcal{Y}}$ as $(z, g(z))$, such that there exists a homeomorphic mapping $\phi = (\phi_x, \phi_y)$ between $(x, f(x))$ and $(z, g(z))$. That is, $x = \phi_x(z)$ and $f(x) = \phi_y \circ g(z)$. As an aside, in topology, $\phi = (\phi_x, \phi_y) : \mathcal{M}_{\mathcal{X},\mathcal{Y}} \to \mathbb{R}^{n^2} \times \mathbb{R}^{n^2}$ corresponds to the local coordinate chart of $\mathcal{M}_{\mathcal{X},\mathcal{Y}}$ at the neighbourhood of $(x, f(x))$. Instead of learning f directly in the ambient space, we wish to learn the diffeomorphism g between $\mathcal{X}$ and $\mathcal{Y}$ in order to take advantage of the low-dimensional nature of the embedded space. Consequently, the process of generating $y = f(x)$ from x can be written as a sequence of function evaluations:

$$f(x) = \phi_y \circ g \circ \phi_x^{-1}(x)$$

For the convenience of later presentation, we notice that given input image x, the output image follows a probability distribution $Q(Y|X = x, f)$, which is a degenerate distribution with point mass at $y = f(x)$.

We now turn to the more realistic scenario where corruption exists in the sensor-domain input and describe the denoising process. Instead of observing the perfect input data x_i, we observe $\tilde{x}_i$, which is a corrupted version of x_i by some known corruption process described by the probability distribution $P(\tilde{X}|X = x)$. To handle this complication, we seek to learn a denoising step $Q(X|\tilde{X} = \tilde{x}, p)$ to our model pipeline, such that our prediction for y is no longer a deterministic value, but a random variable with conditional distribution $P(Y|\tilde{X})$ so that we can properly characterize the prediction uncertainty caused by the corruption process.

Instead of learning this denoising step explicitly, we draw an analogy from the denoising autoencoder and model the joint distribution $P(Y, X, \tilde{X})$ instead. Specifically, in addition to the assumptions (1) and (2) listed above, we also assume (3) that the true distribution $P(X|\tilde{X})$ lies in the semiparametric family $\mathbb{Q}$ defined by its first moment $\mathbb{Q} = \{Q(X|\tilde{X} = \tilde{x}, p) | E(X) = p(\tilde{X})\}$.

We model $P(Y, X, \tilde{X})$ using the decomposition below:

$$Q_{(f,p)}(Y, X, \tilde{X}) = Q(Y|X, f)Q(X|\tilde{X}, p)P(\tilde{X})$$

where $Q(Y|X, f)$ denotes the model for the reconstruction process that we described earlier, $Q(X|\tilde{X}, p)$ is the denoising operator that we seek to learn, and $P(\tilde{X})$ is the empirical distribution of corrupted images. We note that we can combine the models for denoising and reconstruction processes together by collapsing the first two terms on the right-hand side into one term, which gives:

$$Q_{(f,p)}(Y, X, \tilde{X}) = Q(Y, X|\tilde{X}, (f, p))P(\tilde{X})$$

We note that $Y = f(X)$ is a deterministic and homeomorphic mapping of X, so $Q(Y, X|\tilde{X}, (f, p)) = Q(Y|\tilde{X}, (f, p))$ is the predictive distribution of output image y given the noisy input $\tilde{x}$, which is exactly our estimator of interest. Consequently, the model can be written as:

$$Q_{(f,p)}(Y, X, \tilde{X}) = Q(Y|\tilde{X}, (f, p))P(\tilde{X})$$

This completes the definition of our model for the joint distribution.

In the actual training stage, we usually took advantage of the fact that perfect input images x are available, and train the model with $\tilde{x}$ that we generated from $P(\tilde{X}|X = x)$. That is to say, the joint distribution of $(Y, X, \tilde{X})$ observed in training data admits the form:

$$P(Y, X, \tilde{X}) = P(Y|X)P(\tilde{X}|X)P(X)$$

The training proceeds by minimizing the Kullback–Liebler divergence between observed probability $P(Y, X, \tilde{X})$ and our model $Q(Y, X, \tilde{X})$:

$$\mathbb{D}_{\mathrm{KL}}\{P(Y,X,\tilde{X})||Q_{(f,p)}(Y,X,\tilde{X})\}$$

with respect to the function-valued parameters (f, p). As the Kullback–Liebler divergence converges towards 0, $Q(X|\tilde{X},p)$ converges to $P(X|\tilde{X})$, the denoising projection, and at the same time $Q(Y|\tilde{X},(f,p))$ converges to $P(Y|\tilde{X})$.

