

Rethinking Real-World MRI Denoising: Learning from Physical Noise

Sebastian Rassmann¹ , David K ugler¹ , Sascha Brunheim¹ ,
Philipp Ehses¹ , and Martin Reuter^{1,2,3} 

¹ German Center for Neurodegenerative Diseases (DZNE), Bonn, Germany
`{firstname,lastname}@dzne.de`

² Athinoula A. Martinos Center for Biomedical Imaging, Boston, USA

³ Department of Radiology, Harvard Medical School, Boston, USA

Abstract. Magnetic resonance imaging (MRI) inherently suffers from noise, which limits downstream medical analyses. In MRI, noise-free images are unobtainable; therefore, existing denoising approaches formulate surrogate training objectives, compromising between preserving detail and concealing noise, causing domain shifts or incomplete denoising. To enable denoiser training directly on unmodified, noisy images, we exploit repeated acquisitions. This naturally constitutes a *physical* Noise2Noise (pN2N) setting. For unrepeated data, we introduce a diffusion-based re-noiser that synthesizes noisy image pairs, extending pN2N to *Renoise2Noise* (ReN2N). Furthermore, we demonstrate that ReN2N improves generalization to unseen datasets. Additionally, we propose to combine pN2N or ReN2N with optional guidance from co-acquired contrast, yielding four versions of our novel denoising framework: YADO (*You Accurately Denoise real Observations*). Across 14 test conditions, YADO consistently outperforms 17 state-of-the-art baselines, matching the quality of physically noise-suppressed images obtained via brute-force averaging of independent acquisitions. YADO thus establishes practical denoising for *real-world* acquisition settings.

Keywords: Denoising · MRI · Diffusion Model · Data augmentation · Medical image analysis

1 Introduction

Magnetic resonance imaging (MRI) is inherently affected by acquisition noise [1, 21], which degrades image quality and obscures tissue boundaries. This reduces the reliability of downstream analyses [33, 36, 65], especially for subtle, participant-specific structures such as white-matter hyperintensities (WMH) or perivascular spaces (PVS) [3, 60, 63, 73, 87]. Thus, effective denoising, *i.e.* the reduction of noise while maintaining medical information, has substantial potential to improve quantitative biomarkers and clinical decision-making [27, 63]. Here, the fundamental challenge lies in correctly separating the true signal from complex, spatially correlated [1] noise patterns, as found in *real-world* images.

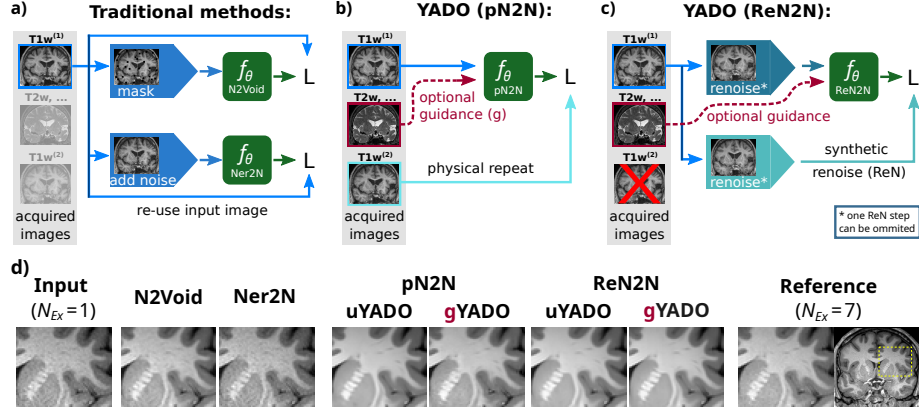


Fig. 1: Rethinking Brain MRI Denoising:

a) While traditional denoiser training relies on degradation (Noise2Void/N2Void: blind-spot masking [4, 37] or Noisier2Noise/Ner2N: noising [53]), **b)** we propose to use physical repeats for *physical* N2N (pN2N) training. **c)** We further generalize pN2N to single-noisy data by *re-noising* images (ReN2N) using a diffusion model. **b,c)** When available, co-acquired contrasts can guide (g) YADO (pN2N- or ReN2N-trained, 4 variants total) which increases total information over unguided (u) denoising. **d)** All YADO versions achieve unprecedented image quality from routine scans, matching the details of physically noise-suppressed reference images (here average of $N_{Ex}=7$ acquisitions).

While classical denoising methods rely on local statistics and hand-crafted priors [48, 49], the adoption of learning-based methods that implicitly capture rich content-specific image statistics led to superior performance [27, 37, 42, 44, 67, 75]. Recent approaches leverage surrogate self-supervised objectives, including blind-spot masking [4, 37, 85] or synthetic degradation [53, 59, 62, 71] (Fig. 1a) to enable denoiser training from single noisy images. However, their restrictive noise assumptions, such as voxel-independence, are violated in MRI [1, 79, 84]. Moreover, masking and degradation both reduce image context or quality, induce domain shifts (artificially corrupted vs. real-world images) and thus incentivize incomplete or biased noise removal at inference time. In contrast, more reliable direct supervision with clean, noise-free reference images in ideal Noise2Clean (N2Clean) training is impossible, since *real-world* MRI is inherently noisy. Zero-shot denoisers (ZSDs) [47, 75, 86] partially circumvent paired-data requirements, but remain computationally expensive and prone to overfitting individual noise realizations [75], which limits their applicability.

Here, we propose previously neglected N2N-like training [44] with *physically repeated*, co-registered image pairs (pN2N, Fig. 1b). We show that pN2N training on repeated acquisitions of the same participant – as available in many large-scale neuroimaging studies¹ – already outperforms state-of-the-art (SOTA) methods.

¹ The FOMO-60k data bundle [10] contains >9000 (42%) T1-weighted images (T1w) with in-session repeats and 65% are accompanied by co-acquired T2w and/or FLAIR.

To extend such training to datasets without physical repeats, we leverage diffusion models (DMs) [24, 68, 70] to simulate re-acquisition generating synthetically noise-decorrelated image pairs for training (Fig. 1c). We term this training paradigm *Renoise2Noise* (ReN2N). In contrast to previous Noisier2Noise (Ner2N) strategies, ReN2N avoids systematic shifts of the input data distribution toward noisier inputs by generating images with similar noise patterns as real images, while preserving information [63]. ReN2N training also serves as an effective data augmentation strategy, even when physical repeats are available. Finally, we observe that most real-world MRI protocols are multi-contrast¹. Although cross-modal information has previously been exploited indirectly to define denoising objectives [15, 78, 83], we show that explicit denoiser conditioning (guiding) on co-acquired contrasts provides complementary information that further improves denoising.

We integrate these insights into a unified structural MRI denoising framework: **YADO** (*You Accurately Denoise real Observations*), designed to reliably denoise *real-world* acquisitions. Taking advantage of the full potential of existing data, we achieve unprecedented image quality from routinely acquired MRI (Fig. 1d).

Contributions

Our contributions are physical Noise2Noise (pN2N), Renoise2Noise (ReN2N), and guided denoising (g). We show superior performance in extensive benchmarking against 17 SOTA approaches across 14 test conditions:

1. **pN2N**: We are the first to leverage independently-acquired image pairs from repeated MRI to formulate a near ideal denoiser training.
2. **ReN2N**: We generalize pN2N training to non-repeated, *i.e.* single-noisy, datasets by re-noising acquired images via a diffusion-based framework.
3. **guided denoising**: We introduce denoiser-conditioning on co-acquired different contrast MRI, thus guiding the denoising with additional available information.
4. We combine these novel approaches into **YADO** and show in extensive experiments, using independently acquired reference scans and challenging downstream evaluations, that it surpasses state-of-the-art performance for MRI denoising.

Overall, our results suggest that structural brain MRI analyses *generally* benefit from pre-processing with YADO.

We publish code and model weights at github.com/Deep-MI/YADO.

2 Theory

The aim in any denoising task is to obtain a denoising function f_θ that approximates the surrogate clean signal X' from a noisy measurement X , *i.e.* to find

$$\arg \min_{\theta} \text{metric}(f_\theta(X), X') \quad \text{with metric } e.g. \text{ MAE, MSE, (1-SSIM)}. \quad (1)$$

However, since MRI is inherently noisy, X' is unknown in practice. Hence, we define the clean image as $X' := \mathbb{E}_t[X_t]$. Assuming an additive zero-mean noise model and incorporating systematic biases [1, 21] into X' , the acquired image X_t decomposes into its deterministic X' and effective observable noise $\mathbf{n}_t$:

$$X_t = \mathbb{E}_s[X_s] + \mathbf{n}_t =: X' + \mathbf{n}_t, \quad \mathbf{n} \sim \mathcal{D}_{\text{Noise}}, \mathbb{E}[\mathbf{n}] := 0, \quad (2)$$

such that $\mathbf{n}_t$ contains all non-systematic effects, *e.g.* thermal measurement noise [1, 21], motion artifacts [84], or registration errors. X' can only be approximated by multiple consecutive MRI acquisitions (*i.e.* individual images $X_1, X_2, \dots$) as

$$\begin{aligned} X' = \mathbb{E}_t[X_t] &= X' + \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{t=1}^N \mathbf{n}_t \\ &\stackrel{N_{\text{Ex}} \gg 1}{\approx} X' + \frac{1}{N_{\text{Ex}}} \sum_{t=1}^{N_{\text{Ex}}} \mathbf{n}_t \\ &= X' + \bar{\mathbf{n}}_{N_{\text{Ex}}} . \end{aligned} \quad (3)$$

To empirically assess the performance of f_θ , the denoised image $\hat{X}' = f_\theta(X' + \mathbf{n}_{\text{in}})$ can be compared to *independently* acquired reference images $X_{\text{ref}} := X' + \mathbf{n}_{\text{ref}}$ with statistically independent noise $\mathbf{n}_{\text{ref}}$. For linear metrics such as the MSE,

$$\text{metric}(f_\theta(X' + \mathbf{n}_{\text{in}}), X' + \mathbf{n}_{\text{ref}}) = \mathbb{E}_{\mathbf{n} \sim \mathcal{D}_{\text{Noise}}} [\text{metric}(f_\theta(X' + \mathbf{n}_{\text{in}}), X') + c_{\mathbf{n}_{\text{ref}}}] , \quad (4)$$

i.e. the inherent acquisition noise in the reference image $\mathbf{n}_{\text{ref}}$ adds a constant aleatoric error $c_{\mathbf{n}_{\text{ref}}} := \text{metric}(X', X_{\text{ref}})$ [63]. This causes a systematic underestimation of denoising performance. Note that if $\mathbf{n}_{\text{in}}$ and $\mathbf{n}_{\text{ref}}$ are statistically dependent, performance rewards partially sustaining $\mathbf{n}_{\text{ref}}$ in the denoiser output, which violates the objective of estimating the clean signal X' (see Suppl. Sec. 8).

