ANATOMY IS THE MANIFOLD: LEARNING INTRINSIC METRICS INSTEAD OF IMAGES FOR PRIOR-GUIDED MRI RECONSTRUCTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Prior-guided multi-contrast MRI uses a fully sampled companion scan to reconstruct an undersampled target contrast. Networks that predict target intensities from the companion can transfer structures absent from the target and lose accuracy under distribution shift. Instead, we train a network to predict, from the companion, edge conductivities on a pixel graph, which act as an intrinsic metric. This metric defines a quadratic prior, and the target measurements form the only data term. As a result, zero measurements give a zero image, and the effect of a change in the companion can be written in closed form. We select among hand-crafted and learned metrics using held-out k-space, and show that under uniform random sampling the held-out residual estimates reconstruction error up to scale and noise. In synthetic lesion tests, learned metrics copy less of companion-only lesions than intensity priors do and recover more of target-only lesions. The learned metric is still sensitive to companion appearance and registration, but randomising both during training reduces the PSNR loss from a one-pixel misalignment to 2.0 dB (3.6–5.0 dB for intensity priors trained in the same way). In distribution, selecting among metrics comes within 0.6 dB of selecting among intensity priors. On each cohort with a scanner, pathology, contrast or anatomy shift, metric selection gains 0.6–2.3 dB over the strongest intensity baseline. On raw four-coil 0.3 T M4Raw data, the learned metric achieves higher PSNR than U-Net, MoDL and TGVN.

1 INTRODUCTION

Clinical MRI often acquires several contrasts of the same anatomy in one examination. A fully sampled contrast can therefore provide structural information for reconstructing another contrast from fewer measurements. Learned prior-guided methods exploit this redundancy and achieve high reconstruction accuracy when training and test data follow similar distributions (Xiang et al., 2019; Zhou & Zhou, 2020; Aggarwal et al., 2019; Atalık et al., 2026).

We ask how a fully sampled companion should influence the reconstruction of an undersampled target. Most learned methods use the companion to predict target *intensities*. Because undersampled measurements leave part of the image undetermined, such a network can introduce structure that the data do not support (Bhadra et al., 2021; Antun et al., 2020). Our experiments show three related problems (Figs. 1 and 3 and Table 1). First, when the scanner and pathology change (BraTS) or the anatomy changes (prostate), IXI-trained intensity priors fall below zero-filling, whereas a classical anatomically weighted prior with no learned parameters (Haldar et al., 2008; Mehranian et al., 2019) beats all seven of them by 2.2–7.5 dB (Table 2). Second, lesions added only to the companion partly appear in the reconstructed target. Third, a one-pixel companion misalignment reduces PSNR by 9–12 dB for learned models trained without randomisation, including ours.

Our approach. We let the companion set only the edge conductivities of a graph on the image grid. We interpret these conductivities as a discrete intrinsic metric of the anatomy. Each contrast is a function on this graph, and the reconstruction trades off data fit against smoothness under the metric. For fixed conductivities, it is the posterior mean of a Gaussian field. The companion affects only the precision matrix of this field, while the right-hand side of the resulting linear system depends on the

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

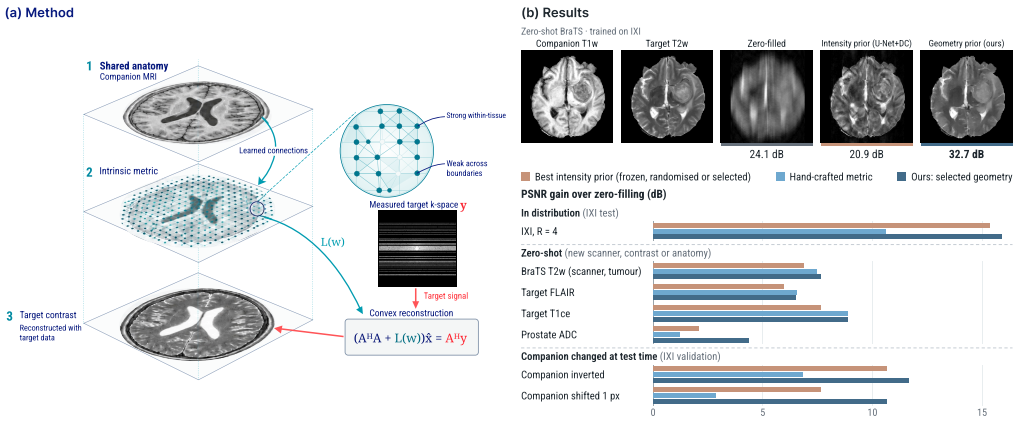

Figure 1: **Reconstruction with a companion-derived metric. (a)** The companion defines the graph conductivities and hence the precision $L(w)$, while the target measurements supply the right-hand side of the reconstruction system. The example uses raw M4Raw k-space with retrospective $R \approx 4$ undersampling. **(b)** Top: on a BraTS slice reconstructed without retraining, the IXI-trained intensity prior has lower PSNR than zero-filling, while held-out selection among our metrics (geometry selection) preserves more of the target contrast. Bottom: PSNR gain over zero-filling under seven test conditions. Orange shows the best evaluated intensity-prior result, including models trained with randomisation and held-out k-space selection. Light blue shows the training-free hand-crafted metric (κ=0.01); dark blue shows geometry selection, which chooses the hand-crafted metric with κ=0.02 for the slice above. Companion perturbations use IXI validation data; BraTS and prostate also differ in companion contrast. Full results are in Tables 1 and 2.

target measurements alone. The anatomically weighted prior mentioned above is the special case with hand-crafted conductivities. Because it is robust to distribution shift, we keep it as one of the candidates for test-time selection, alongside the learned metrics. We then study what this restriction guarantees, whether a learned metric is as robust as the hand-crafted one, and how to choose a metric at test time without a fully sampled reference (Fig. 2).

Contributions. (i) *A zero-measurement guarantee and a lesion probe.* We show that zero measurements yield a zero reconstruction and derive an exact expression for the effect of a companion change on the converged solution (Proposition 1). We also introduce a polarity-resolved probe for any companion-guided method. It inserts a lesion $\pm d$ into the companion and isolates the odd response, which reverses with lesion contrast. For the hand-crafted rule, we prove that the odd response vanishes when the companion is constant on the lesion and its one-pixel neighbourhood (Section B). Compared with MoDL, DuDoRNet and U-Net+DC, the learned metric reduces this response by 25–50% and recovers more of a target-only lesion. (ii) *Contrast and alignment randomisation.* The learned metric is sensitive to companion appearance and alignment. Randomising both during training makes it more robust. For intensity priors trained with the same augmentation, both the in-distribution cost and the loss under a one-pixel companion shift are larger (Section 4.1). (iii) *Held-out k-space selection.* Under uniformly random line sampling, the expected held-out residual ranks candidates by their reconstruction error outside the fit set (Proposition 2). We use this result to choose among hand-crafted, learned and randomised metrics (geometry selection) or among intensity priors (intensity selection). Without target-domain training, geometry selection has higher PSNR than the strongest intensity baseline on every shifted cohort (Table 2), although intensity selection is 0.5 dB better on 2D IXI data at R=8. (iv) *Volumetric and raw-data reconstruction.* The formulation supports Cartesian, radial NUFFT, volumetric and four-coil SENSE operators. The volumetric metric improves on the 2D intensity baselines when slices have complementary masks. On raw 0.3 T M4Raw data, the learned metric gives higher PSNR than U-Net, MoDL and TGVN.

2 Related work

Learned multi-contrast reconstruction. Prior-guided methods incorporate the companion through image- or dual-domain CNNs and transformers (Xiang et al., 2019; Zhou & Zhou, 2020; Feng et al., 2023; Chen et al., 2024), unrolled networks (Aggarwal et al., 2019; Sriram et al., 2020) or scan-specific neural fields (Shen et al., 2024). UniFS conditions on the sampling mask so that one model supports several Cartesian patterns (Li et al., 2025), and conditional score models use a joint multi-contrast image prior (Levac et al., 2023). Recent methods explicitly restrict the information transferred from the companion. MC-VarNet separates common and contrast-specific components (Lei et al., 2023). TGVN learns ambiguous-space consistency for raw multi-coil side information (Atalık et al., 2026). PnP-CoSMo separates shared content from contrast-specific style (Rao et al., 2026b), while CoSMo-RecNet freezes a pretrained content/style model and learns an unrolled refiner (Rao et al., 2026a). All four still use learned image updates, and we compare with each in Section 4.2. Learned reconstruction methods can be unstable (Antun et al., 2020), lose accuracy under distribution shift (Knoll et al., 2019; Darestani et al., 2021) and hallucinate structure (Bhadra et al., 2021). Test-time training can recover part of this loss (Darestani et al., 2022). We study a stricter restriction, where the companion only sets a graph precision, and test which of these failures persist.

Structure-guided and graph priors. Our metric builds on anatomically weighted quadratic priors (Haldar et al., 2008; Mehranian et al., 2019) and other structure-guided regularisers (Bowsher et al., 2004; Wang & Qi, 2015; Ehrhardt & Arridge, 2014; Lai et al., 2017), including directional TV for multi-contrast MRI (Ehrhardt & Betcke, 2016). Prior work also learns precisions, graph weights and regularisation maps by differentiating through reconstruction (Tappen et al., 2007; Zeng et al., 2019; de Lutio et al., 2022; Metzger et al., 2023; Kofler et al., 2023). We adopt this approach to learn an anatomy metric, but our focus is on how the reconstruction depends on the companion and target data, how robust it is to changes in appearance and alignment, and how to select metrics from held-out measurements.

Gaussian and intrinsic representations. Gaussian primitives have been used to represent scenes (Kerbl et al., 2023), X-ray and CT data (Cai et al., 2024; Zha et al., 2024) and MRI volumes (Peng et al., 2026). Heat Kernel Textures adapt primitives to surface geometry using the Laplace–Beltrami heat kernel (Foti et al., 2026). On our anatomy graph, a similar construction gives heat-kernel representations or a Gaussian field defined by its Laplacian precision. In the experiments of Section P, the precision formulation gives higher reconstruction accuracy than the tested sparse-primitive and dense heat-kernel representations.

Training through the solve, selection and randomisation. We use implicit differentiation through the reconstruction system, following work on differentiable optimisation and bilevel learning (Amos & Kolter, 2017; Agrawal et al., 2019; Gilton et al., 2021; Kunisch & Pock, 2013; Chen & Pock, 2017). Our selection rule draws on cross-validation and measurement-based risk estimation (Ward, 2009; Yaman et al., 2020; 2022; Ramani et al., 2012). We compare learned and classical priors using held-out k-space and characterise the unnormalised residual under random line sampling. We also randomise companion appearance and alignment during training (Tobin et al., 2017; Billot et al., 2023). In our experiments, this reduces sensitivity to companion changes and costs the geometry prior less in-distribution accuracy than the matched intensity priors.

3 Geometry-conditioned reconstruction

Problem. Let $y = Ax^\star + \epsilon$ denote undersampled k-space measurements of a complex target image $x^\star$. For single-coil Cartesian sampling, A applies the Fourier transform and sampling mask M. For SENSE, it also multiplies the image by coil sensitivities S; for non-Cartesian sampling, it uses a non-uniform fast Fourier transform (NUFFT). A fully sampled companion contrast g of the same anatomy is available. An *intensity-conditioned* reconstructor $\hat{x} = f_\theta(g, y)$ can use g to predict image components in the null space of A, where the measurements provide no constraint. We instead restrict the companion to determining the *geometry* of a quadratic prior.

Anatomy as a discrete manifold. We represent anatomy by a weighted 8-neighbour graph on the image grid, with conductivities $w_{ab} \geq 0$ and Laplacian $L(w) = D - W$. Here W contains the edge

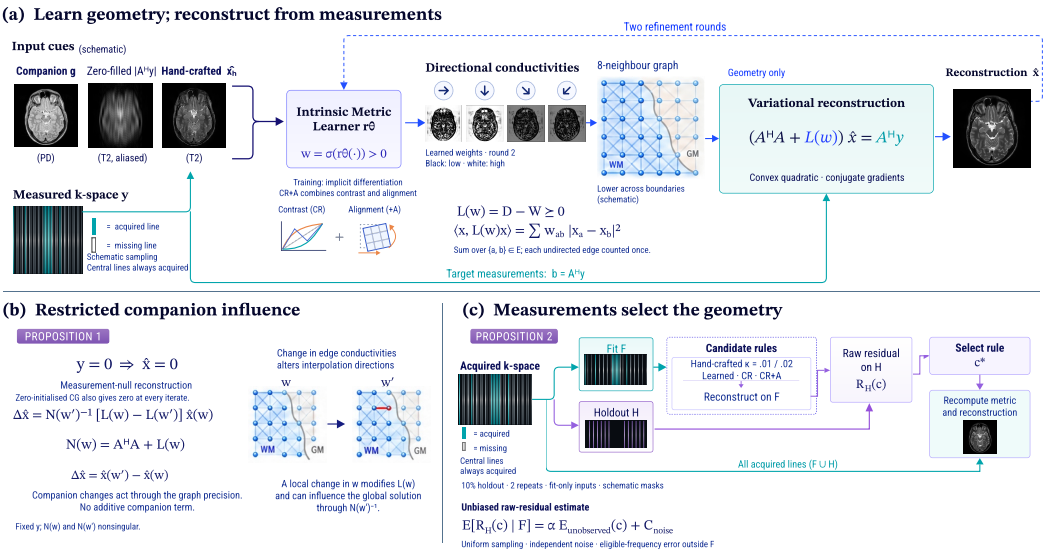

Figure 2: **Method overview.** (a) A network r_θ predicts positive conductivities w on an 8-neighbour anatomy graph in two refinement rounds. Reconstruction uses the system $(A^H A + L(w))\,\hat{x} = A^H y$, with only target measurements on the right-hand side. Training uses implicit differentiation and companion contrast (CR) and alignment (+A) randomisation. (b) Zero measurements yield a zero reconstruction; companion changes act through the graph precision (Proposition 1). (c) Candidates are compared using held-out acquired lines (Proposition 2), and the selected rule is recomputed using all acquired data.

weights and D their row sums on its diagonal. This Laplacian can be viewed as a discrete analogue of an anisotropic Laplace–Beltrami operator (Pang & Cheung, 2017), although the learned weights need not correspond to any smooth continuum tensor. We treat each contrast as a function on the graph and reconstruct it by

$$\hat{x}(w) = \arg\min_x \; \|Ax - y\|_2^2 + \langle x, L(w)\, x\rangle \quad \Longleftrightarrow \quad N(w)\,\hat{x} = A^H y, \qquad N(w) = A^H A + L(w), \tag{1}$$

using conjugate gradients (CG). For fixed w, this is the posterior mean under an intrinsic Gaussian Markov random-field prior with precision $L(w)$ and the Gaussian likelihood corresponding to the data term. Correlations under this prior therefore follow the graph. For example, the associated heat kernel $e^{-tL(w)}\delta_p$ spreads a source at pixel p according to the anatomy metric (Section P). We impose a small positive lower bound on the conductivities so that the graph is connected. The only null mode of $L(w)$ is then the constant image, so $N(w)$ is nonsingular whenever A measures this mode, as ensured by the retained DC coefficient in our single-coil Cartesian masks. Following classical anatomically constrained reconstruction (Haldar et al., 2008; Mehranian et al., 2019), we define the hand-crafted metric

$$w_{ab}^{\text{hand}} = \frac{\lambda_h}{|e_{ab}|^2} \exp\!\big(-|g_a - g_b|/\kappa\big), \tag{2}$$

where $|e_{ab}|$ is the distance between the two pixels. This rule reduces smoothing across companion boundaries while retaining smoothing along them. Its hyperparameters are the edge bandwidth κ and global prior precision λ_h. Because the conductivities are defined on pixel pairs, they place barriers between neighbouring pixels. A per-pixel structure-tensor implementation that we also tested suppresses smoothing across a band of boundary pixels instead, and gives lower accuracy (Section Q).

Properties of the reconstruction. The following holds for any linear forward operator.