There exists a rich literature[46–50] on explicitly learning the stochastic projection p, diffeomorphism g, and the local coordinate chart ϕ. However, we notice that since $(\phi_f, \phi_x, p, g) \in \mathbb{C}^\infty$ (where $\mathbb{C}^\infty$ denotes the set of infinitely differentiable functions), $\hat{f} = \phi_f \circ g \circ \phi_x^{-1} \circ p$ as a whole is a continuously differentiable function on a compact subset of $\mathbb{R}^{n^2}$, and can therefore theoretically be approximated by the universal approximation theorem[51].

Practical considerations. While the idealized conception of the universal approximation theorem requires an infinite number of hidden nodes to achieve a perfect representation of an arbitrary continuous function[52], it has been demonstrated that practical neural network implementations with finite hidden layers achieve a bounded approximation error for functions on a compact set, which applies to all domain transform functions that govern contemporary imaging systems[21,53,54]. Furthermore, the system representation for state-of-the-art image reconstructions are discrete models that can be described exactly by a fully connected network (for example, in the simplest case, an inverse discrete Fourier transform matrix is applied to Nyquist sampled data). In addition, many inverse encoding models for image reconstruction can be represented using limited support (for example, SENSE reconstruction of uniformly or CAIPIRINHA (controlled aliasing in parallel imaging results in higher acceleration) staggered undersampled data results in a block diagonal inverse encoding, TV smoothness constraint for compressed sensing (Poisson matrix) produces a semiseparable inverse encoding with linear complexity, and so on). Using a discrete/finite neural network, AUTOMAP aims to expand upon these analytic models for image reconstruction.

The features identified in the AUTOMAP training process are not generated to create maximal distinction between certain categories of objects or subjects, as is common for many image classification tasks. Instead, our loss function forces minimization of pixelwise error and thus prioritizes reconstruction accuracy. As a result, our features are constrained to serve that purpose (note the resemblance of the deconvolutional kernels of Fig. 3h to Gabor-like edge detectors). This training approach makes AUTOMAP much less likely to produce 'hallucinatory behaviours' that arise from the use of finely tuned category-specific feature representations (as exploited by Google's DeepDream system[55]).

Given knowable systematic defects in the acquisition chain that can be modelled in the signal encoding (for example, measurable magnetic resonance gradient timing delays[56,57]), an appropriate training set can be generated to allow AUTOMAP to compensate these acquisition nonidealities without manually designed postprocessing requiring expert knowledge. However, as with other reconstruction methods, untrained and unaccounted-for acquisition errors (for example, subject motion or voltage or current spikes in hardware) will produce errors in the reconstruction with the current implementation of AUTOMAP. Ideally, to detect and compensate for these and other unpredictable or object-dependent artefacts such as X-ray scattering in computed tomography or chemical shift in MRI would probably require incorporating a more sophisticated discriminator network[58] into AUTOMAP, similar to those used in generative adversarial networks to quantitatively evaluate the quality of the reconstructed image and iteratively adapt the reconstruction process via modification of the AUTOMAP network weights to reduce artefacts.