3 Related Work

3.1 General-purpose denoisers

Learning-based denoising without access to clean targets X' faces the problem of identity decay and therefore requires a suitable training objective (Eq. (1)). Noise2Noise (N2N) [44] showed that supervised training is possible from noisy images alone. However, this requires two observations $X_1 \neq X_2$ of the same clean image X' . To enable single-image training, subsequent methods construct surrogate objectives that obscure the noise. Noisier2Noise (Ner2N) [53] injects additional synthetic noise, with the expectation that training to revert incremental corruptions encourages the removal of real noise during inference. Noise2Score (N2Score) [31] recasts Ner2N as a score function estimate, reconstructing the clean signal via Tweedie’s formula, but requires knowing the noise magnitude.

Blind-spot approaches such as Noise2Void (N2Void) [37] and Noise2Self [4] predict masked pixels from their spatial context. Neighbor2Neighbor (Neigh2-Neigh) [26] improves efficiency through structured sub-sampling. While these

methods rely on voxel-wise noise independence, Structured N2Void [8] (S-N2Void) extends the blind-spot paradigm by enlarging the mask to encounter spatially correlated noise as found in MRI (Suppl. Sec. 11). However, dilating the mask reduces the available contextual information, limiting the range of concealable correlations.

Alternative approaches to self-supervised learning are zero-shot denoisers (ZSDs). Instead of applying a fixed pre-trained model, ZSDs optimize the network parameters directly on the noisy image at inference time. Consequently, they are inherently agnostic to domain shifts and specific noise statistics [4, 47, 50, 75]. A prominent example is Deep Image Prior (DIP) [75], which maps a fixed random input to the noisy observation using a randomly initialized neural network that acts as an implicit image prior. More recently, Pixel2Pixel [47] exploits non-local self-similarity to generate realistic pseudo-instances to handle spatially correlated noise similar to blind-spot methods. Although ZSDs are attractive in settings with extreme data scarcity, they typically incur substantial inference costs and remain susceptible to overfitting the specific noise realization [75].

3.2 Training objectives for brain MRI denoising

Classical denoising methods, such as non-local means (NLM) implemented in ANTs [49] and BM4D [48], exploit self-similarity and non-local structural correlations. However, they rely on hand-crafted priors and simple noise models. Despite these limitations, such approaches remain widely used in practical MRI analysis due to their efficiency, flexibility, robustness, and ease of deployment.

As acquisition noise in MRI is unavoidable ($\bar{\mathbf{n}} \neq 0$ for finite N_{Ex}), learning-based denoisers inherently face the challenge of defining suitable training objectives. Although MRI simulation can provide clean simulated targets for N2Clean training [57, 81] (Eq. (1)), simulated vs. physical images constitute a domain shift. Moreover, the required atlas construction is labor-intensive [13], limiting availability, and constraining learnable anatomical structures or pathologies [5].

Artificially degrading input images following the Ner2N paradigm [53] is common. Most approaches only add white Gaussian or Rician noise [12, 45, 59, 62, 81], whereas BME-X [71] applies more realistic degradations, including simulated motion. As an alternative, Neighboring Slice N2N [85] proposed a blind-spot approach that leverages 3D redundancy in medical images. Previous work has also introduced single-noisy training objectives based on Stein’s unbiased risk estimator (SURE) [57, 58]. However, their applicability to real-world scenarios with spatially correlated non-Gaussian noise remains unclear. Finally, DESN [86] introduced a ZSD based on Deep Image Prior (DIP) [75].

Whereas many works employ regression losses [15, 45, 78, 81], BME-X [71] additionally incorporates a tissue-segmentation loss that encourages intra-tissue smoothness. Other approaches utilize conditional GANs [11, 19, 29, 62] or diffusion models (DMs) [12, 24, 59, 68, 69, 83]. Although GANs and DMs excel at visual realism [43, 63, 66], information fidelity, i.e. precision and information preservation are much more important in medical imaging [6, 63]. In fact, we show that ideal N2N scenarios inherently result in the reproduction of noise from the training

images [59,63], yielding *re*-noisers rather than *de*-noisers (Fig. 10). An exception addressing this property is R2D2+ [12], which proposed to train a score-based DM on artificially degraded images. For inference, the DM is repurposed to reduce the overall noise level below the synthetic noise level seen during training.

3.3 Multi-modality for denoising

Several works exploit multi-modal information to obtain denoisers. N2Contrast [78] employs modality translation to optimize bilateral filters [77]. Patch2Self [15] leverages redundancy across b -vectors to denoise diffusion MRI (dMRI) in a simple ZSD. However, this approach does not generalize to applications where, unlike dMRI, only a few input volumes are available [15,83]. DDM² [83] exploits image translation to obtain clean surrogate targets for DM training. YODA [63] implicitly established a denoiser from image translation as cross-contrast N2N.

Overall, existing multi-modal approaches exclusively used cross-modal noise independence to overcome identity decay. Conversely, here we propose using complementary cross-modal information as an additional condition to guide the denoiser f_θ , enriching general information during training *and* inference.

3.4 Evaluation of brain MRI denoisers

Since noise-free images X' are impossible to acquire (Eq. (3)), evaluating denoiser performance in the *real world* is inherently challenging. It is common practice to resort to artificially degraded acquired images *e.g.* by adding synthetic noise, comparing against the original acquired noisy image (*e.g.* [45,57,62,81,85]) or a smoothed version thereof [71]. However, since re-using the input as the reference image rewards the (partial) reconstruction of the input noise (see Suppl. Sec. 8) and real-world MRI noise is more complex than typically simulated [57,62,71,78,86], white Gaussian/Rician noise (Suppl. Sec. 11), this approach does not constitute evidence for real-world denoising abilities.

Conversely, image pairs from repeated acquisitions inherently possess independent noise manifestations (Suppl. Sec. 11) and thus reflect the properties of real-world MRI noise. Although evaluating with noisy reference images introduces an unavoidable aleatoric error [63], this term is constant across methods (Eq. (3), Suppl. Eq. (12)) and therefore does not affect performance ranking.

Another common evaluation paradigm is application-driven experiments such as contrast-to-noise ratios (CNRs) around known structures of interest or downstream task performance [12,57,63,71,83]. Although such metrics can constitute a proxy for real-world usability [63], they are task- and structure-specific, labels can themselves be affected by noise [63], and the evaluations can be confounded by model design [71] or strong anatomical priors [57,71,83]. In particular, contrast- or homogeneity-based evaluations of conserved anatomies can be dominated by memorized population-level priors rather than image-derived evidence, rendering such evaluations insensitive to denoisers that suppress or hallucinate information *e.g.* about degraded, altered, or missing structures.

4 Method

4.1 Learning with physical noise

Given two repeated acquisitions $X_1 = X' + \mathbf{n}_1$ and $X_2 = X' + \mathbf{n}_2$ with approximately independent noise realizations $\mathbf{n}_1 \perp \mathbf{n}_2$ (Suppl. Sec. 11), we train the denoiser f_θ via

$$\arg \min_{\theta} \mathbb{E}_{\substack{X' \sim P_{X'} \\ \mathbf{n}_{1,2} \sim \mathcal{D}_{\text{Noise}}}} \left[\|f_\theta(X_1) - X_2\|_2^2 \right] \xrightarrow{\text{N2N [44]}} f_\theta(X) = \mathbb{E}_{X' \sim P_{X'}} [X' | X]. \quad (5)$$

Following N2N [44], this objective yields the same optimal estimator as hypothetical N2Clean training. As MRI noise is zero-mean by construction, $\mathbb{E}[X' | X]$ constitutes an unbiased estimate of X' , with residual uncertainty due to noise in X . Note that pN2N training preserves the full information in X , whereas blind spots or adding noise discard input information and introduce domain-shifts.

4.2 Cross-modality conditioning

When co-acquired contrasts C are available, we also condition the denoiser so that the ideal training yields $f_\theta(X, C) = \mathbb{E}[X' | X, C]$. Guiding with C therefore tightens the posterior and improves denoising performance, as different structures exhibit complementary visibility across MRI contrast.

4.3 *Renoise2Noise* (ReN2N) training

To generalize the proposed pN2N denoiser training to single-repeat domains (*e.g.* [2]), we propose *Renoise2Noise* (ReN2N) training. For this, we train conditional DMs [24, 68, 70] on repeated acquisitions, which we then leverage to generate synthetic repeats with decorrelated noise [63] for subsequent YADO training. This generalizes pN2N to single-noisy datasets. Note that DM sampling implicitly first generates a noise-free image estimate $\hat{X}'$, which is subsequently refined by a stochastic imitation of realistic acquisition noise [30, 59, 63]. Following [63], DM sampling can therefore be interpreted analogously to physical MRI as

$$\hat{X}_t^{(\text{DM})} = \text{DM}_\phi(X' + \mathbf{n}_{\text{phys}}) = \hat{X}'^{(\text{DM})} + \hat{\mathbf{n}}_t^{(\text{DM})}. \quad (6)$$

In contrast to common degradation-based training schemes [53], ReN2N preserves input information in expectation, as $\mathbb{E}_t[\hat{X}_t] \approx \hat{X}'$ [63]. When trained on data from similar contrasts and acquisition protocols, ReN2N minimizes domain shifts (Suppl. Sec. 12.4). Beyond generalizing pN2N training to unpaired datasets, we further propose ReN2N as an effective data augmentation strategy boosting performance on unseen datasets.

5 Experiments

5.1 Datasets and evaluation

We considered three large neuroimaging studies for training: OASIS-3 [39] (1 mm T1w), HCP Aging (HCP-A) [7, 18] (0.8 mm T1w), and the Rhineland Study (RS) [34] (0.8 mm T1w/T2w/FLAIR). We used (480/101/169) train/validation/test image pairs for OASIS, and (122/0/33, T1w), (154/0/39, T2w), and (163/0/50, FLAIR) for the RS. We supplemented the latter with 2796 (T1w/T2w) or 748 (FLAIR) non-repeated RS images for pre- or single-noisy training. For YADO training on the non-repeated HCP-A (656/64/0), we formulated pN2N by predicting a left-out echo of the multi-echo (ME) MPRAGE [35, 54] sequence.

To benchmark *real-world* denoising performance, we evaluated multiple external datasets with repeated acquisition: CHDI Track-HD/-ON [72], Kirby [40], HCP Epilepsy (HCP-E) [56], HCP Young Adults (HCP-YA) [76], and repeated motion-corrupted FLAIR [17]. These data included patients with temporal lobe epilepsy (HCP-E, $n=68$) and pre-/post-symptomatic Huntington’s disease (HD) in CHDI (48/31). For CHDI, we denoise low-quality scans (grade 3), while high-quality repeats (grade 1) serve as a reference. Similarly, we denoise still or voluntary-motion (nodd) FLAIR images, comparing to still, motion-corrected references [14, 17]. All image pairs were acquired as in-session or same-day repeats (except HCP-E, where re-acquisition was after up to 14 days).

We report SSIM [80] as the primary metric (PSNR as Suppl. Tabs. 14 and 15). Beyond image-level metrics, we assessed lesion-wise averaged absolute CNR [63] for PVS (T1w/T2w) and WMH (FLAIR) in the RS [46] (Eq. (15) in [63]). We also evaluated downstream WMH segmentation as lesion-detection F_1 -score [38] for *SHIVA* [73, 74]; reliable, automatic PVS segmentation remains unsolved [82]. In total, our evaluation includes images from $n = 673$ participants acquired across all major scanner vendors and 0.7–1 mm resolution (Tabs. 1 and 2).