Proposition 1 (Measurement-null reconstruction and companion response)**.** *For fixed conductivities $w \geq 0$ with $N(w)$ nonsingular, $\hat{x}(w) = N(w)^{-1} A^H y$ is linear in y, contains no additive companion term and satisfies $A^H(y - A\hat{x}) = L(w)\,\hat{x}$ with $L(w) \succeq 0$. If the rule producing w also depends on the target measurements, the complete map may be nonlinear. Nevertheless, $y = 0$ gives $\hat{x} = 0$, as*

does CG initialised at zero. For a companion change $w \to w'$ with y fixed and $N(w')$ nonsingular,

$$\hat{x}(w') - \hat{x}(w) = N(w')^{-1} r, \qquad r_a = \sum_b (w_{ab} - w'_{ab})\big(\hat{x}_a(w) - \hat{x}_b(w)\big).$$

If $\hat{x}(w)$ is flat across all changed edges, then $r = 0$ and $\hat{x}(w') = \hat{x}(w)$ for a change of any size.

Every output is thus the exact Tikhonov solution for some anatomy metric, and this can be checked per pixel up to the CG residual. It cannot contain an additive image independent of y, but a boundary present only in the companion can still change how measured information is interpolated and distort the target (Section 4.3). To study this, we insert a lesion $\pm d$ with both polarities and decompose the reconstruction changes $\Delta_\pm$ into an odd component $\frac{1}{2}(\Delta_+ - \Delta_-)$, which reverses with lesion contrast, and an even component $\frac{1}{2}(\Delta_+ + \Delta_-)$. For the hand-crafted rule (2), the odd part vanishes when the companion is constant around the lesion support (Section B). On heterogeneous tissue, interpolation changes can also enter the odd response, so the probe does not exactly separate copying from interpolation.

Learning the metric. We use a U-Net r_θ to predict four conductivity maps for the 8-neighbour graph. Its inputs are the companion, the zero-filled target $|A^H y|$ and the hand-crafted reconstruction. The output $w_\theta = \sigma(r_\theta(\cdot))$ uses a positive activation σ, subject to the conductivity floor above. We train on fully sampled pairs by minimising $\ell = \|\hat{x}(w_\theta) - x^\star\|^2$. Target measurements can influence the predicted weights, but the network still outputs only conductivities. For the converged solve, let $u = N(w)^{-1} \nabla_{\hat{x}} \ell$. Implicit differentiation gives

$$\partial \ell / \partial w_{ab} = -\operatorname{Re}[(u_a - u_b)^*(\hat{x}_a - \hat{x}_b)], \tag{3}$$

which requires one adjoint CG solve per reconstruction round and supports the NUFFT and multi-coil SENSE operators used here. We use two refinement rounds, feeding the first reconstruction back to the network before the second solve. In practice, we run CG for a fixed number of iterations. The implicit gradient then only approximates the derivative of this truncated solve, and the zero initial guess preserves the zero-measurement property (Section O). For volumes, we add one through-plane edge per voxel, with weights predicted from the companion by a small CNN, and reconstruct the volume jointly. Different masks across slices provide complementary k-space measurements that the through-plane prior can combine.

Contrast-randomised training. A network that predicts conductivities can still overfit companion *appearance*. For example, inverting the foreground of an IXI companion pushes the learned metric below zero-filling but leaves the hand-crafted metric unchanged (Section 4.1). We therefore randomise companion contrast during training. With probability 0.8, we apply a piecewise-linear intensity map T that keeps the background dark and is non-monotone or inverted with probability one half. The training objective is

$$\mathbb{E}_{T,\,M,\,\epsilon} \left\| \hat{x}\big(w_\theta(Tg,\, y)\big) - x^\star \right\|^2, \tag{4}$$

where $w_\theta(Tg, y)$ is predicted from the randomised companion Tg, $|A^H y|$ and the hand-crafted reconstruction built from Tg. The map T changes intensities without spatially warping the companion, although it can alter edge strengths or suppress a boundary by merging intensities. The reconstruction must still fit the same target measurements. We also test target-contrast randomisation and rigid companion motion. Variants are selected on IXI validation (Section C), and the same augmentation is applied to the intensity priors in Section 4.1.

Selecting geometry from held-out k-space. We compare candidate reconstruction rules c by how well they predict held-out measurements. In each of two repeats with fixed random seeds, we hold out 10% of the acquired phase-encoding lines outside the calibration band, denoted by $\mathcal{H}_r$, and reconstruct from the remaining lines $\mathcal{F}_r$. We select

$$c^* = \arg\min_{c \in \mathcal{C}} \; \frac{1}{2} \sum_{r=1}^{2} \frac{\|P_{\mathcal{H}_r}(A\hat{x}_{\mathcal{F}_r}(c) - y)\|_2^2}{\|P_{\mathcal{H}_r} y\|_2^2}; \tag{5}$$

and recompute the selected metric and reconstruction using all acquired lines. Within each repeat, learned metrics are predicted using only the fit data and companion, excluding the held-out target measurements. Selection requires neither ground truth nor target-domain training. Because every candidate satisfies the zero-measurement property, the selected reconstruction does too. The unnormalised residual has the following interpretation.

Proposition 2 (Risk estimation from held-out k-space)**.** *Suppose the non-calibration lines are drawn uniformly without replacement from a pool containing the eligible lines $\mathcal{E}$, and $\mathcal{H}$ is a uniformly random subset of h acquired eligible lines. Assume zero-mean noise with variance σ^2, independent across k-space samples and independent of the companion g, and that each candidate uses only g and the fit measurements $y_{\mathcal{F}}$. Conditional on the fit set $\mathcal{F}$, its measurements and the companion,*

$$\mathbb{E}\big[\|P_{\mathcal{H}}(A\hat{x}_{\mathcal{F}}(c) - y)\|_2^2 \,\big|\, \mathcal{F}, y_{\mathcal{F}}, g\big] = \frac{h}{|\mathcal{E} \setminus \mathcal{F}|} \big\|P_{\mathcal{E}\setminus\mathcal{F}}\, A_{\text{full}}\big(\hat{x}_{\mathcal{F}}(c) - x^\star\big)\big\|_2^2 + h\, n_{\text{s}}\, \sigma^2,$$

where A_{full} is the fully sampled operator and n_{s} the number of samples per line.

Thus, after accounting for a common scale and noise term, the unnormalised residual estimates error on eligible frequencies outside the fit set. Most of these frequencies are unacquired. The normaliser in (5) is shared by all candidates within a repeat and does not change their ranking there. Averaging the normalised residuals across repeats is a practical choice. The proposition does not show that this averaged score is unbiased, either for the fit-set reconstructions or for the selected reconstruction after refitting. The proof relies on the exchangeability of acquired and unacquired lines (Section B). Under variable-density sampling, each held-out residual would need to be weighted by the inverse of its conditional inclusion probability. Our five 2D candidates are two hand-crafted metrics (κ=0.01, 0.02) and learned metrics trained without randomisation, with contrast randomisation, or with contrast and alignment randomisation. The repeat count was chosen on IXI validation (Section E).

4 Experiments

Protocol. We train on **IXI** PD→T2, with 540/28/9 training/validation/test subjects (Imperial College London, 2008). For evaluation without retraining, **BraTS/MSD** provides 15 subjects with a T1w companion and co-registered T2w, FLAIR and T1ce targets, which differ from IXI in scanner and pathology. **MSD prostate** T2→ADC provides 8 test subjects with a different anatomy and contrast pair (Antonelli et al., 2022). For these cohorts, k-space is simulated from the real, normalised 256×256 target magnitude image by a centred orthonormal DFT, so its noise-free spectrum is Hermitian. We add complex Gaussian noise (σ=0.01, a fully sampled image SNR of 33–55) and use random column masks with a 4% calibration band and training acceleration R=8. A separate **M4Raw** experiment uses raw four-coil 0.3 T k-space and 128/30/25 training/validation/test subjects (Lyu et al., 2023). Its SENSE sensitivities and scaling are estimated from 16 retained calibration lines. PSNR (peak 1) and SSIM are computed on magnitude images and averaged first over slices within each subject, then over subjects. Confidence intervals are computed from 10,000 subject-level bootstrap resamples, and Section A gives the preprocessing for each cohort.

Comparators and model settings. Baselines include zero-filling, compressed-sensing TV (Lustig et al., 2007), directional TV (dTV) tuned on IXI validation (Ehrhardt & Betcke, 2016), a NeRP-style per-scan neural field (Shen et al., 2024), U-Net with hard data consistency (U-Net+DC), prior-conditioned MoDL (Aggarwal et al., 2019), DuDoRNet (Zhou & Zhou, 2020), TGVN, MC-VarNet, PnP-CoSMo and CoSMo-RecNet (Atalık et al., 2026; Lei et al., 2023; Rao et al., 2026b;a). All learned baselines use the same IXI split, masks and noise as our model. U-Net, MoDL and TGVN are also trained on raw M4Raw (Section 4.4). The hand-crafted metric uses (κ, λ_h)=$(0.01, 0.3)$ for normalised magnitude data and $(0.03, 0.1)$ for raw M4Raw. Our metric network is a four-level U-Net with 12.5 M parameters and two refinement rounds.

4.1 Sensitivity to companion appearance and alignment

We perturb only the companion, using IXI validation data (Table 1).

Appearance. Because (2) depends only on absolute intensity differences, inverting the companion leaves the hand-crafted metric at 27.4 dB. The learned metric drops from 34.7 to 19.7 dB, below zero-filling, and the intensity priors fall to 15.3–16.1 dB, even though the companion is still aligned. Contrast randomisation (4) improves PSNR under inversion for all three model families trained with it, and the randomised metric reaches 32.3 dB. With the original companion, randomisation costs the metric 0.9 dB, U-Net+DC 1.1 dB and MoDL 2.8 dB.

Table 1: **Sensitivity to companion appearance and alignment** (IXI validation, 28 subjects × 3 slices, R=8; PSNR). Only the companion changes at test time. Underlined values are below zero-filling (20.59 dB). Alignment randomisation uses shifts up to 2 px and rotations up to 3°; the last two columns exceed this range. Randomised rows perturb only the companion during training. The intensity selector also includes the four companion-and-target variants in Table 6.

		appearance		misalignment			larger	
Companion at test time	exact	inverted	random map	1 px	2 px	3°	4 px	6°
U-Net+DC	35.35	16.11	27.18	23.43	20.29	22.09	18.82	20.07
+ contrast rand., companion	34.22	31.24	31.02	24.11	21.46	22.93	20.98	21.51
+ contrast & alignment rand., companion	33.21	30.23	30.32	28.21	26.65	26.77	24.04	25.21
MoDL	35.02	15.30	27.33	23.72	20.49	22.59	18.95	20.29
+ contrast rand., companion	32.25	30.49	29.82	25.40	23.30	24.22	22.03	22.76
+ contrast & alignment rand., companion	31.34	29.34	29.14	27.72	25.86	26.19	23.46	24.69
DuDoRNet	34.17	15.90	27.32	23.99	20.88	22.74	19.32	20.63
Held-out selection, 10 intensity models	35.23	31.17	30.82	28.00	26.50	26.65	23.95	25.05
Hand-crafted metric	27.41	27.41	25.50	23.42	22.08	22.72	20.92	21.52
Learned metric	34.68	19.72	29.38	25.38	22.53	24.17	21.32	22.21
+ contrast rand., companion	33.76	32.30	32.29	27.25	25.01	25.64	23.51	24.10
+ contrast & alignment rand., companion	33.26	31.68	31.86	31.22	30.08	30.02	26.70	27.50
Held-out selection, 5 metrics	34.65	32.25	32.21	31.22	30.08	30.02	26.70	27.50

Table 2: **Generalisation across datasets and contrasts** (PSNR). Models are trained on IXI PD→T2 at R=8 and tested without retraining. ZF: zero-filling. BraTS also changes the companion to T1w; prostate changes anatomy, scanner and both contrasts. *Frozen*: highest PSNR among U-Net+DC, MoDL, DuDoRNet, TGVN, MC-VarNet, PnP-CoSMo and CoSMo-RecNet, or the learned metric, without randomisation. *Randomised*: highest PSNR among eight randomised intensity models or two randomised metrics. These two columns report the best model separately for each cohort. *Selected*: per-slice held-out selection among ten intensity models (Table 6) or five metrics (Table 8). Bold indicates the best result in each row.

			intensity priors			geometry priors (ours)			
Test-time change	ZF	dTV	frozen	randomised	selected	hand-crafted	frozen	randomised	selected
IXI, R=8 (no change)	20.85	29.54	35.54	34.59	**35.58**	27.91	35.04	34.25	35.04
IXI, R=4	21.76	31.86	37.11	35.26	37.06	32.37	**37.66**	37.10	**37.66**
BraTS T2w (scanner, tumour)	24.26	30.56	24.18	31.07	31.10	31.72	27.15	**32.32**	31.92
BraTS FLAIR (+ target contrast)	24.88	30.42	27.51	30.81	30.76	**31.43**	30.10	30.55	31.39
BraTS T1ce (+ target contrast)	24.66	32.03	29.21	32.29	32.27	33.52	32.87	32.14	**33.53**
Prostate ADC (anatomy)	21.30	22.92	20.33	23.36	23.12	22.49	23.37	**25.64**	**25.64**

Alignment. A one-pixel diagonal companion shift reduces PSNR by 9–12 dB for the learned models trained without randomisation. In a graph prior, a displaced companion boundary can suppress smoothing at the wrong location. Adding rigid companion motion during training reduces this sensitivity. The randomised metric reaches 31.2 dB at 1 px and 30.1 dB at 2 px, compared with 33.3 dB when aligned. With the same augmentation, U-Net+DC and MoDL reach 28.2 and 27.7 dB at one pixel, which is 5.0 and 3.6 dB below their aligned results. Our randomised metric loses only 2.0 dB. Beyond the training range, at 4 px and 6°, the randomised metric reaches 26.7 and 27.5 dB, more than 2.2 dB above the strongest intensity candidate. We suspect that the conductivity network learns to discount companion boundaries that disagree with target measurements. Geometry selection chooses a randomised metric on 94–100% of perturbed slices (Table 9) and achieves 1.09–3.58 dB higher PSNR than intensity selection, with positive differences for all 28 subjects.

4.2 Generalisation across datasets and contrasts

All models in Table 2 use their IXI checkpoints without retraining. U-Net+DC, MoDL, DuDoRNet, TGVN, MC-VarNet, PnP-CoSMo and CoSMo-RecNet all have lower PSNR than zero-filling on BraTS T2w and prostate (Tables 7, 15 and 16). For example, TGVN reaches 23.59 dB on T2w and 19.67 dB on prostate, compared with 24.26 and 21.30 dB for zero-filling. The training-free hand-crafted metric exceeds all intensity priors on BraTS by 0.62–1.23 dB.

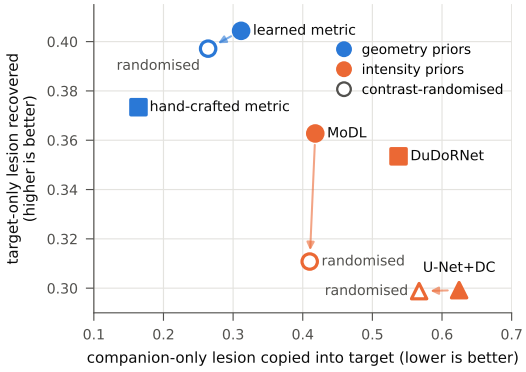

Figure 3: **Lesion recovery and companion response** ($\sigma = 1.5$ pixels, R=8; 28 IXI validation subjects). Arrows indicate companion contrast randomisation (Table 10).

Table 3: **Raw four-coil 0.3 T M4Raw** (25 test subjects $\times$ 3 slices; PSNR / SSIM). All models share the SENSE operator and calibration-derived sensitivities; learned models also share the training split and epoch count. The learned metric's subject-paired PSNR gains over the next-best method are 0.24 dB at R=4 (MoDL; 95% CI $[0.18, 0.30]$) and 0.28 dB at R=8 (U-Net; $[0.20, 0.35]$).