However, in this current implementation, AUTOMAP reconstruction is stable in the presence of typical sampling trajectory variation during an MRI acquisition (Fig. 5a–f). Extending our evaluation to another class of encoding errors, we performed a simulated motion corruption experiment using the TAMER motion encoding model[59]. For standard multi-shot MRI acquisitions, patient motion between each imaging shot creates discrepancies in the encoding that often result in large ringing and blurring artefacts. As can be seen in Fig. 3, AUTOMAP extracts important data interdependency relationships and it is important to examine the robustness of these under realistic patient imaging situations. Using T2-weighted data (from the acquisition associated with Fig. 5g) we have simulated motion corruption using a realistic motion trajectory that was measured during an fMRI scan of a patient with Alzheimer's disease. Specifically, in-plane motion variation was applied between each imaging shot to create representative artefacts which would be seen during a 2D Turbo Spin Echo acquisition with turbo factor 4. The fully sampled data were processed using AUTOMAP and the standard Fourier transform reconstruction. As can be seen in Extended Data Fig. 4, AUTOMAP does not exhibit instability in the presence of data corruption owing to patient motion and shows comparable artefact level and structure to standard reconstructions.

The role of expert knowledge in the AUTOMAP reconstruction framework requires careful consideration, especially for future developments into extended applications. In this paper, our emphasis on withholding domain-specific knowledge is to explore the extreme case where expert knowledge is removed from the system where possible; thus our results are representative of what is probably a conservative limit on performance gains of introducing machine learning into the domain transform problem; we expect the appropriate integration of human domain-expert knowledge to AUTOMAP to yield even greater performance.

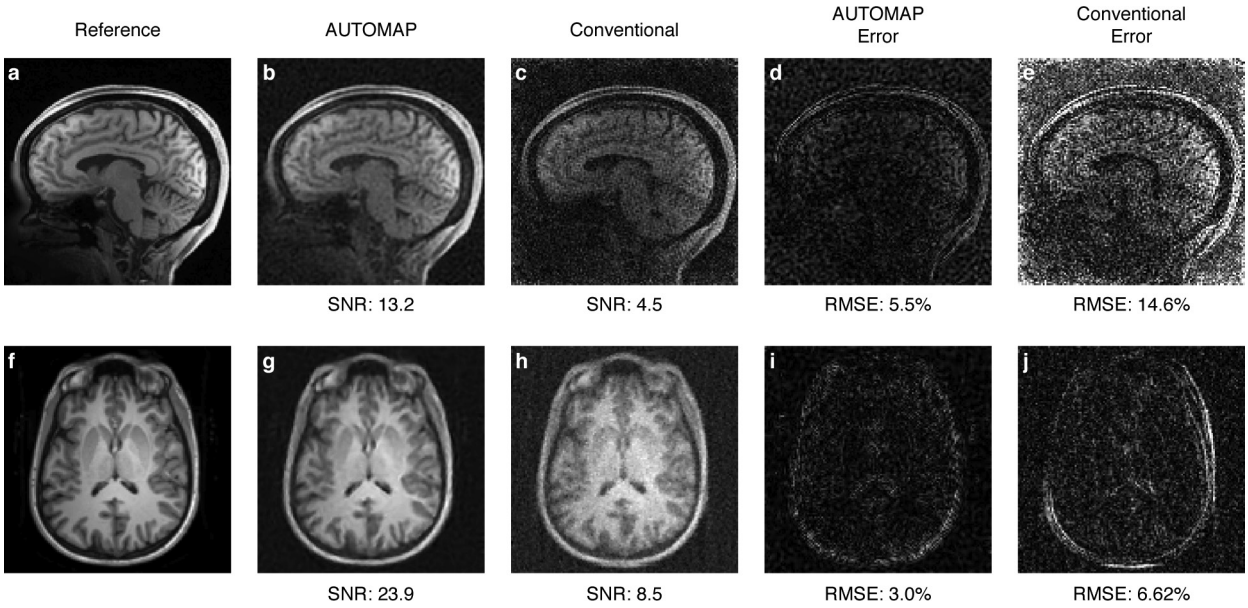
One particular area in which we anticipate benefits from expert knowledge is in the scaling of AUTOMAP to reconstruct higher-resolution images more efficiently. Given the current densely connected layer architecture, the number of weights in these layers scale by n^2 (compared to the fast Fourier transform complexity of the order of $n\log n$, for example), which can be problematic as memory and computation are limited resources. However, it is important to note the role that locality and separability play in image reconstruction. When considering standard MRI reconstruction methods, locality can be easily observed by examining the point spread function that defines the interaction between imaging voxels[60–63]. This strong locality is an important property because it relates to the stability of solving the inverse problem, and has the added advantage of often resulting in nearly or fully decoupled reconstruction problems. As has been demonstrated in the recent MRI literature, decomposition methods such as the alternating direction method of multipliers (ADMM) have been used effectively to combine smaller semiseparable reconstruction problems to solve larger fully coupled optimizations (for example, maximum likelihood)[64,65]. In the presence of hardware limitations, we expect the AUTOMAP framework to be directly employed on domain decomposed subsets to perform larger reconstructions. This would simply require the training of networks that describe the influence of subsets of acquired data to overlapping subsets of voxels. Given the success of methods that take advantage of locality properties (Schwarz alternating method/domain decomposition in linear algebra, Dantzig–Wolfe decomposition for linear programming, ADMM for convex optimization, and so on), we do not envision this being a limiting factor for AUTOMAP. In this light, the ability of AUTOMAP to accurately and robustly represent an inverse encoding model for the reconstruction of real-world multichannel data (Fig. 5g–k) will allow it to serve as a core building block for large-scale reconstructions through the use of domain decomposition, variable splitting and alternating direction methods. Finally, clinically oriented implementations of AUTOMAP would probably be focused on a particular imaging modality or class of acquisitions within a modality, and would thus allow for limiting the universality of the network to obtain compact representations that can be generated though network compression methods such as pruning[66,67].