5.2 Implementation details

YADO denoiser The *YADO denoiser* is implemented as a 2D U-Net [61] with five ResNet stages ([256, 256, 512, 512, 1024] channels) and self-attention ([0, 0, 2, 2, 4] heads), resulting in 213M parameters. We adopted a 2.5D setup using five-slice slabs and randomly sampling axial, coronal, or sagittal planes during training. For inference, we predict and average each plane. Training uses an L_2 loss with voxels outside a FastSurfer [22, 23] brain mask down-weighted by $\varepsilon=0.01$ with an AdamW [32] optimizer (LR 10^{-4} , batch size 96, see Suppl. Sec. 13.1). We concatenate co-acquired contrasts as channels [24] (T1w: T2w, T2w: T1w, FLAIR: T1w+T2w) after registration [16, 20, 25, 64]. We optimized the hyperparameter based on the OASIS-3 validation set, while using the weight’s EMA ($\beta=0.999$) as final model allows training without a validation set.

Renoise2Noise diffusion model The *ReN2N* DM employs the same backbone as YADO with an additional time-step embedding [24]. We train the DM

exclusively on RS data with a $\sqrt{\cos}$ noise schedule [55] for realistic re-noising [63] (Suppl. Sec. 13.2). We use a v -prediction target and train with Adam (LR 10^{-4}) using the same brain-masking loss as YADO. By default, we assume the availability of cross-contrast information to condition the DM like the guided (g)YADO. Yet, we also considered T1w-only ReN (Suppl. Sec. 12.5). For ReN2N training, we sampled four stochastic re-noised instances per input image, with shared late diffusion steps to reduce compute. ReN preprocessing takes *e.g.* 80 days of V100 compute for the 960 OASIS training images (Suppl. Tab. 3). We supplement the synthetic data with the acquired image, randomly selecting input and target from the five instances (Fig. 1c).

5.3 Modeling procedures

To enable a direct comparison of YADO’s pN2N or ReN2N to popular single-noisy objectives, we re-implemented Ner2N [53], N2Score [31], N2Void [4,37], S-N2V [8] (3×3 blind-spot mask), and Neigh2Neigh [26] with the same backbone as YADO and using repeats as individual training instances. To assert fair experiments, we extensively ablated hyper-parameters for all methods (Suppl. Sec. 13.3). In addition, we considered RED-WGAN [62] and regression-only CNN3D [62], BME-X [71], N2Contrast [78], and DDM² [83], for which we additionally considered improved *regression*-sampling [63]. We also implemented YODA [63] (regression only) as an indirect, translation-only denoiser with the same backbone as YADO. To model the inherent disadvantage of pN2N that requiring repeats, we supplemented the paired T2w/FLAIR RS training data with unpaired images (2796/748) for single-image denoising (excluding ReN2N). Yet, we pre-trained YADO and the ReN DM on the respective single-image set using modality translation as a proxy task [63]. BM4D [48], ANTs NLM [49], and the as-is, docker-distributed version of BME-X [71] were evaluated as training-free denoisers. We aggregated relative performance based on $\sqrt{n}$ -weighted averages of ranks within each task/dataset.

6 Results and Discussion

6.1 Denoiser benchmarking on T1w images

We compare 1 mm (OASIS-3-trained) and 0.8 mm (HCP-A) T1w denoising on routine RS images to physically noise-suppressed acquisitions ($N_{\text{Ex}}=7$) in Fig. 2 (more qualitative results in the Suppl.), while a comprehensive quantitative benchmark is provided in Tab. 1. We find that all YADO variants achieved a remarkable visual quality that closely resembles the high-quality ($N_{\text{Ex}}=7$, TA>45 min) reference images. Quantitatively, YADO outperforms all baselines in the rank-based analysis, whereby YADO-variants marked the best performing metric in each task. Remarkably, ReN2N improves over pN2N in 20 of 22 experiments, *i.e.* all but T2w-guided synthesis of 1. the internal OASIS and 2. CNR at PVS. This demonstrates that ReN2N not only extends pN2N training to a single-image domain but, moreover, improves model generalization.

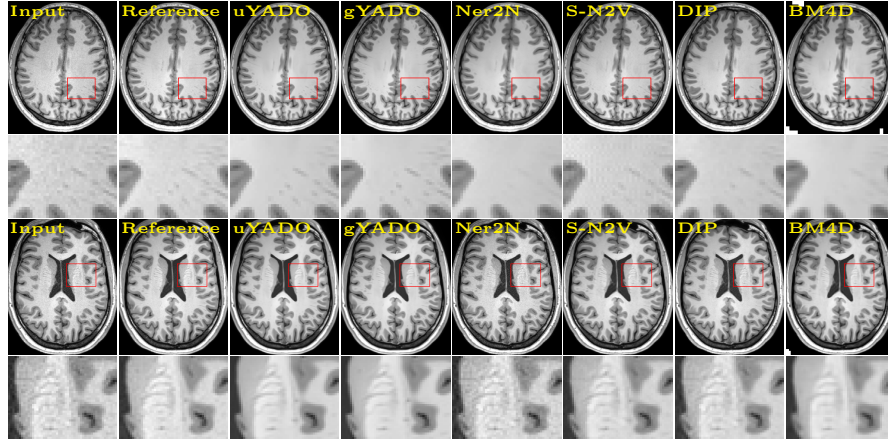


Fig. 2: T1w denoising results (upper: 1 mm, lower: 0.8 mm) for selected methods (more in Suppl., YADO: ReN2N) on routine images ($N_{\text{Ex}}=1$). The shown reference images were physically noise-suppressed ($N_{\text{Ex}}=7$, > 45 min acquisition time).

The comparison of guided (g) to unguided (u) YADO denoising demonstrates that co-acquired T2w images improve performance, except for PVS CNR (non-significant improvement) and Kirby. However, the high anisotropy of Kirby T2w images [41] directly explains this reduction in performance, as the T2w context becomes a liability. We also observed a substantially improved performance of ReN2N over pN2N for HCP-A training. We attribute this to the use of individual echos that possess slightly different contrasts and to the modified scope of removable noise due to the echo-based pN2N training formulation (see Suppl. Sec. 10).

Whereas YADO generally improves image quality and PVS contrast (with only two exceptions) over the Identity baseline (which does nothing), most baseline methods frequently fail to provide a significant improvement. In fact, *7/17 methods ranked worse than the Identity function*. While the principled N2N-derived methods show relatively consistent SSIM improvements, PVS CNR deteriorates in all but 0.8 mm (S-)N2Void [8]. N2Contrast [78], which uses simple, trainable bilateral filters, is the only competing method that improves all image quality measures. However, the benefit over Identity is relatively small and other filter-based methods (ANTs NLM [49], BM4D [48]) rank higher. Surprisingly, we found that BM4D [48] outperforms the most expensive learning-based methods, ranking second only behind S-N2V [8]. In addition, avoiding perceptual realism by removing the GAN-loss from RED-WGAN [62] or conducting *regression* DM-sampling [63] of DDM² [83] improves image quality. Finally, we observe large differences of nominal values for SSIM between datasets, obtaining nominally *worse* results in internal than in most external test sets. We attribute this to the aleatoric uncertainty, *i.e.* dataset-specific $c_{\mathbf{n}_{\text{ref}}}$ in Eq. (4). We trace the disappointing performance of traditional methods back to a trade-off between noise concealment and information preservation, limiting perfor-

Table 1: Quantitative T1w denoising results (% SSIM; except PVS: abs. CNR [63]). Formatting: **Bold**: best or statistically tied ($p \geq 5\%$, Wilcoxon vs. best, Bonferroni-corrected). **Gray background**: not significantly better than Identity. (Parentheses): guidance with anisotropic/low quality T2w. $\dagger$: synthesis and evaluation at 1 mm.