Method	R=4	R=8
Zero-filled	21.51 / 0.496	20.50 / 0.405
Directional TV (SENSE)	22.75 / 0.465	21.19 / 0.403
Hand-crafted metric	22.96 / 0.535	21.10 / 0.427
U-Net	25.11 / 0.565	23.84 / 0.514
MoDL	25.70 / 0.610	23.64 / 0.525
TGVN	25.16 / **0.629**	23.29 / **0.540**
Learned metric (ours)	**25.94** / 0.607	**24.12** / 0.527

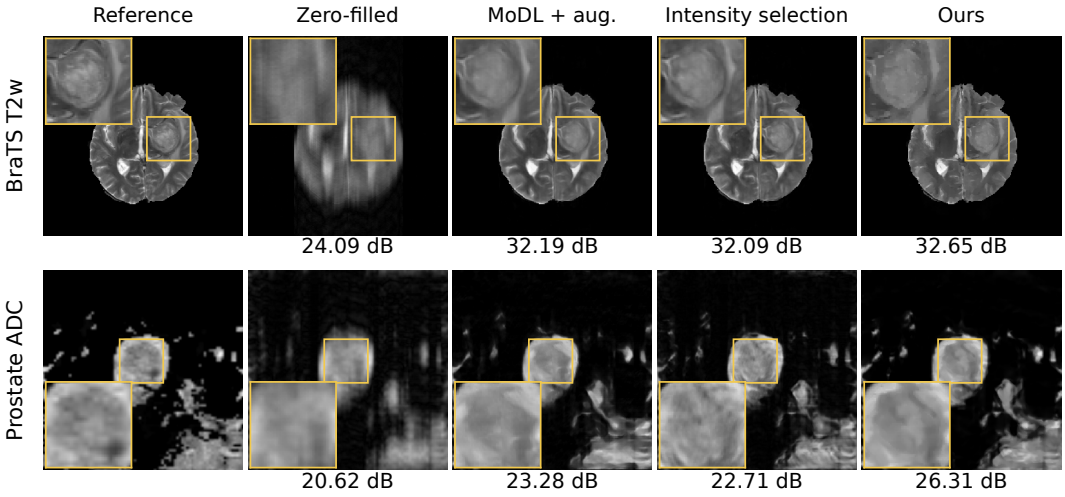

Figure 4: **Zero-shot reconstruction at R=8.** MoDL uses contrast and alignment augmentation. Both selectors use held-out k-space. Insets share a display window. Numbers are full-slice PSNR.

Randomisation substantially improves transfer for both model families. The best randomised intensity prior reaches 31.07 dB on BraTS T2w and 23.36 dB on prostate. For the learned metric, contrast randomisation raises T2w PSNR from 27.15 to 32.32 dB but reduces T1ce PSNR from 32.87 to 32.14 dB. Adding alignment randomisation gives 25.64 dB on prostate. No single augmentation variant is best across all cohorts.

In distribution, held-out selection among the five metrics picks the learned metric on 98% and 93% of IXI slices at R=8 and R=4 and matches its PSNR (35.04 and 37.66 dB). On the shifted cohorts, it exceeds the strongest intensity baseline by 0.82 dB on T2w, 0.59 dB on FLAIR, 1.24 dB on T1ce and 2.27 dB on prostate; all subject-paired 95% confidence intervals are above zero (Section H). This baseline is intensity selection on T2w and, on the other cohorts, MoDL trained with companion and target contrast randomisation and companion alignment randomisation. Figure 4 compares reconstructions on T2w and prostate. In distribution, intensity selection is 0.54 dB better at R=8, while geometry selection is 0.59 dB better at R=4. Selection is not perfect under shift. On BraTS, its mean PSNR regret relative to the per-slice oracle is 0.3–0.7 dB (Table 9).

4.3 Responses to target-only and companion-only lesions

Figure 3 reports the polarity-resolved probe defined in Section 3. For the smallest companion-only lesions, the odd response projected onto the inserted lesion (the copied fraction) is 62%, 54% and 42% of its amplitude for U-Net+DC, DuDoRNet and MoDL, compared with 31% for the learned metric and 16% for the hand-crafted metric. The learned metric also recovers the largest fraction of target-only lesions: 40%, 60% and 82% at $\sigma = 1.5, 3, 6$, respectively. Companion contrast randomisation reduces the copied fraction for all tested models. For the smallest lesion, randomisation also lowers MoDL's recovered target fraction from 0.36 to 0.31, but the randomised metric still recovers 0.40 of the lesion and its copied fraction falls to 0.26. In a separate test using another subject's companion, U-Net+DC falls 2 dB below zero-filling, while the learned metric has similar PSNR to zero-filling (Table 11).

4.4 Volumetric, radial and raw-data reconstruction

Through-plane edges couple slices in a joint reconstruction. On all 388 central IXI test slices with independent random masks at R=8, the volumetric metric reaches 36.16 dB: 0.86 dB above U-Net+DC, 0.94 dB above 2D MoDL and 1.59 dB above MoDL trained in 3D. Three independent training seeds give 36.13 ± 0.03 dB. To separate through-plane regularisation from complementary sampling, we also use one random mask shared across slices. Gains over the slice-wise metric (0.79 dB) and MoDL-3D (1.38 dB) hold for all nine subjects, with 95% confidence intervals above zero. Differences from U-Net+DC and 2D MoDL are no longer statistically significant (Section K). In a volumetric version of the lesion probe, the 3D metric recovers more of target-only lesions and responds less to companion-only lesions than both the 2D metric and MoDL-3D (Table 22). On the central 50% of each BraTS volume, selecting between hand-crafted and learned 3D metrics gives 33.56, 33.33 and 34.93 dB for T2w, FLAIR and T1ce without retraining. MoDL-3D reaches only 23.95 dB on T2w (Table 18).

Radial sampling. The metric network trained on Cartesian masks also works with a golden-angle radial NUFFT operator without retraining. On IXI validation, it reaches 32.2, 35.3 and 38.7 dB with 16, 32 and 64 spokes, at least 3.4 dB above the other methods in Table 14.

Raw multi-coil data. The learned metric has the highest PSNR at both accelerations and exceeds TGVN by 0.78 and 0.83 dB at R=4 and R=8, with positive differences for all 25 subjects (Table 3 and Section L). The margin over the next-best method is smaller: 0.24 dB over MoDL at R=4 and 0.28 dB over U-Net at R=8. TGVN has the highest SSIM at both accelerations, and MoDL's exceeds ours by 0.003 at R=4.

Limitations. The zero-measurement guarantee does not ensure recovery of structures missing from the measurements or exclude artifacts caused by incorrect companion boundaries. The companion-response expression in Proposition 1 is exact only at convergence, but our solver stops CG early. At 129 ms per slice on one GPU, the learned metric is also slower than the intensity priors (8–63 ms; Table 25), and held-out selection adds one reconstruction per candidate and repeat. Alignment randomisation covers small in-plane rigid motion, and larger perturbations still reduce the randomised metric's PSNR by 5.8–6.6 dB. Under stronger noise than in training (σ=0.05), MoDL-3D has 0.56 dB higher PSNR than the volumetric metric, although the metric keeps a higher SSIM (Section T). Except for M4Raw, k-space is simulated from magnitude images under a single-coil model. The test cohorts are small (9 IXI, 15 BraTS and 8 prostate subjects), and the raw-data experiment uses retrospective low-field acquisitions. Finally, Proposition 2 assumes uniform random line sampling.

5 Conclusion

In our experiments, restricting the companion to the geometry of the prior, rather than using it to predict target intensities, reduced lesion copying and improved accuracy when test data differed from the training data. The restriction alone was not enough: the learned metric was also sensitive to companion appearance and alignment, and the gains under shift relied on training randomisation and held-out selection. Future work could estimate registration jointly with the metric and calibrate the Gaussian posterior uncertainty, which underestimates errors in our tests (Section S).

AI USE STATEMENT

Generative AI tools were used primarily to improve the language and readability of the manuscript and to discuss and refine research ideas. The authors evaluated the suggestions and reviewed all AI-assisted revisions. The authors take responsibility for the final content of the paper.

ETHICS STATEMENT

This study uses publicly released research datasets under their respective terms and involves no new data collection. The reconstructions have not been validated for clinical use. The lesion and companion stress tests examine failure modes that may not be apparent from aggregate image-quality metrics.

REPRODUCIBILITY STATEMENT

The appendix describes the data splits, sampling masks, architectures, training schedules, solver settings, checkpoint selection and statistical procedures. Model variants and selection rules were chosen on validation data before test evaluation. All tables are generated from per-case results by scripts. We will release code, trained checkpoints and evaluation scripts.

REFERENCES

Hemant K Aggarwal, Merry P Mani, and Mathews Jacob. MoDL: Model-based deep learning architecture for inverse problems. *IEEE Transactions on Medical Imaging*, 38(2):394–405, 2019. doi: 10.1109/TMI.2018.2865356. 1, 3, 6

Akshay Agrawal, Brandon Amos, Shane Barratt, Stephen Boyd, Steven Diamond, and J. Zico Kolter. Differentiable convex optimization layers. In *Advances in Neural Information Processing Systems*, volume 32, 2019. 3

Brandon Amos and J. Zico Kolter. OptNet: Differentiable optimization as a layer in neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 136–145. PMLR, 2017. 3

Michela Antonelli et al. The medical segmentation decathlon. *Nature Communications*, 13(1):4128, 2022. doi: 10.1038/s41467-022-30695-9. 6, 14

Vegard Antun, Francesco Renna, Clarice Poon, Ben Adcock, and Anders C. Hansen. On instabilities of deep learning in image reconstruction and the potential costs of AI. *Proceedings of the National Academy of Sciences*, 117(48):30088–30095, 2020. doi: 10.1073/pnas.1907377117. 1, 3

Arda Atalık, Sumit Chopra, and Daniel K. Sodickson. A trust-guided approach to MR image reconstruction with side information. *IEEE Transactions on Medical Imaging*, 45(1):190–205, 2026. doi: 10.1109/TMI.2025.3594363. 1, 3, 6, 17, 28

Sayantan Bhadra, Varun A Kelkar, Frank J Brooks, and Mark A Anastasio. On hallucinations in tomographic image reconstruction. *IEEE Transactions on Medical Imaging*, 40(11):3249–3260, 2021. doi: 10.1109/TMI.2021.3077857. 1, 3

Benjamin Billot, Douglas N. Greve, Oula Puonti, Axel Thielscher, Koen Van Leemput, Bruce Fischl, Adrian V. Dalca, and Juan Eugenio Iglesias. SynthSeg: Segmentation of brain MRI scans of any contrast and resolution without retraining. *Medical Image Analysis*, 86:102789, 2023. doi: 10.1016/j.media.2023.102789. 3

James E. Bowsher, Hong Yuan, Laurence W. Hedlund, Timothy G. Turkington, Gamal Akabani, Alexandra Badea, William C. Kurylo, Charles T. Wheeler, Gary P. Cofer, Mark W. Dewhirst, and G. Allan Johnson. Utilizing MRI information to estimate F18-FDG distributions in rat flank tumors. In *2004 IEEE Nuclear Science Symposium Conference Record*, volume 4, pp. 2488–2492, 2004. doi: 10.1109/NSSMIC.2004.1462760. 3

Yuanhao Cai, Yixun Liang, Jiahao Wang, Angtian Wang, Yulun Zhang, Xiaokang Yang, Zongwei Zhou, and Alan Yuille. Radiative Gaussian splatting for efficient X-ray novel view synthesis. In *European Conference on Computer Vision (ECCV)*, pp. 283–299, 2024. doi: 10.1007/978-3-031-73232-4_16. 3

Qi Chen, Xiaohan Xing, Zhen Chen, and Zhiwei Xiong. Accelerated multi-contrast MRI reconstruction via frequency and spatial mutual learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 56–66, 2024. doi: 10.1007/978-3-031-72104-5_6. 3

Yunjin Chen and Thomas Pock. Trainable nonlinear reaction diffusion: A flexible framework for fast and effective image restoration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1256–1272, 2017. doi: 10.1109/TPAMI.2016.2596743. 3

Mohammad Zalbagi Darestani, Akshay S. Chaudhari, and Reinhard Heckel. Measuring robustness in deep learning based compressive sensing. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *PMLR*, pp. 2433–2444, 2021. 3

Mohammad Zalbagi Darestani, Jiayu Liu, and Reinhard Heckel. Test-time training can close the natural distribution shift performance gap in deep learning based compressed sensing. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, volume 162 of *PMLR*, pp. 4754–4776, 2022. 3

Riccardo de Lutio, Alexander Becker, Stefano D'Aronco, Stefania Russo, Jan D. Wegner, and Konrad Schindler. Learning graph regularisation for guided super-resolution. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. doi: 10.1109/CVPR52688.2022.00202. 3, 16

Matthias J. Ehrhardt and Simon R. Arridge. Vector-valued image processing by parallel level sets. *IEEE Transactions on Image Processing*, 23(1):9–18, 2014. doi: 10.1109/TIP.2013.2277775. 3

Matthias J. Ehrhardt and Marta M. Betcke. Multicontrast MRI reconstruction with structure-guided total variation. *SIAM Journal on Imaging Sciences*, 9(3):1084–1106, 2016. doi: 10.1137/15M1047325. 3, 6

Chun-Mei Feng, Yunlu Yan, Geng Chen, Yong Xu, Ying Hu, Ling Shao, and Huazhu Fu. Multi-modal transformer for accelerated MR imaging. *IEEE Transactions on Medical Imaging*, 42(10):2804–2816, 2023. doi: 10.1109/TMI.2022.3180228. 3

Simone Foti, Caner Korkmaz, Stefanos Zafeiriou, and Tolga Birdal. Heat kernel textures: the geodesic Gaussians that do not splat. In *European Conference on Computer Vision (ECCV)*, pp. 306–323, 2026. doi: 10.1007/978-3-032-37595-7_17. 3, 30

Davis Gilton, Gregory Ongie, and Rebecca Willett. Deep equilibrium architectures for inverse problems in imaging. *IEEE Transactions on Computational Imaging*, 7:1123–1133, 2021. doi: 10.1109/TCI.2021.3118944. 3

Justin P. Haldar, Diego Hernando, Sheng-Kwei Song, and Zhi-Pei Liang. Anatomically constrained reconstruction from noisy data. *Magnetic Resonance in Medicine*, 59(4):810–818, 2008. doi: 10.1002/mrm.21536. 1, 3, 4

Imperial College London. IXI dataset, 2008. https://brain-development.org/ixi-dataset/ (accessed 2026-09-14). 6

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D Gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4):139:1–139:14, 2023. doi: 10.1145/3592433. 3

Florian Knoll, Kerstin Hammernik, Erich Kobler, Thomas Pock, Michael P. Recht, and Daniel K. Sodickson. Assessment of the generalization of learned image reconstruction and the potential for transfer learning. *Magnetic Resonance in Medicine*, 81(1):116–128, 2019. doi: 10.1002/mrm.27355. 3

Andreas Kofler, Fabian Altekrüger, Fatima Antarou Ba, Christoph Kolbitsch, Evangelos Papoutsellis, David Schote, Clemens Sirotenko, Felix Frederik Zimmermann, and Kostas Papafitsoros. Learning regularization parameter-maps for variational image reconstruction using deep neural networks and algorithm unrolling. *SIAM Journal on Imaging Sciences*, 16(4):2202–2246, 2023. doi: 10.1137/23M1552486. 3

Karl Kunisch and Thomas Pock. A bilevel optimization approach for parameter learning in variational models. *SIAM Journal on Imaging Sciences*, 6(2):938–983, 2013. doi: 10.1137/120882706. 3

Zongying Lai, Xiaobo Qu, Hengfa Lu, Xi Peng, Di Guo, Yu Yang, Gang Guo, and Zhong Chen. Sparse MRI reconstruction using multi-contrast image guided graph representation. *Magnetic Resonance Imaging*, 43:95–104, 2017. doi: 10.1016/j.mri.2017.07.009. 3

Pengcheng Lei, Faming Fang, Guixu Zhang, and Tieyong Zeng. Decomposition-based variational network for multi-contrast MRI super-resolution and reconstruction. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. doi: 10.1109/ICCV51070.2023.01947. 3, 6, 17

Brett Levac, Ajil Jalal, Kannan Ramchandran, and Jonathan I. Tamir. MRI reconstruction with side information using diffusion models. In *57th Asilomar Conference on Signals, Systems, and Computers*, pp. 1436–1442, 2023. doi: 10.1109/IEEECONF59524.2023.10476807. 3