Data availability. The generic natural images used for training are available from the ImageNet database (http://www.image-net.org/). The brain images used for training and evaluation were obtained from the MGH-USC HCP database (https://db.humanconnectome.org/).

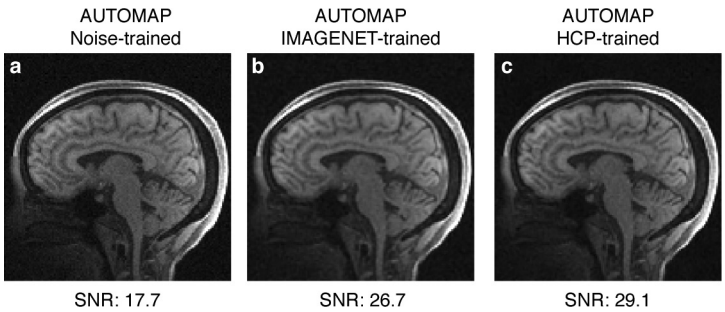
Code availability. Source code is available from the corresponding author upon reasonable request.

34. Fessler, J. A. & Sutton, B. P. Nonuniform fast Fourier transforms using min-max interpolation. *IEEE Trans. Signal Process.* **51,** 560–574 (2003).
35. Kim, D. H., Adalsteinsson, E. & Spielman, D. M. Simple analytic variable density spiral design. *Magn. Reson. Med.* **50,** 214–219 (2003).
36. Uecker, M., Ong, F., Tamir, J. I. & Bahri, D. Berkeley advanced reconstruction toolbox. *Proc. Int. Soc. Magnetic Resonance in Medicine* 2486 (2015).
37. Abadi, M. *et al.* TensorFlow: large-scale machine learning on heterogeneous distributed systems. Preprint at https://arxiv.org/abs/1603.04467 (2016).
38. Nair, V. & Hinton, G. E. Rectified linear units improve restricted Boltzmann machines. In *Proc. 27th Int. Conf. on 'Machine Learning'* 807–814 (ACM, 2010).
39. Makhzani, A. & Frey, B. J. Winner-take-all autoencoders. *Adv. Neural Inf. Process. Syst.* **28,** 2791–2799 (2015).
40. Reeder, S. B. *et al.* Practical approaches to the evaluation of signal-to-noise ratio performance with parallel imaging: application with cardiac imaging and a 32-channel cardiac coil. *Magn. Reson. Med.* **54,** 748–754 (2005).
41. Duyn, J. H., Yang, Y., Frank, J. A. & van der Veen, J. W. Simple correction method for k-space trajectory deviations in MRI. *J. Magn. Reson.* **132,** 150–153 (1998).
42. Pruessmann, K. P., Weiger, M., Scheidegger, M. B. & Boesiger, P. SENSE: sensitivity encoding for fast MRI. *Magn. Reson. Med.* **42,** 952–962 (1999).
43. Saad, Y. & Schultz, M. H. GMRES: a generalized minimal residual algorithm for solving nonsymmetric linear systems. *SIAM J. Sci. Statist. Comput.* **7,** 856–869 (1986).
44. Comtat, C, *et al.* OSEM-3D Reconstruction Strategies for the ECAT HRRT. *IEEE Symp. Conf. Record Nuclear Science* **6,** 3492–3496 (2004).
45. Izquierdo-Garcia, D. *et al.* An SPM8-based approach for attenuation correction combining segmentation and nonrigid template formation: application to simultaneous PET/MR brain imaging. *J. Nucl. Med.* **55,** 1825–1830 (2014).
46. Yu, K. & Zhang, T. Improved local coordinate coding using local tangents. In *Proc. 27th Int. Conf. on 'Machine Learning'* 1215–1222 (ACM, 2010).