	OASIS-3 trained (ca. 1.0 mm)							HCP-A trained (0.8 mm)				all
Test dataset	SSIM (%)							SSIM (%)				rank
Resolution (mm)	OAS3	CHDI	Kirby	HCP-E	-YA	RS	PVS	HCP-E	-YA	RS	PVS	all
Vendor	S	S/P	P	GE	S	S	S	GE	S	S	S	-
n	169	100	21	100	100	33	89	100	100	33	79	512
uYADO (pN2N)	89.39 ± 4.69	94.25 ± 1.54	95.69 ± 0.85	89.65 ± 3.82	94.06 ± 0.69	93.83 ± 2.23	2.63 ± 0.81	87.05 ± 4.67	85.85 ± 1.42	90.44 ± 3.29	2.15 ± 0.40	4.55
uYADO (ReN2N)	89.41 ± 4.87	94.33 ± 1.59	95.99 ± 0.90	90.60 ± 3.38	94.50 ± 0.64	94.37 ± 2.39	2.74 ± 0.75	88.25 ± 4.24	86.66 ± 1.33	91.23 ± 2.95	2.32 ± 0.53	2.09
gYADO (pN2N)	90.03 ± 4.35	94.11 ± 1.42	(94.63) ± 0.76	89.96 ± 3.31	94.11 ± 0.66	94.31 ± 1.59	2.85 ± 1.66	86.87 ± 4.51	86.05 ± 1.43	89.93 ± 2.99	2.25 ± 0.54	4.17
gYADO (ReN2N)	89.77 ± 4.82	94.59 ± 1.47	(95.63) ± 0.78	90.91 ± 3.24	94.59 ± 0.60	94.59 ± 2.25	2.85 ± 1.26	88.74 ± 3.99	86.98 ± 1.28	91.69 ± 2.56	2.29 ± 0.60	1.46
Noisier2N [53]	88.59 ± 4.82	93.53 ± 1.71	95.58 ± 0.90	89.47 ± 3.28	93.91 ± 0.72	93.42 ± 2.60	2.28 ± 0.83	84.63 ± 5.13	82.65 ± 1.65	88.45 ± 3.91	2.10 ± 0.39	9.26
N2Score [31]	89.07 ± 5.10	93.77 ± 1.56	95.21 ± 0.88	87.89 ± 3.53	93.61 ± 0.67	92.75 ± 2.42	1.06 ± 0.53	79.97 ± 23.75	65.71 ± 36.25	90.24 ± 3.08	1.71 ± 0.54	12.24
N2Void [4, 37]	88.60 ± 5.20	93.78 ± 1.76	95.72 ± 0.99	88.19 ± 3.86	90.86 ± 0.76	92.63 ± 2.63	2.24 ± 0.62	86.99 ± 4.33	85.14 ± 1.51	90.11 ± 3.63	2.20 ± 0.47	8.58
S-N2Void [8]	88.69 ± 5.05	93.67 ± 1.74	95.59 ± 1.01	88.43 ± 3.73	91.38 ± 0.70	92.62 ± 2.51	2.17 ± 0.65	87.42 ± 4.26	85.89 ± 1.44	90.56 ± 3.43	2.22 ± 0.37	7.81
Neigh2Neigh [26]	87.94 ± 4.31	93.03 ± 1.53	94.59 ± 0.76	89.08 ± 3.27	93.61 ± 0.68	92.68 ± 2.31	2.31 ± 0.89	87.36 ± 4.04	85.40 ± 1.35	89.63 ± 2.84	1.83 ± 0.36	9.75
RED-WGAN [62]	86.35 ± 5.34	91.38 ± 1.95	93.42 ± 1.13	87.21 ± 3.50	89.48 ± 1.41	90.49 ± 3.19	2.18 ± 0.48	84.70 ± 4.80	83.53 ± 1.38	88.34 ± 3.44	1.88 ± 0.34	15.35
CNN3D [62]	87.79 ± 5.05	93.00 ± 1.82	95.11 ± 1.05	88.86 ± 3.61	92.62 ± 0.93	92.62 ± 2.87	2.25 ± 0.59	86.28 ± 4.63	84.51 ± 1.50	89.31 ± 3.49	1.95 ± 0.35	11.62
DDM ² [83]	51.05 ± 5.47	43.04 ± 6.29	39.29 ± 4.11	35.15 ± 7.08	39.49 ± 4.32	48.27 ± 4.56	0.16 ± 0.46	23.64 ± 4.65	30.25 ± 8.68	39.82 ± 5.29	0.02 ± 0.07	22.00
DDM ² Reg. [63, 83]	87.28 ± 4.58	91.14 ± 1.86	92.73 ± 0.93	84.10 ± 5.82	89.80 ± 1.29	91.28 ± 2.27	1.76 ± 0.67	72.24 ± 6.49	75.28 ± 5.31	84.41 ± 2.90	1.01 ± 0.45	17.70
BME-X [71]	83.10 ± 3.45	87.81 ± 1.64	86.64 ± 1.15	82.87 ± 5.20	88.38 ± 1.01	88.05 ± 2.26	1.55 ± 0.70	82.71 ± 4.68	82.17 ± 1.74	88.14 ± 2.92	1.56 ± 0.34	18.68
N2Contrast [78]	87.32 ± 5.64	93.12 ± 1.99	95.34 ± 1.17	87.98 ± 4.43	90.46 ± 1.07	92.16 ± 3.26	2.42 ± 0.50	85.51 ± 5.26	83.39 ± 1.72	88.97 ± 3.91	2.12 ± 0.35	11.82
YODA [63]	88.29 ± 4.46	87.95 ± 3.15	(70.80) ± 3.78	83.48 ± 4.61	86.87 ± 1.91	89.18 ± 2.67	2.17 ± 0.80	71.19 ± 6.88	80.11 ± 2.77	84.17 ± 3.55	1.69 ± 1.04	17.30
DIP [75]	87.87 ± 5.34	93.27 ± 1.80	95.31 ± 1.16	89.04 ± 4.09	92.23 ± 1.04	92.88 ± 3.07	2.33 ± 0.60	86.77 ± 4.90	84.86 ± 1.67	89.58 ± 3.72	1.97 ± 0.40	9.84
Pixel2Pixel [47]	85.76 ± 3.66	90.81 ± 1.42	92.16 ± 0.73	87.20 ± 3.54	91.39 ± 0.84	90.23 ± 2.22	1.41 ± 0.63	86.20 ± 4.23	83.87 ± 1.43	87.94 ± 2.75	1.21 ± 0.32	16.46
ANTs NLM [49]	87.89 ± 4.58	93.11 ± 1.69	94.97 ± 0.91	88.95 ± 3.75	93.21 ± 0.77	92.62 ± 2.64	2.36 ± 0.66	87.22 ± 4.42	85.42 ± 1.42	89.73 ± 3.24	1.91 ± 0.38	9.49
BM4D [48]	88.28 ± 4.52	93.40 ± 1.67	95.33 ± 0.90	89.37 ± 3.66	93.71 ± 0.73	92.93 ± 2.43	2.13 ± 0.83	87.81 ± 4.18	86.09 ± 1.40	90.27 ± 3.21	1.83 ± 0.39	7.96
BME-X (as-is) [71]	52.72 ± 5.62	55.39 ± 6.04	55.48 ± 3.88	46.47 ± 21.34	68.41 ± 5.72	53.13 ± 4.10	1.53 ± 0.61	35.33 ± 29.69	68.85 ± 5.20	64.45 ± 5.41	1.53 ± 0.61	20.52
Identity	86.91 ± 5.65	92.66 ± 2.09	95.03 ± 1.21	87.09 ± 4.46	89.50 ± 1.04	91.60 ± 3.25	2.33 ± 0.46	84.61 ± 5.18	82.68 ± 1.66	88.47 ± 3.90	2.10 ± 0.37	14.35

mance for degradation-based methods (see Suppl. Sec. 12.1). To further test how Ner2N and S-N2Void training is affected by unconcealed noise, we trained ablated versions performing input degradations, while using physical re-acquisition as a reference akin to pN2N (Sec. 12.2). Although this decorrelation improves the performance for Ner2N and S-N2Void, the overall performance gap to pN2N YADO does not close. Supplementing single-noisy baseline training with ReN images and thus accounting for ReN’s augmentation effect, could not also not achieve the same performance as YADO (Sec. 12.3). Thus, degradation-induced information loss and domain-shifts inevitably limit the performance of single-noisy training objectives such as Ner2N and (S-)N2Void.

In turn, simplifying ReN2N from DM-based renoising to just adding Rician noise to denoised images (cf. Eq. (6)) hampered the performance of the resulting denoiser (Sec. 12.4). This demonstrates that, beyond generally enabling single-noisy N2N-like training by providing training pairs, ReN’s realism imitating real-world noise is key to ReN2N’s performance improvements.

6.2 Generalization of denoisers to T2w and FLAIR

We prioritized the top-performing T1w denoising methods to constrain resource demands in the following experiments. We selected S-N2Void [8], Ner2N [53], DIP [75], and the efficient BM4D [48] and ANTs [49] denoisers. Table 2 and Fig. 3 show quantitative and qualitative results (with $N_{\text{Ex}}=3$ reference FLAIR). Qualitatively, YADO improves the delineation of structures in both FLAIR and T2w (Fig. 3 and Suppl. Fig. 20f). Conversely, Ner2N [53] and S-N2Void [8, 37] remove the noise insufficiently and maintain “structures” visible in the input images but not in the reference: both methods suffer from partial identity decay. Fixing this generalization weakness of Ner2N and blind-spot methods over YADO requires domain-specific hyperparameter optimization (compare Suppl. Sec. 12) equally mandating T2w/FLAIR optimizations for YADO. DIP [75] and BM4D [48], on the other hand, feature stronger noise removal at the expense of blurring, *i.e.* loss of structural information. Quantitatively, YADO constitutes the best performing method in each evaluation. For T2w denoising, we observed that dataset-internal ReN2N training improves generalization to unseen data sources over pN2N, but reduces performance in the internal test set, including the CNR for PVS. However, ReN2N can potentially also leverage

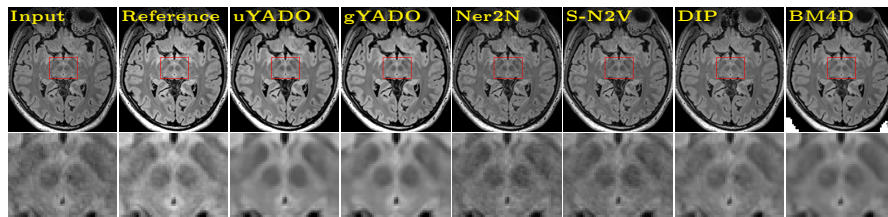


Fig. 3: FLAIR denoising results in the RS (reference: $N_{\text{Ex}}=3$, YADO: ReN2N).

Table 2: Quantitative T2w/FLAIR denoising results (% SSIM; except PVS and WMH: CNR [63]; % F_1 connected components [38] for WMH segmentation with SHIVA [73]). Formatting: **Bold**: best or statistically tied ($p \geq 5\%$, Wilcoxon vs. best, Bonferroni-corrected). Gray background: not significantly better than Identity. (Parentheses): synthesis/guidance with anisotropic/low quality T2w. [†]: synthesis and evaluation at 1 mm.

Test dataset	RS-trained T2w (0.8 mm)				RS-trained FLAIR (0.8 mm)					rank all
	SSIM (%)		CNR		SSIM (%)		CNR		F_1 [73]	
	RS	HCP-E	-YA	PVS	RS	still	nodd	WMH		
Resolution (mm)	0.8	0.8	0.7	0.8	0.8	1.0 [†]	1.0 [†]	1.0		0.8-1
Vendor	S	GE	S	S	S	S	S	S		-
n	39	100	100	79	50	10	10	52		383
uYADO (pN2N)	88.80 ± 10.44	90.21 ± 3.80	89.50 ± 1.48	2.69 ± 0.95	88.36 ± 6.30	84.89 ± 2.53	73.84 ± 6.98	3.14 ± 0.39	61.58 ± 21.53	5.60
uYADO (ReN2N)	88.58 ± 10.43	91.10 ± 3.46	89.76 ± 1.51	2.59 ± 0.99	88.25 ± 6.33	84.94 ± 2.50	73.88 ± 7.10	3.08 ± 0.33	61.02 ± 21.33	4.60
gYADO (pN2N)	89.65 ± 8.60	90.38 ± 3.89	89.24 ± 1.53	3.12 ± 0.87	89.60 ± 5.07	(69.94) ± 2.56	(62.75) ± 5.16	3.40 ± 0.45	64.23 ± 19.13	3.55
gYADO (ReN2N)	88.98 ± 9.69	91.43 ± 3.39	90.28 ± 1.43	2.91 ± 0.85	89.28 ± 5.49	(59.29) ± 3.07	(52.94) ± 4.40	3.33 ± 0.41	62.42 ± 21.01	3.63
Ner2N [53]	87.00 ± 11.66	90.23 ± 4.18	88.58 ± 1.79	2.21 ± 0.78	85.78 ± 7.05	83.21 ± 2.44	70.28 ± 7.57	3.34 ± 0.32	62.82 ± 19.95	7.65
S-N2V [37]	87.69 ± 11.35	90.83 ± 3.70	89.26 ± 1.52	2.38 ± 0.95	86.60 ± 7.02	83.58 ± 2.42	71.45 ± 7.43	3.36 ± 0.34	62.75 ± 20.17	5.52
DIP [75]	87.44 ± 11.41	90.95 ± 3.61	89.79 ± 1.59	2.29 ± 0.88	87.04 ± 6.95	83.90 ± 2.52	71.12 ± 7.56	3.28 ± 0.33	62.68 ± 20.11	5.30
ANTs NLM [49]	86.76 ± 10.47	90.27 ± 3.24	89.05 ± 1.42	2.32 ± 0.80	87.20 ± 6.38	83.52 ± 2.42	72.80 ± 6.78	3.37 ± 0.33	62.89 ± 19.96	6.00
BM4D [48]	87.28 ± 10.81	90.93 ± 3.37	89.68 ± 1.37	2.52 ± 3.18	87.76 ± 6.64	84.51 ± 2.36	73.87 ± 6.73	3.21 ± 0.33	62.83 ± 20.32	4.74
Identity	86.99 ± 11.67	90.14 ± 4.18	88.52 ± 1.79	2.22 ± 0.79	85.73 ± 7.06	83.18 ± 2.44	70.27 ± 7.58	3.37 ± 0.34	62.45 ± 20.19	8.41

unpaired data to increase the dataset and model size exploiting simple scaling laws (see Suppl. Sec. 12.6).