Jialin Li, Yiwei Ren, Kai Pan, Dong Wei, Pujin Cheng, Xian Wu, and Xiaoying Tang. UniFS: Unified multi-contrast MRI reconstruction via frequency-spatial fusion. In *2025 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 1–6, 2025. doi: 10.1109/BIBM66473.2025.11356472. 3

Michael Lustig, David Donoho, and John M Pauly. Sparse MRI: The application of compressed sensing for rapid MR imaging. *Magnetic Resonance in Medicine*, 58(6):1182–1195, 2007. doi: 10.1002/mrm.21391. 6

Mengye Lyu, Lifeng Mei, Shoujin Huang, Sixing Liu, Yi Li, Kexin Yang, Yilong Liu, Yu Dong, Linzheng Dong, and Ed X Wu. M4Raw: a multi-contrast, multi-repetition, multi-channel MRI k-space dataset for low-field MRI research. *Scientific Data*, 10(1):264, 2023. doi: 10.1038/s41597-023-02181-4. 6

Abolfazl Mehranian, Martin A. Belzunce, Colm J. McGinnity, Aurelien Bustin, Claudia Prieto, Alexander Hammers, and Andrew J. Reader. Multi-modal synergistic PET and MR reconstruction using mutually weighted quadratic priors. *Magnetic Resonance in Medicine*, 81(3):2120–2134, 2019. doi: 10.1002/mrm.27521. 1, 3, 4

Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Guided depth super-resolution by deep anisotropic diffusion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18237–18246, 2023. doi: 10.1109/CVPR52729.2023.01749. 3

Jiahao Pang and Gene Cheung. Graph Laplacian regularization for image denoising: Analysis in the continuous domain. *IEEE Transactions on Image Processing*, 26(4):1770–1785, 2017. doi: 10.1109/TIP.2017.2651400. 4

Tengya Peng, Ruyi Zha, Zhen Li, Xiaofeng Liu, and Qing Zou. Three-dimensional MRI reconstruction with 3D Gaussian representations: Tackling the undersampling problem. *IEEE Transactions on Medical Imaging*, 45 (5):1905–1917, 2026. doi: 10.1109/TMI.2025.3642134. 3

Sathish Ramani, Zhihao Liu, Jeffrey Rosen, Jon-Fredrik Nielsen, and Jeffrey A. Fessler. Regularization parameter selection for nonlinear iterative image restoration and MRI reconstruction using GCV and SURE-based methods. *IEEE Transactions on Image Processing*, 21(8):3659–3672, 2012. doi: 10.1109/TIP.2012.2195015. 3

Chinmay Rao, Efe Ilıcak, Matthias J. P. van Osch, Mariya Doneva, Laurens Beljaards, Navid Jabarimani, Nicola Pezzotti, and Marius Staring. Data-efficient networks for multi-contrast MRI reconstruction based on a generalized content/style prior. *arXiv preprint arXiv:2609.01959*, 2026a. doi: 10.48550/arXiv.2609.01959. 3, 6, 17, 30

Chinmay Rao, Matthias van Osch, Nicola Pezzotti, Jeroen de Bresser, Mark van Buchem, Laurens Beljaards, Jakob Meineke, Elwin de Weerdt, Huangling Lu, Mariya Doneva, and Marius Staring. A plug-and-play method for guided multi-contrast MRI reconstruction based on content/style modeling. *Medical Image Analysis*, 113:104160, 2026b. doi: 10.1016/j.media.2026.104160. 3, 6, 17

Liyue Shen, John Pauly, and Lei Xing. NeRP: implicit neural representation learning with prior embedding for sparsely sampled image reconstruction. *IEEE Transactions on Neural Networks and Learning Systems*, 35(1): 770–782, 2024. doi: 10.1109/TNNLS.2022.3177134. 3, 6

Anuroop Sriram, Jure Zbontar, Tullie Murrell, Aaron Defazio, C Lawrence Zitnick, Nafissa Yakubova, Florian Knoll, and Patricia Johnson. End-to-end variational networks for accelerated MRI reconstruction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 64–73, 2020. doi: 10.1007/978-3-030-59713-9_7. 3

Marshall F. Tappen, Ce Liu, Edward H. Adelson, and William T. Freeman. Learning Gaussian conditional random fields for low-level vision. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–8, 2007. doi: 10.1109/CVPR.2007.382979. 3

Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 23–30, 2017. doi: 10.1109/IROS.2017.8202133. 3

Guobao Wang and Jinyi Qi. PET image reconstruction using kernel method. *IEEE Transactions on Medical Imaging*, 34(1):61–71, 2015. doi: 10.1109/TMI.2014.2343916. 3

Rachel Ward. Compressed sensing with cross validation. *IEEE Transactions on Information Theory*, 55(12): 5773–5782, 2009. doi: 10.1109/TIT.2009.2032712. 3

Lei Xiang, Yong Chen, Weitang Chang, Yiqiang Zhan, Weili Lin, Qian Wang, and Dinggang Shen. Deep-learning-based multi-modal fusion for fast MR reconstruction. *IEEE Transactions on Biomedical Engineering*, 66(7):2105–2114, 2019. doi: 10.1109/TBME.2018.2883958. 1, 3

Burhaneddin Yaman, Seyed Amir Hossein Hosseini, Steen Moeller, Jutta Ellermann, Kâmil Uğurbil, and Mehmet Akçakaya. Self-supervised learning of physics-guided reconstruction neural networks without fully sampled reference data. *Magnetic Resonance in Medicine*, 84(6):3172–3191, 2020. doi: 10.1002/mrm.28378. 3

Burhaneddin Yaman, Seyed Amir Hossein Hosseini, and Mehmet Akçakaya. Zero-shot self-supervised learning for MRI reconstruction. In *International Conference on Learning Representations (ICLR)*, 2022. 3

Jin Zeng, Jiahao Pang, Wenxiu Sun, and Gene Cheung. Deep graph Laplacian regularization for robust denoising of real images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1759–1768, 2019. doi: 10.1109/CVPRW.2019.00226. 3

Ruyi Zha, Tao Jun Lin, Yuanhao Cai, Jiwen Cao, Yanhao Zhang, and Hongdong Li. R^2-Gaussian: rectifying radiative Gaussian splatting for tomographic reconstruction. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 44907–44934, 2024. doi: 10.52202/079017-1427. 3

Bo Zhou and S Kevin Zhou. DuDoRNet: learning a dual-domain recurrent network for fast MRI reconstruction with deep T1 prior. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4272–4281, 2020. doi: 10.1109/CVPR42600.2020.00433. 1, 3, 6

A Evaluation protocol

Simulated measurements. Except for M4Raw, we simulate $y = M(Fx^\star + \epsilon)$ from the normalised magnitude image $x^\star \in [0,1]^{256\times256}$ of the target contrast. Here F is the unitary 2D DFT with the zero frequency shifted to the array centre, M keeps the sampled phase-encoding columns, and ϵ is circularly symmetric complex Gaussian noise with $\mathbb{E}|\epsilon_j|^2 = \sigma^2$, that is, standard deviation $\sigma/\sqrt{2}$ in each of the real and imaginary parts. Because F is unitary, a fully sampled image has the same noise level σ. Relative to the mean target intensity over pixels above 0.1, $\sigma = 0.01$ corresponds to an SNR of 33–55 (Table 4). We use the same σ and sampling rule for training, validation and all methods, except in Section T, which tests the volumetric models at $\sigma = 0.03$ and 0.05. Each mask keeps $\mathrm{round}(256/R)$ columns (32 at R=8 and 64 at R=4): the central 10 (4%) and further columns drawn uniformly without replacement from the rest. Masks are not symmetric about the centre. Because $x^\star$ is real, its noise-free spectrum is Hermitian, $(Fx^\star)_{-k} = \overline{(Fx^\star)_k}$, so an acquired column also constrains its mirror when the reconstruction is restricted to real images. Our solver and MoDL impose this restriction on the simulated data. For the complex multi-coil M4Raw data, our solver reconstructs complex images (Section L).

Cohorts and preprocessing. Companion and target volumes are each divided by the 99.5th percentile of their nonzero voxels and clipped to $[0,1]$. Axial slices are then centre-cropped or zero-padded in-plane to 256×256, without resampling (Table 4). MSD Task01 provides co-registered, skull-stripped BraTS volumes on a 1 mm grid, and MSD Task05 provides prostate T2 and ADC on a common grid (Antonelli et al., 2022). Because the ADC maps are computed from diffusion-weighted images and not acquired directly, the prostate cohort tests a change of anatomy and contrast but does not model a realistic ADC acquisition. For IXI test, BraTS and prostate, we evaluate five slices per subject, drawn with a fixed seed from the central 35% (IXI) or 25% (BraTS, prostate) of each volume. Two shorter prostate volumes contribute only four slices each, so the prostate cohort has 38 slices in total. The IXI-validation stress tests use three slices per subject.

Table 4: Simulated cohorts (evaluated subjects). *Matrix*: native in-plane size, zero-padded or centre-cropped to 256^2. *Voxel*: in-plane / slice spacing. *Band*: central fraction of each volume from which evaluation slices are drawn. *SNR*: mean target intensity over pixels above 0.1 divided by σ=0.01, the noise level of a fully sampled image (IXI: test / validation; BraTS: T2w / FLAIR / T1ce). M4Raw uses raw k-space (Section L).

Cohort	Companion → target	matrix	voxel (mm)	slices	band	SNR
IXI	PD → T2	256^2	0.90–0.94 / 1.20–1.25	112–136	35%	33 / 37
BraTS	T1w → T2w, FLAIR, T1ce	240^2	1.0 / 1.0	155	25%	43 / 49 / 55
Prostate	T2 → ADC	256^2, 320^2	0.60–0.75 / 3.0–4.0	15–22	25%	40

Metrics. Reconstructions are converted to magnitude and clipped to $[0,1]$. We compute PSNR with peak 1 and SSIM (scikit-image, data range 1, 7×7 window) over the full 256×256 image, including background. Noise for case i is generated with seed i, so every method receives identical measurements. Metrics are averaged within subjects and then across subjects. For paired comparisons, we average slice-level differences within each subject and bootstrap over subjects with 10,000 resamples. In the figures, axial images are shown with anterior at the top, and through-plane reformats preserve the voxel aspect ratio.

B Proofs

Proposition 1. For fixed w, the normal equations $N(w)\hat{x} = A^H y$, with $N(w) = A^H A + L(w) \succ 0$, give $\hat{x}(w) = N(w)^{-1} A^H y$. This expression is linear in y and has no additive term. The normal equations are equivalent to $A^H(y - A\hat{x}) = L(w)\hat{x}$, and $L(w) = \sum_{(a,b)} w_{ab}(\delta_a - \delta_b)(\delta_a - \delta_b)^\top \succeq 0$ for $w \geq 0$. For a CG iterate x_k, the violation $A^H(y - Ax_k) - L(w)x_k = b - N(w)x_k$ is the CG residual, with $b = A^H y$. Even when w depends on target measurements, $y = 0$ leaves a zero right-hand side and hence a unique zero solution. CG initialised at zero has iterates in $\mathrm{span}\{b, Nb, N^2 b, \dots\}$, where $b = A^H y$, and returns zero when $b = 0$.

For a companion change with fixed y, subtract the two normal equations to obtain $N(w')\big(\hat{x}(w') - \hat{x}(w)\big) = \big(L(w) - L(w')\big)\hat{x}(w)$. At node a, the right-hand side is $\sum_b (w_{ab} - w'_{ab})(\hat{x}_a - \hat{x}_b)$. It is supported on endpoints of changed edges and depends on the reconstructed contrast across them. If $\hat{x}_a(w) = \hat{x}_b(w)$ on every changed edge, the right-hand side vanishes and, since $N(w')$ is nonsingular, $\hat{x}(w') = \hat{x}(w)$ for any size of conductivity change. In general, $\|\hat{x}(w') - \hat{x}(w)\|_2 \le \lambda_{\min}(N(w'))^{-1} \, \| \sum_{(a,b)\in E_\Delta} (w_{ab} - w'_{ab})(\hat{x}_a - \hat{x}_b)(\delta_a - \delta_b)\|_2$, where E_Δ is the set of changed edges and δ_a the indicator vector of pixel a.

Polarity of the hand-crafted response. Let d be supported on Λ, and suppose g is constant on Λ and its one-pixel neighbourhood. For every edge (a, b) touching Λ, $g_a = g_b$, so $|(g \pm d)_a - (g \pm d)_b| = |d_a - d_b|$ for both signs. All other edges are unchanged. Hence $w^{\mathrm{hand}}(g + d) = w^{\mathrm{hand}}(g - d)$, and the two reconstructions coincide: the response is even in lesion polarity and its odd component is zero. If g varies locally, the two lesion polarities can change $|g_a - g_b|$ differently, and the resulting changes in interpolation give a nonzero odd component. The measured odd responses in Section 4.3 include this effect and the effect of clipping intensities to $[0, 1]$.

Proposition 2. Write Ω for the acquired lines. In our masks the non-calibration part Ω_{nc} of Ω is a uniformly random n_{nc}-subset of a pool $\mathcal{P}$ that contains the eligible lines $\mathcal{E}$. The holdout $\mathcal{H}$ is a uniformly random h-subset of $\Omega_{\mathrm{nc}} \cap \mathcal{E}$, and $\mathcal{F} = \Omega \setminus \mathcal{H}$. Fix a realisation of $\mathcal{F}$. Then $\Omega_{\mathrm{nc}} = (\mathcal{F} \cap \mathcal{P}) \cup \mathcal{H}$, so $h = n_{\mathrm{nc}} - |\mathcal{F} \cap \mathcal{P}|$ is determined, and every admissible $\mathcal{H}$ is an h-subset of $\mathcal{E} \setminus \mathcal{F}$. Each such configuration has probability $\binom{|\mathcal{P}|}{n_{\mathrm{nc}}}^{-1} \binom{|\mathcal{F}\cap\mathcal{E}|+h}{h}^{-1}$, which does not depend on which subset $\mathcal{H}$ is. Conditional on $\mathcal{F}$, $\mathcal{H}$ is therefore uniform over the h-subsets of $\mathcal{E} \setminus \mathcal{F}$, and each line $j \in \mathcal{E} \setminus \mathcal{F}$ is held out with probability $h/|\mathcal{E} \setminus \mathcal{F}|$.

Now condition also on the fit measurements $y_{\mathcal{F}}$ and companion g, so that $\hat{x}_{\mathcal{F}}(c)$ is fixed. For $j \notin \mathcal{F}$, $y_j = A_j x^\star + \epsilon_j$, where the noise is independent of the fit measurements and companion, zero-mean and of variance σ^2 per sample. Therefore, $\mathbb{E}[\|A_j \hat{x}_{\mathcal{F}}(c) - y_j\|^2 \mid \mathcal{F}, y_{\mathcal{F}}, g] = \|A_j(\hat{x}_{\mathcal{F}}(c) - x^\star)\|^2 + n_s \sigma^2$. Summing over $j \in \mathcal{E} \setminus \mathcal{F}$ with inclusion probability $h/|\mathcal{E} \setminus \mathcal{F}|$ gives the identity. This set contains the held-out lines and all unacquired eligible lines; at R=8, the latter account for more than 98% of it. For variable-density sampling, inverse-probability weighting requires the conditional holdout probabilities $\pi_j = \Pr(j \in \mathcal{H} \mid \mathcal{F}) > 0$ on the eligible complement; weighting line j's residual by π_j^{-1} gives a Horvitz–Thompson estimate of the corresponding total risk plus noise.

A Monte Carlo check using our mask sampler, the hand-crafted metric and one IXI validation slice (300 independent mask and holdout draws, $\sigma = 0.01$) gives a mean held-out residual of 0.772. The right-hand side predicts 0.726, and the difference is within one standard error of the mean (0.047).

C CONTRAST-RANDOMISED TRAINING

Diagnostic. On IXI validation (28 subjects, 3 slices each, R=8), the learned metric reaches 34.68 dB with the original companion, 19.72 dB after foreground inversion and 29.38 dB after a random monotone intensity transform. The inverted result falls below zero-filling (20.59 dB). The corresponding values for the hand-crafted metric are 27.41, 27.41 and 25.50 dB. Because it depends only on absolute intensity differences, the hand-crafted metric is insensitive to contrast polarity within the foreground.