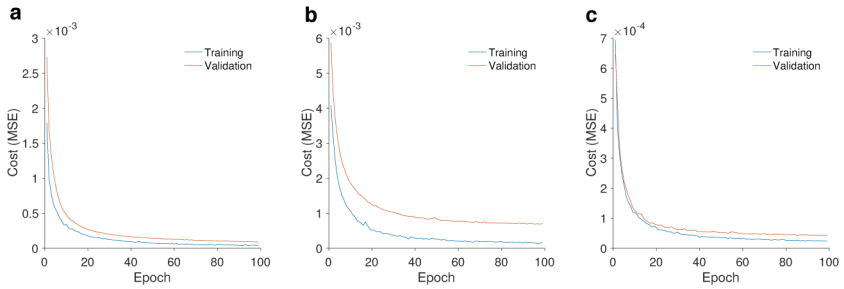
47. Anderes, E. & Coram, M. A general spline representation for nonparametric and semiparametric density estimates using diffeomorphisms. Preprint at https://arxiv.org/abs/1205.5314 (2012).
48. Zhang, M., Singh, N. & Fletcher, P. T. Bayesian estimation of regularization and atlas building in diffeomorphic image registration. *Int. Conf. Inf. Process. Med. Imaging* 37–48 (Springer, 2013).
49. Fishbaugh, J., Prastawa, M., Gerig, G. & Durrleman, S. Geodesic image regression with a sparse parameterization of diffeomorphisms. In *1st Int. Conf. on 'Geometric Science of Information' GSI 2013* (eds Nielsen, F. & Barbaresco, F.) Vol. 8085, 95–102, https://link.springer.com/chapter/10.1007/978-3-642-40020-9_9 (2013).
50. Bernstein, A., Kuleshov, A. & Yanovich, Y. *Manifold Learning in Regression Tasks. Statistical Learning and Data Sciences* 414–423 (Springer, 2015).
51. Hornik, K. Approximation capabilities of multilayer feedforward networks. *Neural Netw.* **4,** 251–257 (1991).
52. Irie, B. & Miyake, S. Capabilities of three-layered perceptrons. In *IEEE Int. Conf. on 'Neural Networks'* Vol. 1, 641–648 (1988).
53. Cybenko, G. Approximation by superpositions of a sigmoidal function. *Math. Contr. Signals Syst.* **2,** 303–314 (1989).
54. Barron, A. R. Approximation and estimation bounds for artificial neural networks. *Mach. Learn.* **14,** 115–133 (1994).
55. Mordvintsev, A., Olah, C. & Tyka, M. DeepDream—a code example for visualizing neural networks. https://research.googleblog.com/2015/07/deepdream-code-example-for-visualizing.html (Google Res, 2015).
56. Addy, N. O., Wu, H. H. & Nishimura, D. G. Simple method for MR gradient system characterization and k-space trajectory estimation. *Magn. Reson. Med.* **68,** 120–129 (2012).
57. Han, H., Ouriadov, A. V., Fordham, E. & Balcom, B. J. Direct measurement of magnetic field gradient waveforms. *Concepts Magn. Reson.* **36A,** 349–360 (2010).
58. Goodfellow, I., Pouget-Abadie, J. & Mirza, M. Generative adversarial nets. *Adv. Neural Inf. Process. Syst.* 2672–2680 (2014).
59. Haskell, M., Cauley, S. F. & Wald, L. L. TArgeted Motion Estimation and Reduction (TAMER): data consistency based motion mitigation for MRI using a reduced model joint optimization. *IEEE Trans. Med. Imaging* PP, 99, http://doi.org/10.1109/TMI.2018.2791482 (2018).
60. Fessler, J. A., Lee, S., Olafsson, V. T., Shi, H. R. & Noll, D. C. Toeplitz-based iterative image reconstruction for MRI with correction for magnetic field inhomogeneity. *IEEE Trans. Signal Process.* **53,** 3393–3402 (2005).
61. Cauley, S. F. *et al.* Fast reconstruction for multichannel compressed sensing using a hierarchically semiseparable solver. *Magn. Reson. Med.* **73,** 1034–1040 (2015).
62. Xi, Y., Xia, J., Cauley, S. & Balakrishnan, V. Superfast and stable structured solvers for Toeplitz least squares via randomized sampling. *SIAM J. Matrix Anal. Appl.* **35,** 44–72 (2014).
63. Xia, J., Chandrasekaran, S., Gu, M. & Li, X. S. Fast algorithms for hierarchically semiseparable matrices. *Numer. Linear Algebra Appl.* **17,** 953–976 (2010).
64. Weller, D. S., Ramani, S. & Fessler, J. A. Augmented Lagrangian with variable splitting for faster non-Cartesian L1-SPIRiT MR image reconstruction. *IEEE Trans. Med. Imaging* **33,** 351–361 (2014).
65. Zhao, B., Setsompop, K., Ye, H., Cauley, S. F. & Wald, L. L. Maximum likelihood reconstruction for magnetic resonance fingerprinting. *IEEE Trans. Med. Imaging* **35,** 1812–1823 (2016).
66. Han, S., Mao, H. & Dally, W. J. Deep compression: compressing deep neural networks with pruning, trained quantization and Huffman coding. Preprint at https://arxiv.org/abs/1510.00149 (2015).
67. Hu, H., Peng, R., Tai, Y.-W. & Tang, C.-K. Network trimming: a data-driven neuron pruning approach towards efficient deep architectures. Preprint at https://arxiv.org/abs/1607.03250 (2016).