For FLAIR images, the unguided YADO versions generalize from 0.8 mm to regular (still), external 1 mm acquisitions [17], whereas voluntary motion deteriorated absolute and relative performance (tie with BM4D). Anisotropic T2w acquisitions [17] polluted guided YADO denoising (compare Kirby in Tab. 1). Yet, suitable, less corrupted T2w images would likely remedy motion corruption in other scenarios. In WMH-based analyses, we find that no denoiser significantly improved CNR, whereas only ANTs NLM denoising significantly improves downstream WMH detectability with SHIVA [73]. However, guided pN2N YADO achieves the best detectability, matching ANTs in a paired comparison. Conversely, the remaining YADO methods reduce WMH CNR and detectability, which we attribute to YADO denoising deviating more from the reference input images by imitating training 0.8 mm rather than inference 1 mm image appearances, contrary to the more preservative competing methods. Moreover, manual reference annotation can be biased by noise [63], which might partially penalize

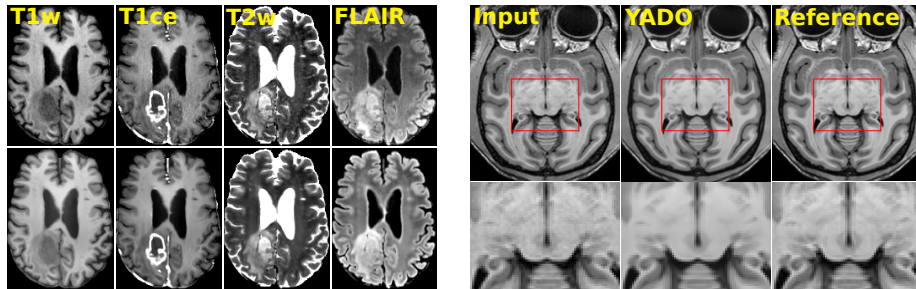


Fig. 4: YADO in two OOD scenarios: tumor images in BraTS [51] (left) and monkey brain images [9] (right; Input: $N_{\text{Ex}}=1$, Reference: $N_{\text{Ex}}=4$).

correct noise removal, especially since WMH shape and anatomy are more complex and diverse than PVS [5]. Whereas the best-case gYADO (pN2N) is able to over-compensate for this handicap, the noise-polluted reference labels might generally disfavor YADO which – unlike competing methods partially sustaining the input noise – deviates from their noisy inputs.

6.3 Generalization to out-of-distribution (OOD) data

To more broadly assess YADO for OOD data beyond HD (CHDI in Tab. 1), we tested guided, ReN2N-trained YADO on BraTS tumor images [2], including contrast-enhanced T1w (T1ce) denoising. Exemplary results are provided in Fig. 4a. Although glioma and, even more so, T1ce contrast deviate substantially from the training corpus, YADO sustains tumor anatomy and signal from contrast agent. Note that a quantitative performance evaluation is not possible due to a lack of independent reference images.

To further exaggerate the domain shift, we also denoise monkey brain images [9] (Fig. 4b). We observe that YADO removes noise while maintaining evident anatomical signal, demonstrating that YADO’s learned anatomical and structural priors are not tied to healthy human brain but leverage more subtle information. Beyond out-of-the-box generalization, this indicates that YADO checkpoints might constitute a valuable starting point for fine-tuning on other contrast modalities, body parts, disease types, or even species like monkeys or mice.

7 Conclusion

In this paper, we *re-think* MRI denoising by revisiting fundamental noise theory and evaluation practices. We derive a well-behaved training scenario, pN2N, directly grounded in established N2N theory [44], and introduce an unbiased evaluation framework that avoids rewarding noise replication. By evaluating on independently acquired *real* images, we show that several recent learning-based approaches fail to consistently surpass classical non-local filtering methods [48, 49], and some even underperform the identity mapping. We attribute

this gap in real-world performance to optimizing for synthetic, well-behaved noise distributions rather than accounting for the complexity of physical acquisition noise. This has been potentially amplified by innovation biases in previous benchmarking [28, 63].

By training on physical images with *real-world* noise, we establish YADO as a denoiser that removes all non-systematic effects, *i.e.* the true acquisition noise. Both pN2N and ReN2N achieve remarkable quality from routine acquisitions, as YADO-denoised images even match structural detail of physically, brute-force noise-suppressed images ($N_{\text{Ex}} = 3, 7$) while exceeding them in smoothness. Conditioning on co-acquired contrasts further boosts performance by exploiting complementary anatomical information already present in standard protocols. Whereas physical repeats are less common, even small repeat datasets – such as those collected in pilot or protocol-optimization studies [34, 76] – suffice to train (or fine-tune) powerful domain-specific YADO variants. When repeats are unavailable, ReN2N generalizes pN2N to unpaired data yet, at considerable one-time computational costs during training. Given the substantial gains in real-world image quality, we consider this trade-off well-justified.

Furthermore, the adoption of YADO in its published form promises improved denoising performance in many settings: While future work may improve gYADO’s flexibility (*e.g.* varying resolutions [23]) and resilience to low-quality [17, 41] or partial/different conditioning configurations (*e.g.* T1w+FLAIR, without T2w [52]) to leverage the full potential of YADO, uYADO already significantly improves image quality. However, with numerous results on the main structural contrasts (T1w/T2w/FLAIR), generalization to unseen sequences (*e.g.* T1ce), and pathological subpopulations (*e.g.* epilepsy and HD), the practical utility of YADO for MRI denoising *in the wild* is already empirically established.

Acknowledgments Data were provided in part by OASIS-3, PI: T. Benzinger, D. Marcus, J. Morris; NIH P30 AG066444, P50 AG00561, P30 NS09857781, P01 AG026276, P01 AG003991, R01 AG043434, UL1 TR000448, R01 EB009352; by the HCP, WU–Minn Consortium (PIs: David Van Essen and Kamil Ugurbil; 1U54MH091657), funded by the NIH Blueprint for Neuroscience Research (via 16 NIH Institutes and Centers) and by the McDonnell Center for Systems Neuroscience at WashU; by the NIA (NIH; U01AG052564) and supported by funds provided by the McDonnell Center for Systems Neuroscience, WashU; by study participants of TRACK-HD/ON generously contributed and made available by Prof. S. Tabrizi and Dr. E. Wild (UCL) and funded by the CHDI Foundation. HCP-Aging was obtained from DOI:10.15154/1520707. The authors thank the Rhineland Study group (PI: M.M.B. Breteler, D. Früh) for supporting the data acquisition and management, and all study participants.

This work was supported in part by German Federal Ministry of Research, Technology and Space (BMFTR) under Grant 031L0206; in part by NIH under Grant R01MH131586, Grant R01MH130899, and Grant R01AG064027; in part by the Chan Zuckerberg Initiative (CZI) under Grant EOSS5 2022-252594; and in part by the Helmholtz Foundation Model Initiative (Human Radiome Project).

References

1. Aja-Fernández, S., Tristán-Vega, A.: A review on statistical noise models for magnetic resonance imaging. Tech. Rep (2013)
2. Baid, U., et al.: The rsna-asnr-miccai braTS 2021 benchmark on brain tumor segmentation and radiogenomic classification. arXiv:2107.02314 (2021)
3. Barnes, S.R., Haacke, E.M., Ayaz, M., Boikov, A.S., Kirsch, W., Kido, D.: Semiautomated detection of cerebral microbleeds in magnetic resonance images. *Magnetic resonance imaging* **29**(6), 844–852 (2011)
4. Batson, J., Royer, L.: Noise2self: Blind denoising by self-supervision. In: International conference on machine learning. pp. 524–533. PMLR (2019)
5. Bernal, J., Valdés-Hernández, M.D., Escudero, J., Duarte, R., Ballerini, L., Bastin, M.E., Deary, I.J., Thrippleton, M.J., Touyz, R.M., Wardlaw, J.M.: Assessment of perivascular space filtering methods using a three-dimensional computational model. *Magnetic resonance imaging* **93**, 33–51 (2022)
6. Blau, Y., Michaeli, T.: The perception-distortion tradeoff. In: CVPR (2018)
7. Bookheimer, S.Y., et al.: The lifespan human connectome project in aging: An overview. *NeuroImage* **185**, 335–348 (2019). <https://doi.org/10.1016/j.neuroimage.2018.10.009>
8. Broaddus, C., Krull, A., Weigert, M., Schmidt, U., Myers, G.: Removing structured noise with self-supervised blind-spot networks. In: 2020 IEEE 17th international symposium on biomedical imaging (ISBI). pp. 159–163. IEEE (2020)
9. Castro, V.D., Cottureau, B., Durand, J.B.: "structural, diffusion and rs-functional MRI data in macaque monkeys" (2022). <https://doi.org/doi:10.18112/openneuro.ds003989.v1.0.0>
10. Cerri, S., Munk, A., Ambsdorf, J., Machnio, J., Llambias, S.N., Nersesjan, V., Krag, C.H., Liu, P., García, P.R., Ghazi, M.M., et al.: A large-scale heterogeneous 3D magnetic resonance brain imaging dataset for self-supervised learning. arXiv preprint arXiv:2506.14432 (2025)
11. Chen, Z., Pawar, K., Ekanayake, M., Pain, C., Zhong, S., Egan, G.F.: Deep learning for image enhancement and correction in magnetic resonance imaging—state-of-the-art and challenges. *Journal of Digital Imaging* **36**(1), 204–230 (2023)
12. Chung, H., Lee, E.S., Ye, J.C.: Mr image denoising and super-resolution using regularized reverse diffusion. *IEEE T-MI* **42**(4), 922–934 (2022)
13. Collins, D.L., Zijdenbos, A.P., Kollokian, V., Sled, J.G., Kabani, N.J., Holmes, C.J., Evans, A.C.: Design and construction of a realistic digital brain phantom. *IEEE transactions on medical imaging* **17**(3), 463–468 (2002)
14. Eichhorn, H., Frost, R., Kongsgaard, A., Kettless, K., Slipsager, J., Glimberg, S., Shekhrjka, N., Tisdall, M., Wighton, P., Borgwardt, L.: Evaluating the performance of markerless prospective motion correction and selective reacquisition in a general clinical protocol for brain mri. *PsyArXiv* (2022)
15. Fadnavis, S., Batson, J., Garyfallidis, E.: Patch2self: Denoising diffusion mri with self-supervised learning. *NeurIPS* **33**, 16293–16303 (2020)
16. Fischl, B.: Freesurfer. *NeuroImage* **62**(2), 774–781 (2012)
17. Ganz, M., Eichhorn, H.: "datasets with and without deliberate head movements for evaluating the performance of markerless prospective motion correction and selective reacquisition in a general clinical protocol for brain mri" (2022). <https://doi.org/doi:10.18112/openneuro.ds004332.v1.0.0>
18. Glasser, M.F., et al.: The human connectome project’s neuroimaging approach. *Nature neuroscience* **19**(9), 1175–1187 (2016)