Variants and selection. We compare four variants: (i) companion-only contrast randomisation, (ii) companion and target contrast randomisation, (iii) variant (ii) with random 90° rotations and flips, and (iv) variant (ii) trained from scratch for 40 epochs. The piecewise-linear map T is non-monotone or inverted with probability one half and keeps the background dark. It is applied to the companion with probability 0.8 and, when applicable, to the target with probability 0.5. Fine-tuned variants start from the model trained without randomisation and train for 12 epochs at learning rate $5 \cdot 10^{-5}$. We keep the checkpoint with the highest mean validation PSNR over original and inverted companions.

We select the variant with the highest mean IXI-validation PSNR over original, inverted and randomly transformed companions, subject to losing at most 1 dB with the original companion. Only companion-only randomisation satisfies this constraint (Table 5). Randomising the target as well reduces in-distribution PSNR by more than 1 dB but improves transfer to BraTS T2w, FLAIR and T1ce (33.6,

32.7 and 35.0 dB) and to prostate (up to 25.5 dB). These variants are not included in the five-metric selector. A second training seed gives the same PSNR on the 28-slice checkpoint-selection subset (33.17 dB in both runs; 31.77 and 31.78 dB with inversion). The two seeds differ by at most 0.17 dB on IXI test, the zero-shot cohorts and the validation companion stress tests.

Alignment randomisation and affinity head. The contrast-and-alignment variant adds random rigid companion motion ($|t| \leq 2$ pixels per axis, $|\theta| \leq 3°$) with probability 0.5 and uses the same fine-tuning schedule. We also test an output head that derives conductivities from feature affinities (de Lutio et al., 2022). With the same 40-epoch schedule, this head reaches a PSNR/SSIM of 34.14 dB/0.949 on IXI validation, slightly below the 34.62 dB/0.952 of direct conductivity prediction.

Table 5: Contrast-randomisation variants of the learned metric (PSNR). Selection maximises mean IXI-validation PSNR over original, inverted and randomly transformed companions, subject to a loss of at most 1 dB with the original companion. Test columns use no target-domain retraining. The second seed was not used for variant selection.

	IXI validation (selection)				zero-shot test				
Randomised	exact	inverted	random map	criterion	T2w	FLAIR	T1ce	prostate	selected
none (final model)	34.68	19.72	29.38	27.93	27.15	30.10	32.87	23.37	
companion	33.76	32.30	32.29	32.78	32.32	30.55	32.14	24.75	✓
second training seed	33.76	32.30	32.28	32.78	32.24	30.53	32.31	24.67	
companion + target	32.67	31.22	31.38	31.75	33.57	32.72	34.98	24.97	
companion + target + rotations/flips	32.51	31.08	31.21	31.60	33.55	32.66	35.05	25.11	
companion + target, from scratch	32.06	31.66	31.01	31.58	33.53	32.44	34.86	25.54	

Randomised intensity priors. We fine-tune U-Net+DC and MoDL with the same augmentation code and number of epochs as the randomised metrics. For each network, we use companion-only or companion-and-target contrast randomisation, with or without alignment randomisation, which gives eight variants in total. All improve transfer relative to training without randomisation (Table 6), with PSNR gains of 4.3–8.2 dB on BraTS T2w and 2.4–3.5 dB on prostate. In exchange, their IXI test PSNR at R=8 drops by 1.0–4.6 dB. The contrast-randomised and contrast-and-alignment-randomised metrics lose 0.8 and 1.2 dB (Table 8).

Table 6: Intensity-prior candidates (PSNR). U-Net+DC and MoDL each contribute a model trained without randomisation and four fine-tuned variants. Fine-tuning uses the same augmentation code and number of epochs as the randomised metrics: companion-only or companion-and-target contrast randomisation, with or without rigid companion motion. Bold indicates the best single candidate per column. Held-out selection compares all ten candidates per slice using the same rule as geometry selection (Table 8).

	IXI test		BraTS (zero-shot)			
Intensity-prior candidate	R=8	R=4	T2w	FLAIR	T1ce	Prostate
U-Net+DC, frozen	**35.54**	36.36	23.32	26.55	28.22	19.62
contrast rand., companion	34.59	35.26	27.66	27.40	28.32	22.02
contrast rand., companion and target	33.75	33.94	28.92	28.81	30.42	22.25
contrast & alignment rand., companion	33.66	34.39	28.29	27.40	28.25	22.78
contrast & alignment rand., companion and target	32.68	33.18	29.30	29.01	30.64	23.16
MoDL, frozen	35.50	**37.11**	22.84	26.91	28.54	19.88
contrast rand., companion	32.89	33.73	29.35	29.00	29.97	22.69
contrast rand., companion and target	31.82	32.09	30.97	30.62	32.24	23.01
contrast & alignment rand., companion	32.01	32.74	30.15	29.61	30.66	23.16
contrast & alignment rand., companion and target	30.95	31.24	**31.07**	**30.81**	**32.29**	**23.36**
Held-out selection (10 models)	35.58	37.06	31.10	30.76	32.27	23.12

D ADDITIONAL BASELINE TRAINING DETAILS

TGVN, MC-VarNet, PnP-CoSMo and CoSMo-RecNet use the same IXI training split, masks, noise and R=8 acceleration as the other models. Checkpoints and hyperparameters are selected on IXI validation. We use the released implementations of TGVN, MC-VarNet and PnP-CoSMo and implement CoSMo-RecNet from its description. Table 7 reports results on the cohorts in Table 2.

TGVN. We keep the architecture and loss of Atalık et al. (2026) and supply the fully sampled companion as side-information k-space. Training uses Adam with learning rate $3 \cdot 10^{-4}$ and decay 0.98. For the single-coil IXI data, we use the nearly constant sensitivity map $1 + 10^{-3}\cos(2\pi x/97)\cos(2\pi y/131)$ to keep TGVN's preconditioned CG step well defined. TGVN reaches a PSNR/SSIM of 34.32/0.960 and 34.69/0.962 on IXI test at R=8 and R=4. Without retraining, it obtains 23.59/0.702 on BraTS T2w, 26.14/0.784 on FLAIR, 27.09/0.820 on T1ce and 19.67/0.238 on prostate. Its tumour-ROI PSNR on BraTS T2w is 18.65 dB, compared with 21.20 dB for zero-filling.

MC-VarNet. Training uses the L1 loss and optimiser of Lei et al. (2023): Adam with cosine learning-rate decay from 10^{-4} to 10^{-5}. The data-consistency layer inserts measured complex k-space into every feature channel and retains the real part.

PnP-CoSMo. We train the content/style model on IXI PD and T2 slices using the losses and optimisers of Rao et al. (2026b) for the unpaired and paired stages. Reconstruction uses 200 iterations. The content-refinement step size selected on IXI validation is 10^{-3} at both accelerations.

CoSMo-RecNet. We implement the four-block network of Rao et al. (2026a) using the unpaired content/style model above. It comprises wavelet-convolution domain adaptors, a content refiner, an MLP that predicts the content update weight and a CG data-consistency step. The network has 183k learnable parameters and is trained with an L1 loss.

MC-VarNet, PnP-CoSMo and CoSMo-RecNet reach 33.83, 32.46 and 30.08 dB on IXI test at R=8, respectively. All three have lower PSNR than zero-filling on BraTS T2w and prostate. Geometry selection outperforms all three for every subject in every cohort of Table 7. The smallest subject-paired mean gain, 1.21 dB (95% confidence interval $[0.89, 1.51]$), is over MC-VarNet on IXI at R=8.

Table 7: **Comparison with recent multi-contrast methods** (PSNR / SSIM). All models are trained on IXI PD→T2 at R=8 and evaluated without retraining on the cohorts in Table 2. Implementation and training details are in Section D. Bold indicates the best PSNR and SSIM in each column.

	IXI test		BraTS (zero-shot)			
Method	R=8	R=4	T2w	FLAIR	T1ce	Prostate ADC
Zero-filling	20.85 / 0.557	21.76 / 0.579	24.26 / 0.590	24.88 / 0.595	24.66 / 0.598	21.30 / 0.306
TGVN	34.32 / **0.960**	34.69 / 0.962	23.59 / 0.702	26.14 / 0.784	27.09 / 0.820	19.67 / 0.238
MC-VarNet	33.83 / 0.958	34.99 / 0.962	23.96 / 0.780	27.38 / 0.855	28.77 / 0.893	19.92 / 0.376
PnP-CoSMo	32.46 / 0.914	34.55 / 0.900	22.99 / 0.584	25.10 / 0.656	25.04 / 0.672	19.28 / 0.211
CoSMo-RecNet	30.08 / 0.927	30.76 / 0.930	21.75 / 0.407	24.37 / 0.421	24.88 / 0.429	17.59 / 0.216
Learned metric (ours)	**35.04** / 0.956	**37.66 / 0.968**	27.15 / 0.847	30.10 / 0.890	32.87 / 0.930	23.37 / 0.583
Held-out selection, 5 metrics (ours)	**35.04** / 0.956	**37.66 / 0.968**	**31.92 / 0.917**	**31.39 / 0.903**	**33.53 / 0.932**	**25.64 / 0.716**

E HELD-OUT K-SPACE SELECTION

Selection uses two repeats with fixed random seeds. Each repeat holds out 10% of the acquired lines outside a $\pm 2\%$ band around the k-space centre. We reconstruct each candidate from the remaining lines and recompute the learned conductivities from those data. The candidate with the smallest mean normalised held-out residual is then reconstructed using all acquired lines. The repeat count was chosen on IXI validation. For intensity selection, the networks also see only the fit data, through hard data consistency in U-Net and the learned data-consistency layer in MoDL. Table 8 compares

candidate sets of increasing size; Table 9 reports selection frequencies, agreement with the per-slice PSNR oracle and regret relative to that oracle. Adding the randomised metrics to the candidate set costs at most 0.01 dB on IXI test and 0.03 dB on IXI validation but improves results on BraTS T2w, prostate and perturbed companions.

Table 8: Geometry candidates and held-out selection with increasing candidate-set size (PSNR). The three-metric selector uses two hand-crafted metrics and the learned metric trained without randomisation. The four-metric selector adds contrast randomisation; the five-metric selector adds contrast and alignment randomisation and is used in Tables 1 and 2.

	test cohorts						IXI val., companion changed			
Candidate or selector	IXI R=8	IXI R=4	T2w	FLAIR	T1ce	Prostate	inverted	1 px	2 px	3°
Hand-crafted metric, κ=0.01	27.91	32.37	31.72	31.43	33.52	22.49	27.41	23.42	22.08	22.72
Hand-crafted metric, κ=0.02	28.17	31.60	31.29	31.29	33.37	23.05	27.23	23.82	22.40	23.08
Learned metric	35.04	37.66	27.15	30.10	32.87	23.37	19.72	25.38	22.53	24.17
+ contrast rand., companion	34.25	37.10	32.32	30.55	32.14	24.75	32.30	27.25	25.01	25.64
+ contrast & alignment rand., companion	33.82	36.70	32.27	30.17	32.06	25.64	31.68	31.22	30.08	30.02
Held-out selection (3 metrics)	35.04	37.66	31.55	31.40	33.54	23.38	27.23	25.24	22.54	24.09
+ contrast-randomised metric (4 metrics)	35.04	37.66	31.91	31.41	33.53	24.60	32.30	27.25	24.98	25.64
+ contrast-and-alignment-randomised metric (5 metrics)	35.04	37.66	31.92	31.39	33.53	25.64	32.25	31.22	30.08	30.02

Table 9: Held-out selection among five metrics. Columns report selection frequencies, agreement with the per-slice PSNR oracle, and mean and 95th-percentile PSNR regret relative to that oracle. The oracle is the candidate with the highest ground-truth PSNR on each slice. IXI-validation rows vary only the companion at test time (Table 1); percentages may not sum to 100 because of rounding.

	selected (% of slices)						regret (dB)	
Cohort	hand-crafted κ=0.01	hand-crafted κ=0.02	learned	contrast rand.	contrast & alignment rand.	oracle (%)	mean	p95
IXI test R=8	0	0	98	2	0	98	0.00	0.00
IXI test R=4	0	0	93	7	0	89	0.01	0.06
BraTS T2w	24	45	0	16	15	35	0.66	2.56
BraTS FLAIR	23	63	7	5	3	41	0.28	1.07
BraTS T1ce	20	61	16	1	1	37	0.29	1.14
Prostate ADC	0	0	0	0	100	100	0.00	0.00
IXI val, exact companion	0	0	98	1	1	98	0.03	0.00
IXI val, inverted companion	0	0	0	92	8	92	0.04	0.43
IXI val, random intensity map	0	0	6	86	8	88	0.08	0.49
IXI val, 1 px shift	0	0	0	0	100	100	0.00	0.00
IXI val, 2 px shift	0	0	0	0	100	100	0.00	0.00
IXI val, 3° rotation	0	0	0	0	100	100	0.00	0.00
IXI val, 4 px shift	0	0	0	0	100	100	0.00	0.00
IXI val, 6° rotation	0	0	0	0	100	100	0.00	0.00

F LESION PROBES

For each of the 28 IXI validation subjects, we sample a lesion centre within tissue (companion > 0.3, target > 0.2, at least 40 pixels from the image border). Gaussian lesions have amplitude 0.3 and width $\sigma \in \{1.5, 3, 6\}$ pixels. We add each lesion with both signs and clip intensities to $[0, 1]$. Target-only lesions alter the simulated k-space, whereas companion-only lesions leave the target measurements unchanged. From the reconstruction changes $\Delta_{\pm}$, we compute the recovered target fraction $\langle \Delta^{\text{target}}_{\text{odd}}, d\rangle / \|d\|^2$, the copied fraction $\langle \Delta^{\text{comp}}_{\text{odd}}, d\rangle / \|d\|^2$, and the energy of the even companion response relative to lesion energy (Table 10). We report the odd and even components separately to distinguish responses that depend on lesion polarity from responses shared by both polarities. The hand-crafted metric has the largest even response but the smallest copied fraction. Figure 5 shows individual responses.

Table 10: Polarity-resolved lesion probe (28 IXI validation subjects, R=8). *Gain*: recovered fraction of a target-only lesion. *Copied*: odd response to a companion-only lesion, projected onto the lesion and normalised by its squared norm. *Even*: energy of the even companion response relative to lesion energy. Randomised variants follow Tables 5 and 6.

	σ=1.5			σ=3			σ=6		
Method	gain↑	copied↓	even	gain↑	copied↓	even	gain↑	copied↓	even
U-Net+DC	0.30	0.62	0.17	0.50	0.40	0.09	0.76	0.16	0.05
+ contrast rand., companion	0.30	0.57	0.12	0.49	0.37	0.07	0.75	0.14	0.03
+ contrast rand., companion and target	0.30	0.54	0.11	0.49	0.35	0.07	0.74	0.14	0.03
DuDoRNet	0.35	0.54	0.12	0.52	0.36	0.05	0.75	0.16	0.04
MoDL	0.36	0.42	0.09	0.54	0.32	0.05	0.76	0.15	0.03
+ contrast rand., companion	0.31	0.41	0.12	0.51	0.29	0.09	0.75	0.12	0.03
+ contrast rand., companion and target	0.31	0.39	0.12	0.51	0.28	0.09	0.75	0.12	0.04
Hand-crafted metric	0.37	0.16	0.25	0.56	0.12	0.12	0.78	0.05	0.05
Learned metric	0.40	0.31	0.08	0.60	0.22	0.06	0.82	0.08	0.04
+ contrast rand., companion	0.40	0.26	0.07	0.59	0.20	0.05	0.81	0.08	0.02
+ contrast rand., companion and target	0.39	0.25	0.07	0.58	0.19	0.07	0.80	0.08	0.03

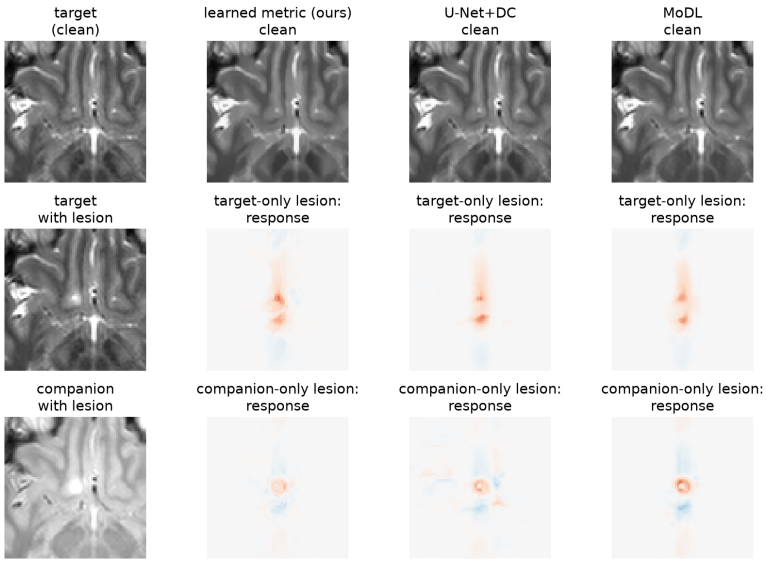

Figure 5: Lesion responses on one IXI validation slice (R=8, cropped views). A bright Gaussian lesion ($\sigma = 3$ pixels, amplitude 0.35) is added to the target only (row 2) or the companion only (row 3). Maps show the resulting reconstruction changes. Target measurements are unchanged in row 3. Table 10 quantifies odd and even responses using both lesion polarities.