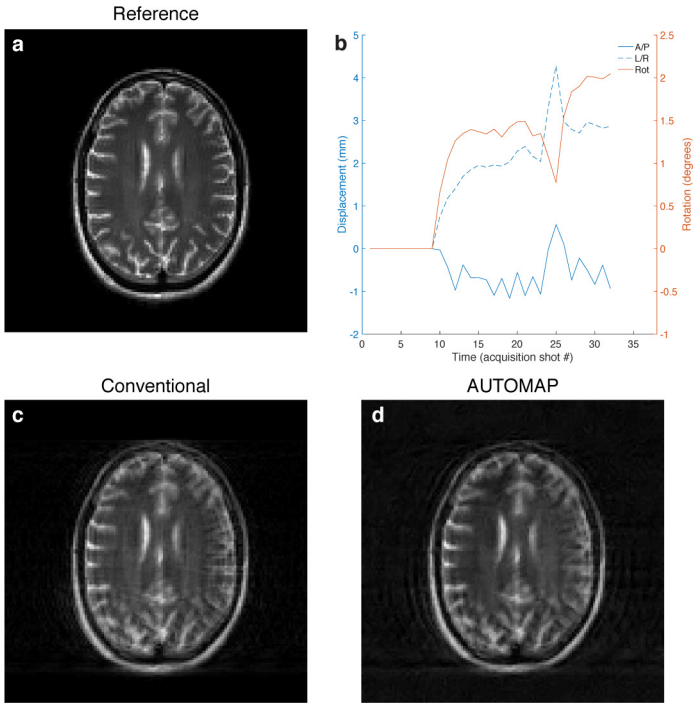
Extended Data Figure 1 | Reconstruction performance of AUTOMAP in low-signal-to-noise-ratio regimes. Reference brain images were encoded into sensor-domain sampling strategies with high levels of additive white Gaussian noise and reconstructed using both AUTOMAP and conventional approaches: **a**–**e**, spiral *k*-space encoding compared with conjugate-gradient SENSE reconstruction with NUFFT regridding; **f**–**j**, Radon projection encoding compared with model-based iterative reconstruction. Image magnitude signal-to-noise ratios (SNRs) and error maps (with root mean squared error calculations) with respect to reference ground truth images are also shown. For each encoding experiment, both error maps are windowed to the same level.