19. Goodfellow, I., et al.: Generative adversarial nets. In: NeurIPS (2014)
20. Greve, D.N., Fischl, B.: Accurate and robust brain image alignment using boundary-based registration. *NeuroImage* **48**(1), 63–72 (2009)
21. Gudbjartsson, H., Patz, S.: The Rician distribution of noisy MRI data. *Magn. Reson. Med.* **34**(6), 910–914 (1995)
22. Henschel, L., Conjeti, S., Estrada, S., Diers, K., Fischl, B., Reuter, M.: FastSurfer - A fast and accurate deep learning based neuroimaging pipeline. *NeuroImage* **219**, 117012 (2020)
23. Henschel, L., Kügler, D., Reuter, M.: FastSurferVINN: Building resolution-independence into deep learning segmentation methods—a solution for highres brain MRI. *NeuroImage* **251**, 118933 (2022)
24. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: NeurIPS (2020)
25. Hoopes, A., Mora, J.S., Dalca, A.V., Fischl, B., Hoffmann, M.: Synthstrip: Skull-stripping for any brain image. *NeuroImage* **260**, 119474 (2022)
26. Huang, T., Li, S., Jia, X., Lu, H., Liu, J.: Neighbor2neighbor: Self-supervised denoising from single noisy images. In: CPVR. pp. 14781–14790 (2021)
27. Ilesanmi, A.E., Ilesanmi, T.O.: Methods for image denoising using convolutional neural network: a review. *Complex & Intelligent Systems* **7**(5), 2179–2198 (2021)
28. Isensee, F., Jäger, P.F., Full, P.M., Vollmuth, P., Maier-Hein, K.H.: nnU-Net for brain tumor segmentation. In: MICCAI (2021)
29. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: CVPR (2017)
30. Karras, T., Aittala, M., Aila, T., Laine, S.: Elucidating the design space of diffusion-based generative models. *NeurIPS* **35**, 26565–26577 (2022)
31. Kim, K., Ye, J.C.: Noise2score: tweedie’s approach to self-supervised image denoising without clean images. *Advances in Neural Information Processing Systems* **34**, 864–874 (2021)
32. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. *arXiv:1412.6980* (2014)
33. Klauschen, F., Goldman, A., Barra, V., Meyer-Lindenberg, A., Lundervold, A.: Evaluation of automated brain mr image segmentation and volumetry methods. Tech. rep., Wiley Online Library (2009)
34. Koch, A., et al.: Versatile mri acquisition and processing protocol for population-based neuroimaging. *Nature Protocols* pp. 1–23 (2024)
35. Van der Kouwe, A.J., Benner, T., Salat, D.H., Fischl, B.: Brain morphometry with multiecho mprage. *Neuroimage* **40**(2), 559–569 (2008)
36. Krueger, G., Granziera, C., Jack Jr, C.R., Gunter, J.L., Littmann, A., Mortamet, B., Kannengiesser, S., Sorensen, A.G., Ward, C.P., Reyes, D.A., et al.: Effects of mri scan acceleration on brain volume measurement consistency. *Journal of Magnetic Resonance Imaging* **36**(5), 1234–1240 (2012)
37. Krull, A., Buchholz, T.O., Jug, F.: Noise2void-learning denoising from single noisy images. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2129–2137 (2019)
38. Kuijf, H.J., et al.: Standardized assessment of automatic segmentation of white matter hyperintensities and results of the wmh segmentation challenge. *IEEE T-MI* **38**(11), 2556–2568 (2019)
39. LaMontagne, P.J., Benzinger, T.L., Morris, J.C., Keefe, S., Hornbeck, R., Xiong, C., Grant, E., Hassenstab, J., Moulder, K., Vlassenko, A.G., et al.: Oasis-3: longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and alzheimer disease. *medrxiv* pp. 2019–12 (2019)

40. Landman, B.A., Huang, A.J., Gifford, A., Vikram, D.S., Lim, I.A.L., Farrell, J.A.D., Bogovic, J.A., Hua, J., Chen, M., Jarso, S., et al.: Multi-parametric neuroimaging reproducibility: a 3-T resource study. *NeuroImage* **54**(4), 2854–2866 (2011). <https://doi.org/10.1016/j.neuroimage.2010.11.047>
41. Landman, B.A., Huang, A.J., Gifford, A., Vikram, D.S., Lim, I.A.L., Farrell, J.A., Bogovic, J.A., Hua, J., Chen, M., Jarso, S., et al.: Multi-parametric neuroimaging reproducibility: a 3-T resource study. *NeuroImage* **54**(4), 2854–2866 (2011)
42. Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al.: Photo-realistic single image super-resolution using a generative adversarial network. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 4681–4690 (2017)
43. Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al.: Photo-realistic single image super-resolution using a generative adversarial network. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 4681–4690 (2017)
44. Lehtinen, J., et al.: Noise2noise: Learning image restoration without clean data. In: *ICML* (2018)
45. Li, S., Zhou, J., Liang, D., Liu, Q.: Mri denoising using progressively distribution-based neural network. *Magnetic resonance imaging* **71**, 55–68 (2020)
46. Lohner, V., et al.: Relation between sex, menopause, and white matter hyperintensities: the Rhineland study. *Neurology* **99**(9), e935–e943 (2022)
47. Ma, Q., Jiang, J., Zhou, X., Liang, P., Liu, X., Ma, J.: Pixel2pixel: A pixelwise approach for zero-shot single image denoising. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2025)
48. Maggioni, M., Katkovnik, V., Egiazarian, K., Foi, A.: Nonlocal transform-domain filter for volumetric data denoising and reconstruction. *IEEE transactions on image processing* **22**(1), 119–133 (2012)
49. Manjón, J.V., Coupé, P., Martí-Bonmatí, L., Collins, D.L., Robles, M.: Adaptive non-local means denoising of mr images with spatially varying noise levels. *Journal of Magnetic Resonance Imaging* **31**(1), 192–203 (2010)
50. Mansour, Y., Heckel, R.: Zero-shot noise2noise: Efficient image denoising without any data. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 14018–14027 (2023)
51. Menze, B.H., et al.: The multimodal brain tumor image segmentation benchmark (brats). *IEEE T-MI* **34**(10), 1993–2024 (2014)
52. Miller, K.L., Alfaro-Almagro, F., Bangerter, N.K., Thomas, D.L., Yacoub, E., Xu, J., Bartsch, A.J., Jbabdi, S., Sotiropoulos, S.N., Andersson, J.L., et al.: Multimodal population brain imaging in the uk biobank prospective epidemiological study. *Nature neuroscience* **19**(11), 1523–1536 (2016)
53. Moran, N., Schmidt, D., Zhong, Y., Coady, P.: Noisier2noise: Learning to denoise from unpaired noisy data. In: *CVPR*. pp. 12064–12072 (2020)
54. Mugler III, J.P., Brookeman, J.R.: Three-dimensional magnetization-prepared rapid gradient-echo imaging (3d mp rage). *Magnetic resonance in medicine* **15**(1), 152–157 (1990)
55. Nichol, A.Q., Dhariwal, P.: Improved denoising diffusion probabilistic models. In: *International conference on machine learning*. pp. 8162–8171. PMLR (2021)
56. NIMH Data Archive, Connectome Coordination Facility: The epilepsy connectome project (ecp). NIMH Data Archive collection (2019). <https://doi.org/10.15154/4wpp-ya55>, collection DOI with NIH award U01NS093650. Accessed 2026-02-26.

57. Pfaff, L., Hossbach, J., Preuhs, E., Wagner, F., Arroyo Camejo, S., Kannengiesser, S., Nickel, D., Wuerfl, T., Maier, A.: Self-supervised mri denoising: leveraging stein’s unbiased risk estimator and spatially resolved noise maps. *Scientific Reports* **13**(1), 22629 (2023)
58. Pfaff, L., Wagner, F., Hossbach, J., Preuhs, E., Maul, N., Thies, M., Denzinger, F., Nickel, D., Wuerfl, T., Maier, A.: Robust multi-contrast mri denoising using trainable bilateral filters without noise-free targets. In: 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI). pp. 1–5. IEEE (2023)
59. Pfaff, L., et al.: No-New-Denoiser: A critical analysis of diffusion models for medical image denoising. In: MICCAI (2024)
60. Pham, W., Lynch, M., Spitz, G., O’Brien, T., Vivash, L., Sinclair, B., Law, M.: A critical guide to the automated quantification of perivascular spaces in magnetic resonance imaging. *Frontiers in neuroscience* **16**, 1021311 (2022)
61. Pinaya, W.H., et al.: Generative ai for medical imaging: extending the monai framework. *arXiv:2307.15208* (2023)
62. Ran, M., Hu, J., Chen, Y., Chen, H., Sun, H., Zhou, J., Zhang, Y.: Denoising of 3d magnetic resonance images using a residual encoder–decoder wasserstein generative adversarial network. *Medical image analysis* **55**, 165–180 (2019)
63. Rassmann, S., K ugler, D., Ewert, C., Reuter, M.: Regression is all you need for medical image translation. *IEEE Transactions on Medical Imaging* (2026)
64. Reuter, M., Rosas, H.D., Fischl, B.: Highly accurate inverse consistent registration: a robust approach. *Neuroimage* **53**(4), 1181–1196 (2010)
65. Reuter, M., Tisdall, M.D., Qureshi, A., Buckner, R.L., van der Kouwe, A.J., Fischl, B.: Head motion during mri acquisition reduces gray matter volume and thickness estimates. *Neuroimage* **107**, 107–115 (2015)
66. Saharia, C., et al.: Image super-resolution via iterative refinement. *IEEE PAMI* **45**(4), 4713–4726 (2022)
67. Saharia, C., et al.: Palette: Image-to-image diffusion models. In: ACM SIGGRAPH (2022)
68. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: ICML. pp. 2256–2265 (2015)
69. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: ICLR (2021)
70. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. In: ICLR (2021)
71. Sun, Y., Wang, L., Li, G., Lin, W., Wang, L.: A foundation model for enhancing magnetic resonance images and downstream segmentation, registration and diagnostic tasks. *Nature Biomedical Engineering* **9**(4), 521–538 (2025)
72. Tabrizi, S., et al.: Biological and clinical changes in premanifest and early stage huntington’s disease in the track-hd study: the 12-month longitudinal analysis. *The Lancet Neurology* **10**(1), 31–42 (2011)
73. Tsuchida, A., Boutinaud, P., Verrecchia, V., Tzourio, C., Debette, S., Joliot, M.: Early detection of white matter hyperintensities using shiva-wmh detector. *Human Brain Mapping* **45**(1), e26548 (2024)
74. Tustison, N.J., et al.: N4ITK: improved N3 bias correction. *IEEE T-MI* **29**(6), 1310–1320 (2010)
75. Ulyanov, D., Vedaldi, A., Lempitsky, V.: Deep image prior. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 9446–9454 (2018)