G COMPANION STRESS TESTS

Appearance and alignment. At test time, we perturb the companion by foreground inversion, a random monotone piecewise-linear intensity map, diagonal translations of 1, 2 and 4 pixels, or rotations of 3 and 6 degrees (with bilinear resampling). Target k-space, masks and noise stay fixed. Table 1 reports the frozen and companion-randomised models and both selectors. The intensity selector also includes the four companion-and-target variants. Table 9 lists the selected geometry candidates.

Companion quality and anatomy. In Table 11, we degrade the companion or replace it with that of another subject (IXI validation, target acceleration $R{=}8$). U-Net+DC is 0.7 dB more accurate than the learned metric with the exact companion but 2.0–2.3 dB less accurate in all five perturbed settings. A companion that is itself undersampled at $R{=}8$ and zero-filled still helps both methods, which reach 23.84 and 21.89 dB, against 20.59 dB for target zero-filling. With another subject's companion, the learned metric drops to 20.92 dB, close to zero-filling, and U-Net+DC to 18.64 dB. The hand-crafted metric, which has no learned parameters, also falls below zero-filling (20.23 dB).

Table 11: Sensitivity to companion quality and anatomy (IXI validation, target $R{=}8$; PSNR). Companion degradation is applied only at test time. The $R{=}4$ and $R{=}8$ columns use undersampled, zero-filled companions; *other subject* supplies another subject's companion. Target measurements are held fixed.

Companion	exact	$R{=}4$ zero-filled	$R{=}8$ zero-filled	noise $\sigma{=}0.05$	noise $\sigma{=}0.1$	other subject
U-Net+DC	35.35	22.88	21.89	29.11	25.18	18.64
Hand-crafted metric	27.41	23.34	22.16	23.89	21.87	20.23
Learned metric (ours)	34.68	24.84	23.84	31.40	27.46	20.92
Zero-filled (no companion)			20.59			

H SUBJECT-PAIRED STATISTICS

All intervals are 95% subject-level bootstrap intervals (10,000 resamples), and counts give the number of subjects that favour the first method.

Classical baseline (dTV). On IXI test, the learned metric improves PSNR over dTV by 5.51 dB ($[5.26, 5.75]$, 9/9 subjects). On zero-shot BraTS T2w, the hand-crafted metric is 1.16 dB better than dTV ($[0.99, 1.33]$, 15/15), and on prostate the learned metric is 0.45 dB better ($[0.21, 0.73]$, 7/8). On raw M4Raw, the learned metric's gains are 3.20 dB at $R{=}4$ ($[3.06, 3.33]$) and 2.93 dB at $R{=}8$ ($[2.79, 3.05]$), with positive differences for all 25 subjects. With dTV as a fourth candidate, the selector reaches 31.49, 31.37 and 33.51 dB on T2w, FLAIR and T1ce, slightly below selection without dTV, and picks dTV on only 5 of 75 T2w slices.

Target contrasts and anatomy. Relative to DuDoRNet, the three-metric selector improves BraTS PSNR by 7.37 dB on T2w ($[6.29, 8.56]$), 3.89 dB on FLAIR ($[3.31, 4.47]$) and 4.33 dB on T1ce ($[3.23, 5.53]$), with positive differences for all 15 subjects on every target. On zero-shot prostate, the learned metric exceeds the strongest intensity prior trained without randomisation by 3.04 dB ($[2.74, 3.32]$, 8/8 subjects).

Randomised models and selectors. Against the strongest intensity result for each shifted cohort, five-metric selection gains 0.82 dB on T2w ($[0.45, 1.18]$), 0.59 dB on FLAIR ($[0.21, 0.97]$), 1.24 dB on T1ce ($[0.79, 1.72]$) and 2.27 dB on prostate ($[1.72, 2.84]$, 8/8 subjects). The comparator is intensity selection on T2w and, on the other cohorts, MoDL trained with companion and target contrast randomisation and companion alignment randomisation. Against intensity selection alone, the four gains are 0.82, 0.64, 1.26 and 2.51 dB, with intervals above zero. On IXI test at $R{=}8$, intensity selection is 0.54 dB better ($[0.01, 1.06]$). On IXI validation, geometry selection improves on intensity selection under every companion perturbation in Table 1. Gains range from 1.09 dB for inversion to 3.58 dB for a 2 px shift, including 2.75 dB at 4 px and 2.45 dB at 6°. All 28 subjects have positive differences.

Volumetric reconstruction. With independent random masks on the 388 central IXI test slices, the volumetric metric reaches 36.16 dB. Subject-paired gains are 1.27 dB over the 2D metric ($[1.15, 1.38]$, 9/9 subjects), 0.86 dB over U-Net+DC ($[0.36, 1.45]$, 8/9) and 0.94 dB over 2D MoDL ($[0.63, 1.30]$, 9/9). With a shared random mask, gains remain positive over the 2D metric (0.79 dB, $[0.67, 0.92]$, 9/9) and MoDL-3D (1.38 dB, $[1.01, 1.75]$, 9/9).

I In-distribution and BraTS details

Table 12 gives the per-method in-distribution results on five slices per IXI subject; the volumetric comparison on identical full-volume support is in Section K. Table 13 adds tumour-ROI PSNR on BraTS T2w. All frozen intensity priors are below zero-filling inside the tumour, whereas the geometry priors are above it.

Table 12: In-distribution results (PSNR / SSIM, subject-level mean) on IXI PD→T2. dTV parameters are tuned on IXI validation at R=8. All learned models are trained once at R=8. Variants of our model are compared in Table 24; the volumetric model in Tables 18 and 20.

Method	IXI val R=4	IXI val R=8	IXI test R=4	IXI test R=8
Zero-filled	21.51 / 0.580	20.59 / 0.556	21.76 / 0.579	20.85 / 0.557
TV (CS)	23.60 / 0.733	21.46 / 0.660	23.93 / 0.735	21.74 / 0.662
Directional TV (companion-guided)	31.17 / 0.905	28.86 / 0.874	31.86 / 0.915	29.54 / 0.885
NeRP-style INR (prior embedding)	29.21 / 0.846	26.59 / 0.813	29.58 / 0.852	26.88 / 0.818
U-Net (no data consistency)	34.71 / 0.964	34.59 / **0.962**	34.90 / 0.962	34.82 / 0.961
U-Net+DC	36.15 / 0.904	**35.26** / 0.929	36.36 / 0.904	**35.54** / 0.931
MoDL	36.65 / **0.967**	34.93 / 0.961	37.11 / **0.970**	35.50 / **0.964**
DuDoRNet	36.41 / 0.958	34.22 / 0.952	37.06 / 0.962	34.97 / 0.957
Hand-crafted metric	31.64 / 0.932	27.40 / 0.885	32.37 / 0.940	27.91 / 0.893
Learned metric (ours)	**37.36** / 0.966	34.62 / 0.952	**37.66** / 0.968	35.04 / 0.956

Table 13: BraTS/MSD T2w reconstruction (15 subjects, 75 slices, R=8). The upper blocks use IXI-trained models without retraining; the final block fine-tunes each model on 30 BraTS training subjects. Tumour-ROI PSNR is computed within the expert segmentation on the 64 slices with more than 50 tumour pixels.

Method	PSNR	SSIM	tumour-ROI PSNR
Zero-filled	24.26	0.590	21.20
TV (CS)	26.06	0.782	22.08
Directional TV (companion-guided)	30.56	0.862	24.18
NeRP-style INR (prior embedding)	25.40	0.755	–
U-Net+DC	23.32	0.566	18.10
MoDL	22.84	0.730	18.02
DuDoRNet	24.18	0.757	18.99
Hand-crafted metric	31.72	0.923	24.66
Learned metric (ours)	27.15	0.847	21.42
U-Net+DC, fine-tuned on 30 BraTS subjects	30.68	0.776	23.66
MoDL, fine-tuned on 30 BraTS subjects	32.66	0.947	25.44
Learned metric, fine-tuned on 30 BraTS subjects (ours)	33.85	0.950	26.37

J Additional transfer results

Table 14 evaluates the Cartesian-trained metric with a golden-angle radial NUFFT operator. Tables 15 and 16 report PSNR and SSIM on BraTS and prostate, including zero-shot models and models fine-tuned for each BraTS target or trained on prostate data. Table 16 also reports gland-ROI PSNR. These tables use the three-metric selector of Table 17, which has no randomised candidates. Results with

randomisation and the five-metric selector are in Tables 2, 6 and 8. Figures 6 to 9 show representative reconstructions.

Main-text qualitative examples. Figure 4 shows `BRATS_037`, slice 25, and `prostate_40`, slice 4 (zero-based). In each cohort, we choose the subject nearest the median subject-mean PSNR gain over the strongest cohort-level intensity strategy, then show its median evaluated slice; ties are broken by subject identifier and slice index. The augmented MoDL uses companion and target contrast randomisation and companion alignment randomisation. Intensity selection chooses contrast-randomised MoDL on the BraTS slice and U-Net+DC with companion contrast and alignment randomisation on the prostate slice. Geometry selection chooses the hand-crafted metric with κ=0.02 and the contrast-and-alignment-randomised metric, respectively. Both selectors use held-out measurements and then reconstruct from all acquired lines. Insets show 56×56-pixel regions centred on the dataset's tumour or prostate-gland annotation. Within each row, all panels share the same orientation, display window and inset location. The full images are not cropped or filtered; PSNR is computed before adding display annotations.

Table 14: Golden-angle radial sampling on IXI validation (PSNR / SSIM). The metric network is trained on Cartesian R=8 data and evaluated with a radial NUFFT operator without retraining.

Method	16 spokes	32 spokes	64 spokes
Zero-filled	20.23 / 0.265	21.35 / 0.316	25.31 / 0.453
U-Net (no data consistency)	28.75 / 0.917	29.77 / 0.924	32.87 / 0.950
Hand-crafted metric	26.06 / 0.850	29.38 / 0.907	33.99 / 0.959
Learned metric (ours)	32.17 / 0.923	35.31 / 0.952	38.69 / 0.973

Table 15: BraTS reconstruction across three target contrasts (15 subjects, R=8; PSNR / SSIM). All targets use the same T1w companion. Zero-shot models use their IXI PD→T2 checkpoints, and dTV keeps its IXI-validation parameters. The three-metric selector is defined in Table 17; randomised models and five-metric selection are reported in Tables 2, 6 and 8. The final block fine-tunes each method on 30 BraTS training subjects separately for each target contrast. The volumetric row uses the central 50% of each volume; all other rows use five slices per subject. Bold indicates the best PSNR and SSIM within the zero-shot 2D and fine-tuned groups separately.

Method	T2w	FLAIR	T1ce
Zero-filled	24.26 / 0.590	24.88 / 0.595	24.66 / 0.598
TV (CS)	26.06 / 0.782	26.83 / 0.790	26.96 / 0.795
Directional TV (companion-guided)	30.56 / 0.862	30.42 / 0.844	32.03 / 0.861
U-Net+DC	23.32 / 0.566	26.55 / 0.647	28.22 / 0.704
MoDL	22.84 / 0.730	26.91 / 0.835	28.54 / 0.884
DuDoRNet	24.18 / 0.757	27.51 / 0.835	29.21 / 0.875
Hand-crafted metric	**31.72 / 0.923**	**31.43 / 0.911**	33.52 / **0.936**
Learned metric (ours)	27.15 / 0.847	30.10 / 0.890	32.87 / 0.930
Held-out selection, 3 metrics (ours)	31.55 / 0.913	31.40 / 0.904	**33.54** / 0.932
Held-out selection, 3 metrics + dTV	31.49 / 0.911	31.37 / 0.902	33.51 / 0.930
Held-out selection, volumetric (ours)	33.56 / 0.941	33.33 / 0.931	34.93 / 0.946
U-Net+DC, fine-tuned *per target contrast*	30.36 / 0.769	29.88 / 0.744	31.77 / 0.790
MoDL, fine-tuned *per target contrast*	31.69 / 0.933	30.94 / 0.909	33.37 / 0.938
Learned metric, fine-tuned *per target contrast* (ours)	**34.26 / 0.951**	**33.13 / 0.931**	**36.06 / 0.959**

Table 16: MSD prostate T2→ADC reconstruction at R=8 (8 test subjects, 38 slices). Zero-shot models use IXI brain PD→T2 checkpoints; dTV uses its IXI-validation parameters. Results with randomisation and five-metric selection are in Tables 2, 6 and 8. The lower block trains each method on 20 prostate subjects. Gland-ROI PSNR uses the expert segmentation of the peripheral and transition zones. Bold indicates the best whole-image PSNR and SSIM within each block.

Method	PSNR / SSIM	gland-ROI PSNR
Zero-filled	21.30 / 0.306	21.42
TV (CS)	22.84 / 0.550	22.31
Directional TV (companion-guided)	22.92 / 0.533	22.57
U-Net+DC	19.62 / 0.199	18.81
MoDL	19.88 / 0.326	17.56
DuDoRNet	20.33 / 0.408	17.59
Hand-crafted metric	22.49 / 0.468	22.01
Learned metric (ours)	23.37 / **0.583**	22.09
Held-out selection, 3 metrics (ours)	**23.38** / 0.577	–
Held-out selection, 3 metrics + dTV	**23.38** / 0.577	–
U-Net+DC, trained on prostate	23.92 / 0.453	23.72
MoDL, trained on prostate	25.70 / 0.757	24.87
Learned metric, trained on prostate (ours)	**27.04 / 0.780**	25.69

Table 17: Held-out selection without randomised models (R=8). In 2D, selection compares two hand-crafted metrics and the learned metric; the additional dTV row includes dTV as a fourth candidate. In 3D, it compares the hand-crafted and learned volumetric metrics. Two holdouts with fixed random seeds each withhold 10% of acquired lines outside the calibration band. Candidates are reconstructed from the fit lines and ranked by mean normalised held-out residual. BraTS uses a T1w companion for all three targets, without target-domain training. The 3D columns use the central 50% of each volume; other columns use five slices per subject.