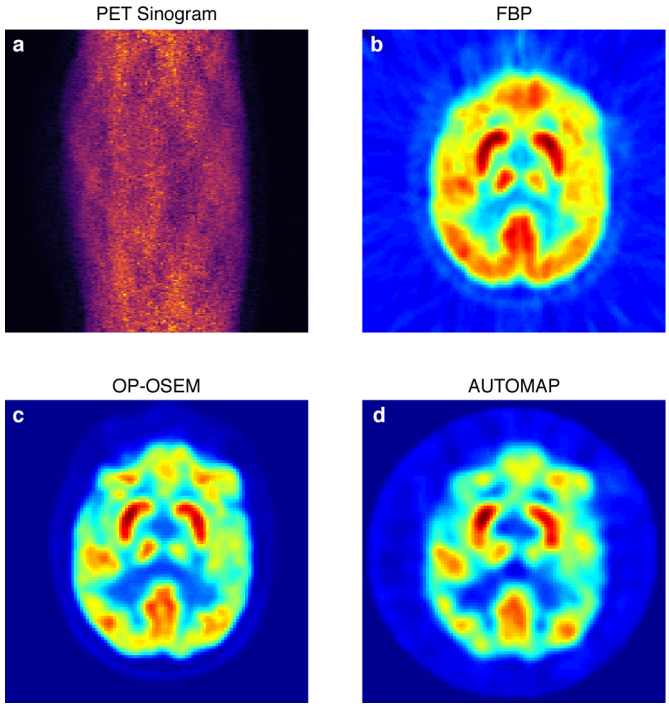
Extended Data Figure 2 | Effect of training corpus on image reconstruction. a–**c**, AUTOMAP was trained using sensor-image pairs of Cartesian Fourier encoded corpora derived from either ImageNet, HCP brain images, or random-valued Gaussian noise without any real-world image structure. Each trained network was then used to reconstruct a noise-corrupted Cartesian *k*-space brain dataset. The signal-to-noise ratio (SNR) of the reconstructed images is shown. The apparent intensity discontinuity in the region above the eyes is due to the masking process used to de-identify the data in the HCP protocol (see Methods for more details).



Extended Data Figure 3 | Training curves of optimizer loss convergence. Mean squared error (MSE) loss was minimized with stochastic gradient descent using the RMSProp algorithm and plotted here against training epoch count for: **a**, Cartesian Fourier encoding on IMAGENET corpus; **b**, spiral Fourier encoding on IMAGENET corpus; and **c**, Cartesian undersampled Fourier encoding on HCP brain corpus. The validation error tracks the training error without upward divergence, demonstrating a stable training regime with good bias-variance tradeoff.



Extended Data Figure 4 | Reconstruction of motion-corrupted MRI. **a**, T2-weighted reference image acquired at 3 T with a turbo spin-echo sequence. **b**, Three-dimensional motion trajectories measured during an Alzheimer's patient study. **c**, **d**, These motion trajectories were used to corrupt the *k*-space of this reference image, and it was reconstructed without motion compensation using inverse Fourier transform (**c**) and AUTOMAP (**d**). Both images show comparable artefact level and structure, demonstrating the stability of AUTOMAP reconstruction in the presence of unanticipated subject motion. A/P refers to anterior and posterior translational motion, L/R refers to left and right translational motion.



Extended Data Figure 5 | Reconstruction of PET scanner data. a–d, Human FDG PET sinogram data (**a**) was reconstructed using (**b**) filtered back projection (FBP), (**c**) OP-OSEM and (**d**) AUTOMAP.