76. Van Essen, D.C., Smith, S.M., Barch, D.M., Behrens, T.E., Yacoub, E., Ugurbil, K., Consortium, W.M.H., et al.: The wu-minn human connectome project: an overview. *NeuroImage* **80**, 62–79 (2013)
77. Wagner, F., Thies, M., Denzinger, F., Gu, M., Patwari, M., Ploner, S., Maul, N., Pfaff, L., Huang, Y., Maier, A.: Trainable joint bilateral filters for enhanced prediction stability in low-dose ct. *Scientific Reports* **12**(1), 17540 (2022)
78. Wagner, F., Thies, M., Pfaff, L., Maul, N., Pechmann, S., Gu, M., Utz, J., Aust, O., Weidner, D., Neag, G., et al.: Noise2contrast: Multi-contrast fusion enables self-supervised tomographic image denoising. In: *International Conference on Information Processing in Medical Imaging*. pp. 771–782. Springer (2023)
79. Wang, D., Doddrell, D.M., Cowin, G.: A novel phantom and method for comprehensive 3-dimensional measurement and correction of geometric distortion in magnetic resonance imaging. *Magnetic resonance imaging* **22**(4), 529–542 (2004)
80. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* **13**(4), 600–612 (2004)
81. Wu, L., Hu, S., Liu, C.: Denoising of 3d brain mr images with parallel residual learning of convolutional neural network using global and local feature extraction. *Computational Intelligence and Neuroscience* **2021**(1), 5577956 (2021)
82. Wu, Y., Zhang, Y., Dong, Z., Ji, F., Tan, A.S., Tan, G., Tang, S., Chen, H., Chen, Z., Ng, E.K.K., et al.: Standardized evaluation of automatic methods for perivascular spaces segmentation in mri-miccai 2024 challenge results. *arXiv preprint arXiv:2512.18197* (2025)
83. Xiang, T., Yurt, M., Syed, A.B., Setsompop, K., Chaudhari, A.: DDM²: Self-supervised diffusion MRI denoising with generative diffusion models. In: *The Eleventh International Conference on Learning Representations* (2023), <https://openreview.net/forum?id=0vqjc50HfcC>
84. Zaitsev, M., Maclaren, J., Herbst, M.: Motion artifacts in MRI: A complex problem with many partial solutions. *Journal of Magnetic Resonance Imaging* **42**(4), 887–901 (2015)
85. Zhou, L., Zhou, Z., Huang, X., Wang, H., Zhang, X., Li, G.: Neighboring slice noise2noise: Self-supervised medical image denoising from single noisy image volume. *arXiv preprint arXiv:2411.10831* (2024)
86. Zhu, Y., Pan, X., Lv, T., Liu, Y., Li, L.: Desn: An unsupervised mr image denoising network with deep image prior. *Theoretical Computer Science* **880**, 97–110 (2021)
87. Zong, X., Nanavati, S., Hung, S.C., Li, T., Lin, W.: Effects of motion and retrospective motion correction on the visualization and quantification of perivascular spaces in ultrahigh resolution T2-weighted images at 7T. *Magnetic Resonance in Medicine* **86**(4), 1944–1955 (2021)

Supplementary

8 Analyzing the effect of noisy references images on denoiser evaluation

To illustrate the effect of noise in model evaluation, we assume here the MSE as metric. While the same effect follows for the PSNR directly, similar observations could be made for the SSIM.

We denote the true error of f_θ w.r.t. a clean image X' as

$$\text{error}_{X_{\text{in}}, X'}(f_\theta) := f_\theta(X_{\text{in}}) - X'. \quad (7)$$

We can then decompose the MSE w.r.t. a noisy reference image $X_{\text{ref}} = X' + \mathbf{n}_{\text{ref}}$ into

$$\text{MSE}(f(X), X) = \mathbb{E}_{X_{\text{in}}, \text{ref} \sim P_X} [\|f_\theta(X_{\text{in}}) - X_{\text{ref}}\|_2^2] \quad (8)$$

$$= \mathbb{E} [\|f_\theta(X' + \mathbf{n}_{\text{in}}) - (X' + \mathbf{n}_{\text{ref}})\|_2^2] \quad (9)$$

$$= \mathbb{E} [\|X' + \text{error}_{X_{\text{in}}, X'}(f_\theta) - X' - \mathbf{n}_{\text{ref}}\|_2^2] \quad (10)$$

$$= \mathbb{E} [\|\text{error}_{X_{\text{in}}, X'}(f_\theta)\|_2^2] \quad (11)$$

$$+ \mathbb{E} [\|\mathbf{n}_{\text{ref}}\|_2^2] \quad (12)$$

$$- 2 \mathbb{E} [(\text{error}_{X_{\text{in}}, X'}(f_\theta))^\top \mathbf{n}_{\text{ref}}] \quad (13)$$

While the term (11) represents the true epistemic error (*i.e.* $\text{MSE}(f_\theta(X), X')$), (12) constitutes the fixed aleatoric error due to $\mathbf{n}_{\text{ref}}$. Note that this error is practically unavoidable due to ubiquitous noise in the reference image (*i.e.* $\mathbf{n}_{\text{ref}} \gg 0$ as $N_{\text{Ex}} < \infty$). Hence, even an ideal denoiser $f_{\text{ideal}}(X) = X'$, *i.e.* one with $\text{error}_{X_{\text{in}}, X'}(f) = 0$, cannot avoid this error, and therefore this term corresponds to the aleatoric noise $c_{\bar{\mathbf{n}}_{\text{ref}}}$ in Eq. (4).

Conversely, if we assume that $\mathbf{n}_{\text{in}}, \mathbf{n}_{\text{ref}}$ are dependent, $\mathbf{n}_{\text{ref}}$ can be partially known, *i.e.* $\exists f : \mathbb{E} [(\text{error}_{X_{\text{in}}, X'}(f))^\top \mathbf{n}_{\text{ref}}] \stackrel{!}{>} 0$, such that the covariance term (13) becomes non-zero. *E.g.* assuming $\mathbf{n}_{\text{in}} \approx \mathbf{n}_{\text{ref}} > 0$, an identity denoiser $f_{\text{I}}(X) = X$ would be evaluated as

$$\text{MSE}(f_{\text{I}}(X_{\text{in}}), X_{\text{ref}}) = \mathbb{E} [\|f_{\text{I}}(X_{\text{in}}) - X_{\text{ref}}\|_2^2] \quad (14)$$

$$= \mathbb{E} [\|(X' + \mathbf{n}_{\text{in}}) - (X' + \mathbf{n}_{\text{ref}})\|_2^2] \quad (15)$$

$$= \mathbb{E} [\|\mathbf{n}_{\text{in}} - \mathbf{n}_{\text{ref}}\|_2^2] \quad (16)$$

$$\approx 0. \quad (17)$$

Thus, this evaluation would indicate an excellent performance for f_I .
In contrast, analyzing the ideal denoiser f_{ideal} in the same scenario yields

$$\text{MSE}(f_{\text{ideal}}(X_{\text{in}}, X_{\text{ref}}) = \text{MSE}(f_{\text{ideal}}(X_{\text{in}}, X') + c_{\tilde{\mathbf{n}}_{\text{ref}}} = c_{\tilde{\mathbf{n}}_{\text{ref}}}, \quad (18)$$

i.e. the aleatoric error term (12) as MSE. Yet, in the evaluation for f_I we would get

$$0 \stackrel{(14)}{\approx} \text{MSE}(f_I(X_{\text{in}}, X_{\text{ref}}) \quad (19)$$

$$< \mathbb{E}[\|\mathbf{n}_{\text{ref}}\|_2^2] = c_{\mathbf{n}_{\text{ref}}} \quad (20)$$

$$= \text{MSE}(f_{\text{ideal}}(X_{\text{in}}, X_{\text{ref}}). \quad \not\Leftarrow \quad (21)$$

Thus, in such an evaluation, simply doing nothing would achieve a lower MSE (*i.e.* higher PSNR) than a hypothetical denoiser f_{ideal} that predicts the perfect clean image X' .

However, let $\tilde{X}_{\text{ref}} = X' + \tilde{\mathbf{n}}_{\text{ref}}$ be an independent reference image, with $\tilde{\mathbf{n}}_{\text{ref}} \perp \mathbf{n}_{\text{in}}$ and $\mathbb{E}[\tilde{\mathbf{n}}_{\text{ref}}] = 0$. Then

$$\text{MSE}(f_I(X_{\text{in}}, \tilde{X}_{\text{ref}}) = \text{MSE}(f_I(X_{\text{in}}, X') + c_{\tilde{\mathbf{n}}_{\text{ref}}}, \quad (22)$$

as the noise is irrecoverable, *i.e.* the term (13) becomes 0. Hence,

$$\text{MSE}(f_{\text{ideal}}(X_{\text{in}}, \tilde{X}_{\text{ref}}) = c_{\tilde{\mathbf{n}}_{\text{ref}}} < \mathbb{E}[\|\mathbf{n}_{\text{in}}\|_2^2] + c_{\tilde{\mathbf{n}}_{\text{ref}}} = \text{MSE}(f_I(X_{\text{in}}, \tilde{X}_{\text{ref}}), \quad (23)$$

restoring correct ordering by eliminating the bias induced by correlated reference noise.

Similarly, assuming f_{θ} is evaluated by adding synthetic noise $\mathbf{n}_{\text{syn}}$ to an image $X = X' + \mathbf{n}_{\text{phys}}$ but re-using it as $X = X_{\text{ref}}$. This evaluates

$$\text{MSE}(f_{\theta}(X_{\text{in}}, X_{\text{ref}}) = \text{MSE}(f_{\theta}(X' + \mathbf{n}_{\text{phys}} + \mathbf{n}_{\text{syn}}, X' + \mathbf{n}_{\text{phys}}), \quad (24)$$

and thereby only assesses the ability to remove synthetic noise $\mathbf{n}_{\text{syn}}$, which does not constitute a proof for real-world denoising abilities, *i.e.* actually removing $\mathbf{n}_{\text{phys}}$ to restore X' . Moreover, assuming stronger effective physical than synthetic noise, *i.e.* $\mathbb{E}[\|\mathbf{n}_{\text{phys}}\|_2^2] > \mathbb{E}[\|\mathbf{n}_{\text{syn}}\|_2^2]$, the identity denoiser $f_I(X) = X$ outperforms the ideal denoiser $f_{\text{ideal}}(X) = X'$, since

$$\text{MSE}(f_I(X_{\text{in}}, X_{\text{ref}}) = \mathbb{E}[\|\mathbf{n}_{\text{syn}}\|_2^2] \quad (25)$$

$$< \mathbb{E}[\|\mathbf{n}_{\text{phys}}\|_2^2] \quad (26)$$

$$= \text{MSE}(f_{\text{ideal}}(X_{\text{in}}, X_{\text{ref}}). \quad \not\Leftarrow \quad (27)$$

Thus, again, doing nothing would be evaluated better than perfect denoising.