Candidate	IXI val	IXI test	BraTS T2w	BraTS FLAIR	BraTS T1ce	IXI test, 3D	BraTS, 3D
Hand-crafted metric, κ=0.01	27.40 / 0.885	27.91 / 0.893	31.72 / 0.923	31.43 / 0.911	33.52 / 0.936	30.90 / 0.930	33.83 / 0.942
Hand-crafted metric, κ=0.02	27.60 / 0.878	28.17 / 0.886	31.29 / 0.905	31.29 / 0.900	33.37 / 0.928	–	–
Learned metric	34.62 / 0.952	35.04 / 0.956	27.15 / 0.847	30.10 / 0.890	32.87 / 0.930	36.16 / 0.965	29.23 / 0.896
Directional TV (validation-tuned)	28.86 / 0.874	29.54 / 0.885	30.56 / 0.862	30.42 / 0.844	32.03 / 0.861	–	–
Held-out metric selection (ours)	34.62 / 0.952	35.04 / 0.956	31.55 / 0.913	31.40 / 0.904	33.54 / 0.932	36.16 / 0.965	33.56 / 0.941
Held-out selection, metrics + dTV	34.62 / 0.952	35.04 / 0.956	31.49 / 0.911	31.37 / 0.902	33.51 / 0.930	–	–
Metric selection chooses learned (%)	100	100	0	8	17	100	7
+ dTV selection picks dTV (%)	0	0	7	4	4	–	–

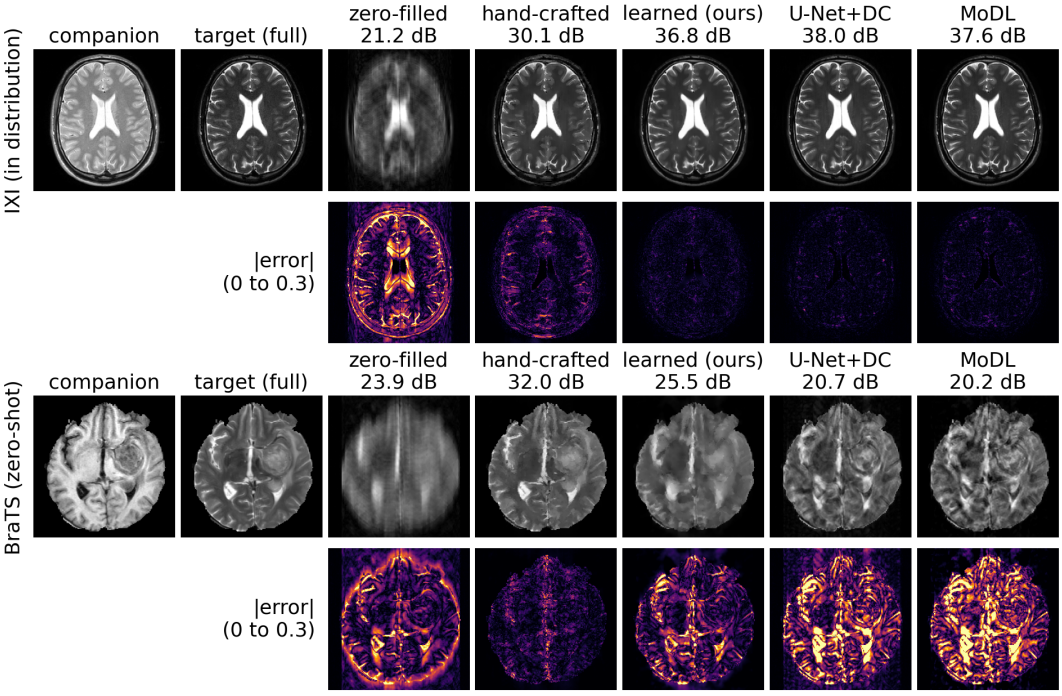

Figure 6: Reconstructions and error maps at R=8. Top: IXI validation. Bottom: BraTS T2w, reconstructed without retraining. The intensity priors have larger errors in the tumour region and lower PSNR than zero-filling on the BraTS example. The learned metric also loses accuracy under this shift. Held-out selection chooses the hand-crafted metric for this slice.

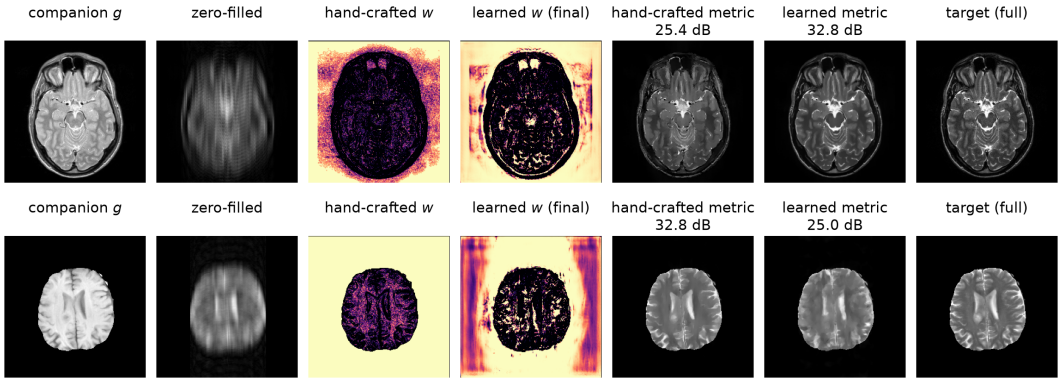

Figure 7: Anatomy-derived conductivities and reconstructions. Columns show the companion, zero-filled target, hand-crafted and learned conductivities, reconstructions and fully sampled target. Conductivity maps display the minimum over directions; darker values indicate less smoothing. On the IXI example (top), the learned metric recovers sharper detail. On the BraTS example (bottom), the hand-crafted metric is more accurate and is the one chosen by held-out selection.

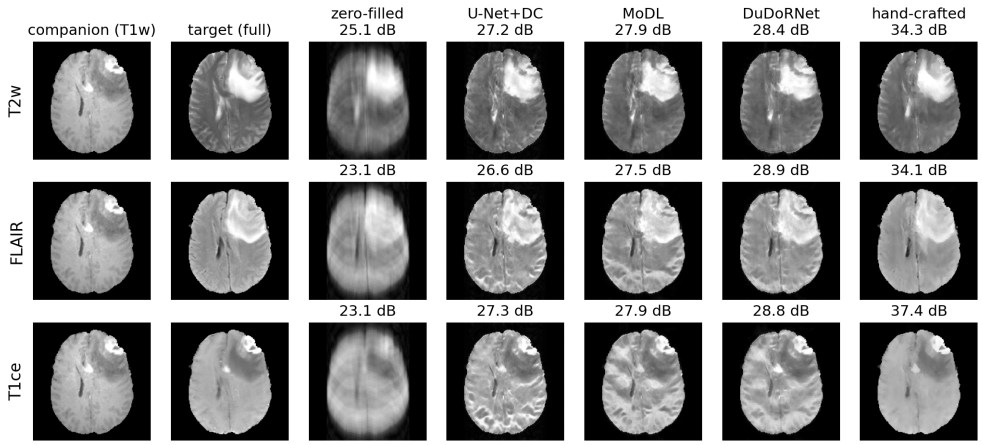

Figure 8: **Reconstructing three target contrasts from one companion.** Each row shows a different target contrast of the same subject at R=8, with a fixed T1w companion. Learned models use their IXI PD→T2 checkpoints. The intensity priors recover overall anatomy but introduce errors in tissue contrast. The geometry reconstruction uses the hand-crafted metric. Results for learned metrics and held-out selection are in Tables 2 and 8.

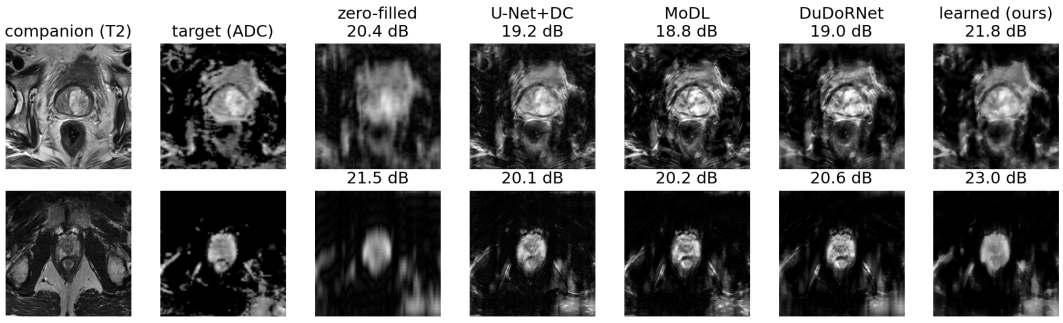

Figure 9: **Zero-shot reconstruction of prostate ADC.** A fully sampled T2 companion guides reconstruction at R=8. All learned models use their IXI brain checkpoints. The intensity priors introduce contrast errors and have lower PSNR than zero-filling. The learned metric preserves more of the target contrast.

K ADDITIONAL VOLUMETRIC RESULTS

Table 18 compares slice-wise and volumetric reconstruction on IXI and BraTS. Table 19 reports three independent runs of the full 2D and 3D training procedure, Table 20 varies acceleration, and Table 22 extends the lesion probe to 3D lesions. All methods in Table 21 are evaluated on the same 388 central IXI test slices.

Independent masks allow a volumetric method to combine complementary measurements across slices. Using one random mask for all slices removes this advantage but keeps the marginal mask distribution. With a shared mask, the five slice-wise methods change by at most 0.1 dB in subject-paired mean PSNR, while the volumetric metric decreases by 0.56 dB (95% confidence interval $[0.26, 0.90]$). It remains 0.79 dB above the 2D metric ($[0.67, 0.92]$, 9/9 subjects) and 1.38 dB above MoDL-3D ($[1.01, 1.75]$, 9/9). Its mean PSNR is also higher than that of U-Net+DC (by 0.28 dB, $[-0.38, 0.91]$, 6/9) and 2D MoDL (by 0.34 dB, $[-0.04, 0.69]$, 7/9), but these differences are not statistically significant. 2D MoDL has a 0.002 higher SSIM. A shared equispaced mask gives similar results. Both through-plane regularisation and complementary sampling therefore contribute to the gain under independent masks. Figure 10 shows a sagittal reformat.

Table 18: Volumetric reconstruction at R=8 with a different random mask per slice (central 50% of each volume; 28 IXI validation, 9 IXI test, and 15 BraTS subjects).

Method	IXI val	IXI test	BraTS zero-shot
Hand-crafted metric, 2D slice-wise	27.42 / 0.884	27.93 / 0.892	31.84 / 0.923
Hand-crafted metric, 3D volume	30.34 / 0.923	30.90 / 0.930	33.83 / 0.942
Learned metric, 2D slice-wise	34.60 / 0.951	34.90 / 0.956	27.24 / 0.845
Learned metric, 3D volume (ours; end-to-end from the 2D model)	35.85 / 0.961	36.16 / 0.965	29.23 / 0.896
MoDL-3D (trained in 3D)	34.14 / 0.955	34.57 / 0.960	23.95 / 0.762

Table 19: Seed replication of the final end-to-end 3D metric. Seeds 0/1/2 independently determine the 2D initialisation and the subsequent six-epoch 3D fine-tune. Subject-level mean PSNR / SSIM; the standard deviation is across seeds.

Split	Seed 0	Seed 1	Seed 2	Mean $\pm$ std.
IXI test R=8	36.16 / 0.965	36.10 / 0.964	36.14 / 0.964	36.13 $\pm$ 0.03 / 0.965 $\pm$ 0.000
IXI val R=8	35.85 / 0.961	35.77 / 0.960	35.82 / 0.960	35.81 $\pm$ 0.03 / 0.960 $\pm$ 0.000

Table 20: Volumetric reconstruction across accelerations (IXI test PSNR, 9 subjects, central 50% of every volume, one random mask per slice). Every row aggregates the same slices; all learned models were trained at R=8 only.

Method	R=4	R=8
Hand-crafted metric, 3D	34.42	30.90
Learned metric, 2D slice-wise	37.47	34.90
Learned metric, 3D (ours)	38.52	36.16
MoDL-3D (trained in 3D)	35.27	34.57

Table 21: Volumetric comparison on the same 388 slices from the central 50% of IXI test volumes (R=8; subject-level mean PSNR / SSIM). Independent random masks provide complementary measurements across slices. Shared random masks remove this complementarity but keep the marginal mask distribution. Shared equispaced masks also change the sampling pattern. Bold indicates the best PSNR and SSIM in each condition.

Method	independent random / slice	shared random	shared equispaced
Zero-filled	20.95 / 0.556	20.94 / 0.554	20.90 / 0.560
Learned metric, 2D (ours)	34.90 / 0.956	34.81 / 0.955	34.46 / 0.954
U-Net+DC	35.30 / 0.930	35.32 / 0.930	35.23 / 0.931
MoDL	35.23 / 0.964	35.26 / **0.964**	35.03 / **0.963**
DuDoRNet	34.50 / 0.955	34.51 / 0.955	34.27 / 0.953
MoDL-3D (trained in 3D)	34.57 / 0.960	34.22 / 0.959	34.06 / 0.958
Learned metric, volumetric (ours)	**36.16 / 0.965**	**35.60** / 0.962	**35.32** / 0.961

Table 22: Volumetric lesion probe on 16-slice sub-volumes from 28 IXI validation subjects (R=8). Gaussian lesions are added to the target only (gain: recovered fraction) or companion only (resp.: response energy relative to lesion energy). Unlike Table 10, this experiment does not separate lesion polarities, so the companion response includes all interpolation effects.

Method	σ=1.5		σ=3	
	gain↑	resp.↓	gain↑	resp.↓
Learned metric, 2D (ours)	0.37	0.40	0.58	0.28
Learned metric, 3D (ours)	0.42	0.27	0.63	0.19
MoDL-3D (trained in 3D)	0.28	0.52	0.47	0.37

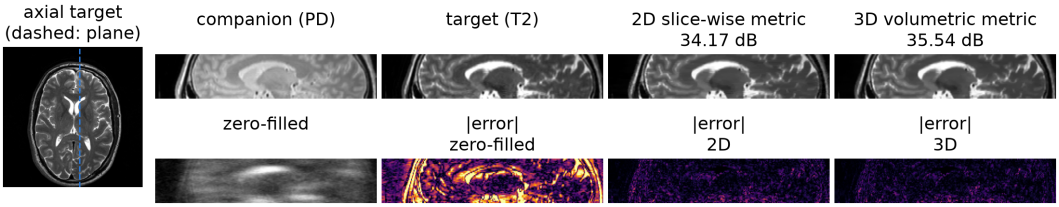

Figure 10: Sagittal reformat of the central 50% of an IXI test volume, with superior at the top and anterior to the left. Each axial slice uses a different random mask; the dashed line marks the reformat plane. The volumetric metric reduces the through-plane inconsistencies of slice-wise reconstruction and increases PSNR from 34.17 to 35.54 dB on this volume.

L RAW LOW-FIELD K-SPACE EXPERIMENTS

The M4Raw experiment uses all four raw target coils. For each mask, we estimate coil sensitivities from the 16 retained autocalibration signal (ACS) lines of the target. The fully sampled target acquisition provides the references for both training and evaluation. All learned models use the same M4Raw training split and number of epochs, with checkpoints selected by validation PSNR. U-Net, MoDL and the learned metric are initialised from their single-channel M4Raw checkpoints, while TGVN is trained with the architecture, loss and optimiser settings of Atalık et al. (2026). We evaluate three fixed slices from each of the 25 test subjects. A separate single-channel experiment combines the coils adaptively before undersampling (Table 23).

Table 23: Single-channel M4Raw reconstruction (0.3 T, T1→T2, 25 test subjects; PSNR / SSIM). Coils are adaptively combined into a virtual complex channel before retrospective undersampling. Learned models are fine-tuned on M4Raw training data, with checkpoints selected on separate validation subjects and identical masks used for test evaluation.

Method	M4Raw R=4	M4Raw R=8
Zero-filled	21.32 / 0.460	20.33 / 0.371
Hand-crafted metric	22.18 / 0.469	20.77 / 0.376
U-Net+DC	25.31 / 0.580	23.67 / 0.506
MoDL	25.73 / **0.604**	23.78 / **0.539**
Learned metric, fine-tuned (ours)	**25.77** / 0.581	**24.16** / 0.523

M ABLATIONS

Table 24: Ablations on IXI validation, R=8 (subject-level means). Width is the base channel count of the metric U-Net, and rounds is the number of refinement rounds. Indented rows modify the nearest non-indented row above them as stated. The Gaussian edge term replaces $\exp(-|g_a - g_b|/\kappa)$ in (2) by $\exp(-(|g_a - g_b|/\kappa)^2)$. Development runs were trained for 20 to 40 epochs; rows marked fixed λ use a unit global conductivity scale, the others learn it.

Variant	Width	Rounds	PSNR	SSIM
Euclidean Laplacian (no anatomy)	–	–	21.12	0.626
Structure-tensor metric (per-pixel)	–	–	23.26	0.741
Hand-crafted metric from zero-filled target only	–	–	21.72	0.678
Hand-crafted metric, Gaussian edge term, κ=0.03	–	–	27.36	0.869
Hand-crafted metric, κ=0.01 (default)	–	–	27.40	0.885
Learned, input: companion only	32	1	31.58	0.925
Learned, input: zero-filled target only	32	1	27.14	0.860
Learned, input: companion + zero-filled	32	1	31.77	0.926
+ decorrelation augmentation	32	1	31.76	0.928
+ intensity/geometric augmentation	32	1	31.56	0.926
learned modulation of the hand-crafted metric	32	1	28.08	0.893
Learned, input: + hand-crafted reconstruction	48	1	32.70	0.937
+ local and non-local edges	48	1	33.41	0.944
+ second refinement round	48	2	33.96	0.947
+ second round and synthesised target as input, fixed λ	48	2	34.09	0.949
64 channels, fixed λ	64	1	33.11	0.943
64 channels, fixed λ, second round (final)	64	2	34.62	0.952

Table 24 lists the ablations we ran during model development. A Euclidean Laplacian without companion information improves little on zero-filling. The per-pixel structure-tensor metric is about 4 dB below pairwise conductivities, and a hand-crafted metric derived only from the zero-filled target is about as accurate as TV. The Laplacian prior $\langle x, Lx\rangle$ also outperforms the heat-kernel and sparse-primitive alternatives (Section P). Among the network inputs, the companion provides most of

the useful geometry. Adding the hand-crafted reconstruction as an input and widening the network together improve PSNR by 0.9 dB, although this comparison does not separate the two changes.

Lesion-decorrelation augmentation and mild intensity augmentation (random gamma in $[0.6, 1.6]$, companion scaling and flips) do not improve in-distribution or zero-shot accuracy. Exchanging companion and target roles for half the training samples gives 31.5 dB in distribution and 27.0 dB on BraTS after five epochs, compared with the fully trained model's BraTS result of 27.2 dB. Mild gamma and scaling augmentations preserve contrast ordering. The non-monotone and inverted maps of Section C change appearance more strongly and do improve robustness to inversion.

N Negative results

Unless stated otherwise, these variants were evaluated on IXI validation at $R{=}8$.

Additional spatial context. Adding the two neighbouring companion slices as input channels slightly lowers PSNR, from 32.7 dB for the corresponding single-slice model to 32.5 dB (one round, 48 channels; Table 24). Adding 16 non-local candidate edges per pixel, chosen by nearest neighbours in companion-patch space, gives 33.4 dB on validation and 34.1 dB on test. This is no better than adding a second refinement round and requires ten times the computation.

Alternative metric parameterisations. Learning a multiplicative correction to the hand-crafted metric retains its robustness but gives lower PSNR than directly learning conductivities with the same inputs and width (28.1 vs. 31.8 dB). Replacing the hand-crafted Laplacian by the spectral polynomial $L + \beta L^2$ brings no improvement.

Additional refinement rounds. Fine-tuning a third round for 12 epochs gives a test PSNR similar to that of the two-round model (35.03 vs. 35.04 dB at $R{=}8$ and 37.60 vs. 37.66 dB at $R{=}4$). Applying three or four rounds at test time to the model trained with two lowers PSNR on the 20-slice validation subset of Section O from 35.45 dB to 33.09 and 34.10 dB.

Noise-dependent regularisation. Rescaling all volumetric conductivities by $(\sigma/0.01)^{\gamma}$ at test time lowers IXI test PSNR at $\sigma{=}0.03$ from 34.9 dB without rescaling (Table 26) to 34.5 dB for $\gamma{=}1$ and 31.8 dB for $\gamma{=}2$.

Coil sensitivity estimation. Smoothed full-image sensitivity maps on M4Raw give a relative four-coil forward error of $\|S\hat{m} - I_c\|/\|I_c\| \approx 0.35$ and reduce PSNR by 3 dB. All SENSE comparisons therefore use sensitivities estimated from the ACS lines.

Volumetric reconstruction on thick slices. On prostate, the zero-shot volumetric metric reaches 23.52 dB and slice-wise reconstruction 23.37 dB. Held-out selection chooses the learned volumetric metric on all eight volumes. The 0.15 dB gain is much smaller than on brain volumes (about 1.3 dB), consistent with the weak through-plane correlation of 3.6–4 mm slices. On M4Raw, an earlier single-channel model with hand-crafted through-plane conductivities from the T1 companion gives 25.1 dB at $R{=}4$ and 23.4 dB at $R{=}8$, 0.1 dB below slice-wise reconstruction in both cases. This experiment uses thick slices, a noisy 0.3 T companion and no learned through-plane network.

O Implementation details

Solver. We apply CG to the normal equations with a zero initial guess: 40 iterations for 2D training, 60 for 3D training and 2D evaluation, and 80 for volumetric evaluation. Because the conductivities are trained for an early-stopped solve, running CG to convergence lowers accuracy. On a 20-slice validation subset (10 subjects $\times$ 2 slices), the final 2D model gives 33.54, 34.99, 35.45, 35.06, 34.40 and 34.35 dB at 30, 40, 60, 80, 150 and 300 iterations, respectively, while the median relative residual $\|A^H y - N(w)\hat{x}\|/\|A^H y\|$ decreases from $1.9 \cdot 10^{-3}$ at 60 iterations to $4.8 \cdot 10^{-5}$ at 150 and $8.9 \cdot 10^{-7}$ at 300.

The 8-neighbour graph stores four conductivity maps (right, down, down-right and down-left), with diagonal weights divided by squared edge length. Shifted tensor operations apply the Laplacian without custom kernels.

Table 25: Model size and inference time (batch 1, 256×256, R=8, median of 20 CUDA-synchronised runs on one RTX A6000, 3D methods per 46-slice volume). Two independent runs agreed to within 3% for the 2D methods and 10% for MoDL-3D.

Method	Parameters	Inference time
U-Net+DC	14.79 M	8 ms / slice
MoDL	0.37 M	23 ms / slice
DuDoRNet	0.65 M	63 ms / slice
Hand-crafted metric	none	40 ms / slice
Learned metric, 2D (ours)	12.51 M	129 ms / slice
MoDL-3D	0.63 M	1.70 s / volume (46 slices)
Learned metric, 3D (ours)	12.51 M	1.93 s / volume (46 slices)

Implicit gradient. For $\ell(\hat{x})$, with $\hat{x} = N(w)^{-1}b$ and $N(w) = A^H A + L(w)$, we solve $N(w)u = \nabla_{\hat{x}}\ell$ by CG. The gradient is $\partial\ell/\partial w = -\nabla_w \operatorname{Re}\langle u, N(w)\hat{x}\rangle$, treating u and $\hat{x}$ as constants when differentiating this scalar. For a Laplacian penalty, this gives (3). The penalty acts on both real and imaginary components of complex images; conjugation has no effect for real images. The implementation also supports per-pixel ridge terms, spectral polynomials and other linear operators with adjoints. We use `torchkbnufft` for radial trajectories. On multi-coil SENSE data, a finite-difference directional derivative agrees with the analytic gradient to within 0.8%.

Network and training. The final 2D metric network is a four-level U-Net (GroupNorm, SiLU, base width 64, 12.51 M parameters) whose last layer is initialised to a constant conductivity. Its inputs are the companion, zero-filled target and hand-crafted reconstruction, followed by the previous reconstruction in the second refinement round. The global conductivity scale is fixed at 1. Training uses Adam for 40 epochs with learning rate 10^{-4}, cosine decay, batch size 4, gradient clipping at 1, and eight random slices from each of 540 subjects (4320 slices/epoch). We initialise the 3D model from this checkpoint and jointly fine-tune the in-plane network and a two-layer through-plane network for six epochs on depth-8 sub-volumes, with learning rates $2\cdot10^{-5}$ and $5\cdot10^{-4}$. We select 3D checkpoints by PSNR on a fixed 14-subject subset of IXI validation. Model sizes and inference times are in Table 25.

Baselines. U-Net uses base width 64; MoDL uses five unrolled iterations, a 64-channel residual CNN and a learned data-consistency weight. Both use the same IXI split, masks and noise as our model and reach 35.3 and 34.9 dB on validation. On M4Raw, they are fine-tuned on complex k-space with their original training settings. TGVN keeps its architecture and loss and uses the 128/30/25 M4Raw split and the target-ACS sensitivity maps shared by all SENSE methods. The published M4Raw results of CoSMo-RecNet (Rao et al., 2026a) use foreground-only metrics and different training-set sizes, so they are not directly comparable to Table 3.

P Prior families on the anatomy manifold

The anatomy graph also defines heat-kernel representations related to Gaussian primitives. In Euclidean space, the heat kernel is a Gaussian with variance $2t$, and $e^{-tL}\delta_p$ is its discrete counterpart on the graph, analogous to the geometry-adapted kernels used by Heat Kernel Textures (Foti et al., 2026). We compare three representations on four IXI validation slices at R=8: (i) the Laplacian-precision penalty $\langle x, Lx\rangle$; (ii) a dense multi-scale field $x = \sum_j e^{-t_j L}a_j$ with a ridge penalty on amplitudes, using Chebyshev polynomial approximations of the heat operators and up to four scales $t \in \{0.25, \dots, 16\}$; and (iii) sparse geodesic Gaussian primitives, with one source per SLIC cell, 1.6k–6.8k primitives, and optimised positions and scales.

The Laplacian penalty reaches 29.5 dB. The best heat-kernel field reaches 27.4 dB, with three scales and the Gaussian edge term defined in Table 24 (κ=0.03). With the default metric, the best results

using one, three and four scales are 22.9, 25.0 and 26.2 dB. Sparse primitives also have limited fitting capacity on a metric with strong boundaries. When fitted to the fully sampled target, they give at best 24 dB with 3.7k primitives and 17 dB at κ=0.01. Because low conductivities restrict diffusion between regions, images on such a metric are difficult to represent with a sparse set of sources. We therefore use the Laplacian penalty in all other experiments.

Q STRUCTURE-TENSOR VERSUS PAIRWISE METRIC

In our structure-tensor implementation, assigning conductivity per pixel creates a band of low conductivity 2–3 pixels wide around tissue boundaries. On the diagnostic slice, this metric gives 24.9 dB with Tikhonov reconstruction and at most 27 dB when fitting the fully sampled image with a dense heat-kernel field. Pairwise conductivities (2) place the barrier between pixels and give 29.8 dB on the same slice.

R DATA DEPENDENCE

In Fig. 11, we hold the companion fixed and reconstruct from zero measurements, another subject's k-space or the target's own k-space.

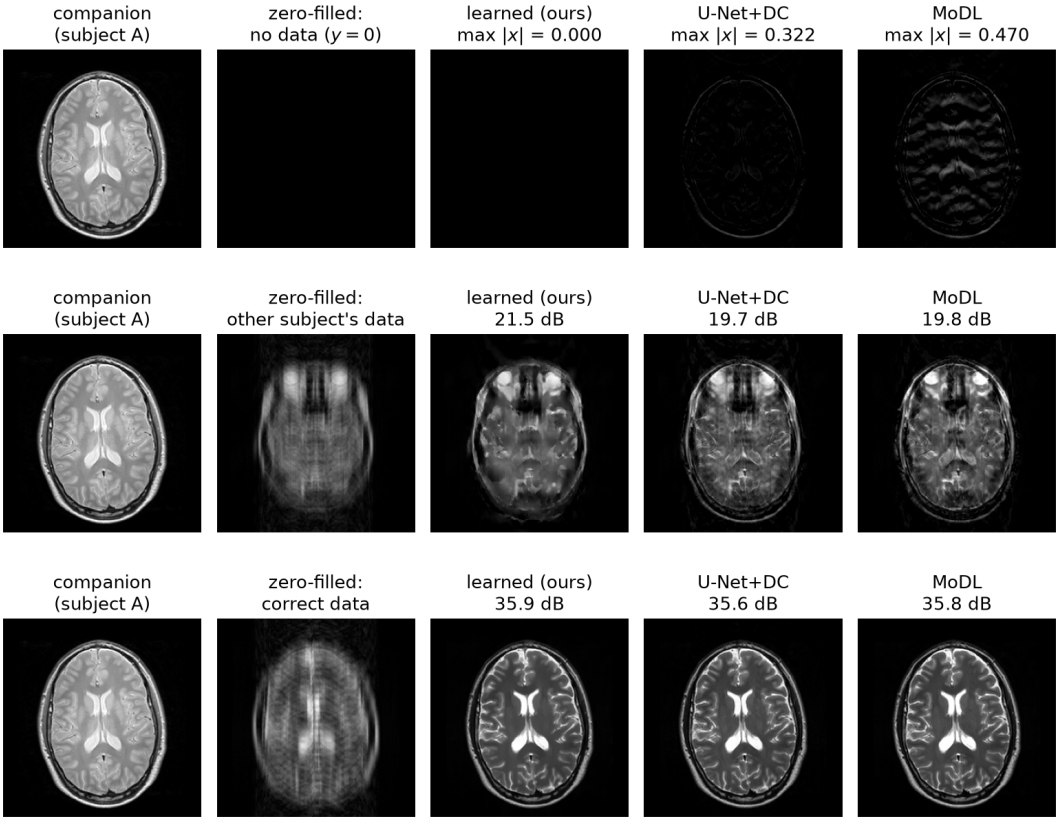

Figure 11: Data dependence (IXI validation, R=8). With y=0, the final learned metric returns exactly zero (Proposition 1), whereas U-Net+DC and MoDL return anatomy-shaped images. With another subject's k-space, all methods follow the data, and PSNR is computed against the subject whose k-space is used.

S Posterior uncertainty

For a fixed metric, a Gaussian likelihood with covariance qI and an intrinsic prior with precision $L(w)/q$ give posterior covariance $qN(w)^{-1}$. This conditional covariance omits uncertainty in the metric. We estimate its diagonal using 16 Hutchinson probes, each requiring an additional CG solve. Predicted standard deviations are largest near tissue boundaries and in weakly constrained regions, but underestimate errors on IXI validation. Absolute error falls within one standard deviation for 35% of brain pixels and within two for 59%. A scale factor of 2.27, fitted on half of the validation subjects, increases coverage to 62% and 82% on the other half, still below the nominal 68% and 95%. The pixelwise correlation between predicted standard deviation and absolute error is only 0.17.

T Robustness to noise

Both volumetric models are trained at σ=0.01 and evaluated on the same full-volume slices. At σ=0.03, the volumetric metric exceeds MoDL-3D by 0.64 dB in PSNR (95% confidence interval $[0.33, 0.97]$, 9/9 subjects) and 0.0068 in SSIM. At σ=0.05, MoDL-3D has 0.56 dB higher PSNR ($[0.19, 0.89]$, 7/9), while the metric has 0.0092 higher SSIM ($[0.0082, 0.0102]$, 9/9; Table 26).

Table 26: Full-volume robustness on IXI test at R=8 (PSNR / SSIM). Both methods use the same 388 central slices, random masks and noise realisations; both were trained only at σ=0.01.

Method	σ=0.01	σ=0.03	σ=0.05
Learned metric, 3D (ours)	**36.16 / 0.965**	**34.94 / 0.955**	33.26 / **0.937**
MoDL-3D (trained in 3D)	34.57 / 0.960	34.30 / 0.949	**33.82** / 0.928

U Reversed contrast roles

When we retrain with T2 as the companion and PD as the target, the learned metric reaches 34.57 dB on validation and 34.66 dB on test. Test PSNR is 27.18 dB for the hand-crafted metric, 20.94 dB for TV and 20.09 dB for zero-filling (Table 27). The volumetric model reaches 36.30 dB on both splits, 1.7 and 1.6 dB above the corresponding 2D results. These comparisons use different slice sets: five slices per subject for 2D and the central 50% of each volume for 3D. The analogous differences in the original contrast direction are 1.2 and 1.1 dB (Tables 12 and 18). Performance relative to the classical baselines is similar in both directions.

Table 27: Reversed contrast roles at R=8: T2 is the fully sampled companion and PD the undersampled target (the opposite of the main experiments). The metric network was retrained in this direction with the same settings. The 2D rows use five slices per subject, the volumetric row the central 50% of each volume.

Method	IXI val	IXI test
Zero-filled	20.05 / 0.555	20.09 / 0.550
TV (CS)	20.91 / 0.653	20.94 / 0.649
Hand-crafted metric	26.95 / 0.863	27.18 / 0.870
Learned metric (ours)	34.57 / 0.949	34.66 / 0.953
Learned metric, volumetric (ours)	36.30 / 0.963	36.30 / 0.966