9 Comparison of denoiser requirements

Table 3: Comparison of requirements at training and inference time. VRAM and time requirements (Nvidia V100) are stated for OASIS-3 (960 training samples, 1 mm resolution). repeat/m.m.: requires physical multi-image acquisition/multi-modality input for training/inference.

[†] additionally requires CPU-heavy motion simulation (ca. 3d on 144 core server node)

Requirements	repeat m.m.		Training		Inference			
	repeat	m.m.	time	VRAM [GB]	repeat	m.m.	time	VRAM [GB]
gYADO (pN2N)	y	n	11d	5.0	n	n	16 s	1.8
uYADO (pN2N)	y	y	11d		n	y		
g/uYADO (ReN2N)	n	y / n	(11+80)d		n	y / n		
N2Void [8, 37]								
Neigh2Neigh [26]	n	n	11d	5.0	n	n	16 s	1.8
Ner2N [53]								
N2Score [31]								
BME-X [71]	n	n	28d [†]	3.0	n	n	15s	1.2
RED-WGAN [62]	n	n	13d	3.7	n	n	5 s	0.6
CNN3D [62]	n	n	7d	3.3	n	n	5 s	0.6
DDM ² [83] (Diff.)	n	y	(1+2)d	4.0	n	y	15 min	1.0
DDM ² [83] (Regr. [63])					n	y	20 s	1.0
N2Contrast [78]	n	y	24d	7.8	n	n	1 s	0.3
YODA [63]	n	y	11d	5.0	n	n	16 s	1.8
DIP [75, 86]			-		n	n	> 1h	1.3
Pixel2Pixel [47]			-		n	n	> 1h	0.7
ANTs NLM [49]			-		n	n	1 min	-
BM4D [48]			-		n	n	8 min	-
BME-X (as-is/full) [71]			-		n	n	13 min	16

10 Denoiser learning from ME-MPRAGE

Unlike default physical N2N (pN2N) training on repeated data from OASIS-3 and the RS, we used individual ME-MPRAGE echos for training on HCP-A. This inherently shifts the scope of training data, as random k -space corruption due *e.g.* motion is shared between echos and, thus, inter-predictable.

As a result, echo-based pN2N training encourages replicating these artifacts (Fig. 5), which is consistent with the theory established in (5). While motion rings are protocol-specific depending on *e.g.* k -space traversal, we observed that similar motion corruptions are also sustained or even enhanced in other datasets (Fig. 5, lower). Nevertheless, areas unaffected by motion rings were cleaned, *i.e.* the suppression of other noise sources such as thermal noise remained functional. Note that motion is random and irrecoverable, *i.e.* constitutes noise (*i.e.* part of $\mathbf{n}$) rather than signal (X' in Eq. (2)). Thus, in an unbiased, re-acquisition based evaluation, replicating the artifacts lowers the assessed performance.

Furthermore, note that all competing methods were similarly unable to remove motion ringing artifacts, except BME-X which is particularly trained on simulated motion. We attribute this denoiser failure to the spatial correlation of the ringing (Fig. 5). However, ReN2N-trained YADO reduced (internal HCP-A, Fig. 5, upper) or completely removed motion rings (RS, Fig. 5, lower) indicating that correlated motion noise is usually reduced/completely destroyed in the ReN augmentation, making motion unpredictable during training and, thus, inference. Note that, in contrast to BME-X, ReN2N YADO avoids blurring and reconstructs more details.

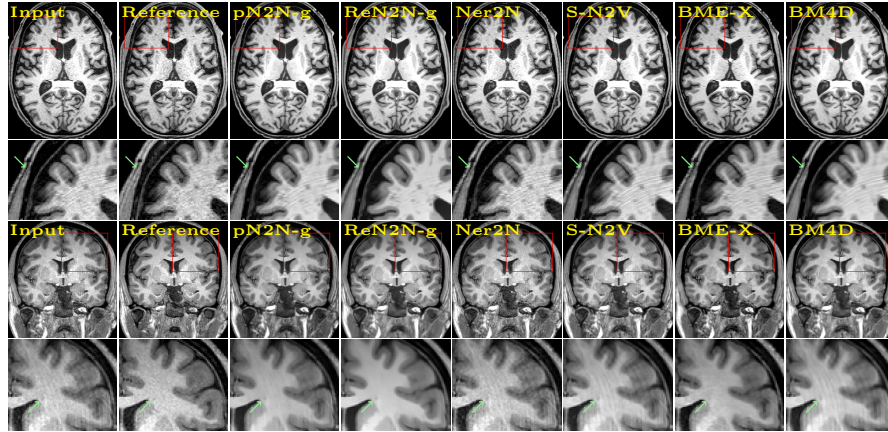


Fig. 5: Predictability of motion artifact in ME-MPRAGE training. Denoising results for guided YADO trained as physical N2N (pN2N) or ReN2N (both guided) is compared to selected competing denoisers on high-motion images from the HCP-A (upper) and RS (lower). For HCP-A, we used echoes 2-4 as input, and echo 1 as reference (same excitation), whereas in the RS we show an independent re-acquisition.

Furthermore, note that individual echos show a shifting contrast *e.g.* due to $T2^*$ effects (compare the dura mater in Fig. 5, upper, between input e2-4 vs. the reference e1). Thus, ME-MPRAGE denoising training becomes predicting a different set of echos, incentivizing a slight miss-match between the in- and output contrast. This contrast deviation might be exacerbated in the presence of T2w guidance similar to Kirby, explaining the partial under-performance of guided vs. unguided pN2N training as reported in Tab. 1.

Whereas, generally, this indicates a inferior performance for within-acquisition denoising, we note that this training formulation possesses the inherent advantage of temporal consistency as time between echos is in the range of milliseconds rather than minutes between re-acquisitions. Thus, echo-based training does not rely on registration avoiding the resulting uncertainties or errors (which, however, cannot be evaluated using repeats), and is, in principle, able to cope with non-static anatomies such as the choroid plexus in the brain, or, more generally, heart or lung MRI.

11 Empirical analysis of noise distribution

To empirically characterize the noise distribution $\mathcal{D}_{\text{Noise}}$ (cf. Eq. (2)), we calculated residuals on the $N_{\text{Ex}} = 8$ T1w dataset (RS). To this end, each individual acquisition X_t was compared to the leave-one-out pseudo-clean average, $\bar{\mathbf{n}}_t = X_t - \frac{1}{N_{\text{Ex}}-1} \sum_{i=0, i \neq t}^{N_{\text{Ex}}} X_i$. After local z-normalization using smoothed σ maps to account for spatial variations of noise scales (*e.g.* due to pre-scan normalization), the residuals approximately followed a Gaussian distribution (Fig. 6). This is consistent with Rician magnitude noise approaching a Gaussian distribution at high SNR [21].

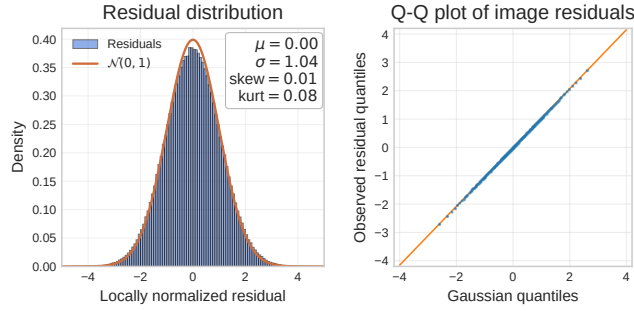


Fig. 6: Histogram and quantile-quantile (Q-Q) plot for a Gaussian fit to locally normalized noise residuals ($\bar{\mathbf{n}}_{t,j}$). The analysis was performed on the repeated ($N_{\text{Ex}}=8$) RS T1w dataset ($n = 2$).

We then assessed the spatial autocorrelation of the residual noise by evaluating the correlation between $\bar{\mathbf{n}}_{t,j}$ and $\bar{\mathbf{n}}_{t,j+l}$ for increasing spatial offsets l .

We found a spatial autocorrelation of noise (Fig. 7, left) with $r \approx 0.30$ for neighboring voxels and still $r \approx 0.06$ five voxels away, confirming that effective MRI noise is spatially correlated. We further analyzed pairwise difference maps between repeated acquisitions. While spatial correlation remained evident within individual difference maps, correlations between independent difference maps were scattered around zero (Fig. 7, right).

However, this analysis is inherently limited since the mandatory image registration requires resampling. Nevertheless, measurable correlations persist over several voxels, suggesting that interpolation alone does not fully explain the observed behavior.

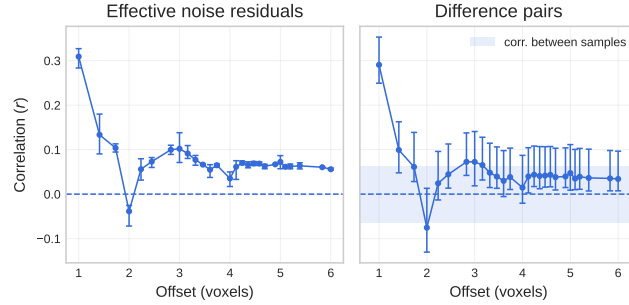


Fig. 7: Analysis of spatial auto-correlation of effective physical noise using residuals w.r.t. to pseudo-clean ($N_{\text{Ex}} = 7$, left) or pairwise-difference (right) in the RS $N_{\text{Ex}} = 8$ T1w dataset ($n = 2$).

Importantly, spatial correlations within an image violate assumptions underlying single-image training like N2V, whereas N2N training requires only independence between the noise realizations of input and reference images [44], which is approximately satisfied by independent physical re-acquisitions. Therefore, these findings support the preference for the proposed repeated acquisition learning over synthetic degradations for training on single-noisy data.

12 Ablations

12.1 Correlation concealment vs. information preservation tradeoff

Training objectives based on blind-spot masking [4, 37] or noising [53] require the selection of the appropriate severeness of the respective degradation. Here we show the effect of this hyperparameter for N2Void [37] (*i.e.* blind-spot size [8]) and Ner2N [53] (*i.e.* magnitude of added Rician noise), comparing not only against an independently acquired reference image, but also against the input image itself (Tab. 4 and Fig. 8). Note that the Identity denoiser constitutes the trivial solution for degradation-free training (blind-spot size 0 or Ner2N with $\sigma = 0$).

We observe that weak degradation (blind-spot size 1 / $\sigma = 1\%$) results in similar in- and outputs, *i.e.* insufficient noise removal or (partial) identity decay, while stronger degradations remedy noise replication as evidenced by decreasing similarity of denoised images and input images. Although stronger degradation improves the actual denoising effect (*i.e.* higher similarity to independent reference images), this effect reverses for strong degradations. We attribute this to increasing the domain-shift between strongly degraded training and real acquired images. In addition, large blind spots induce artifacts during denoising (Fig. 8). Thus, the degradation strength creates a trade-off between the constraining domain-shift vs. sufficient concealment of noise in the input image.

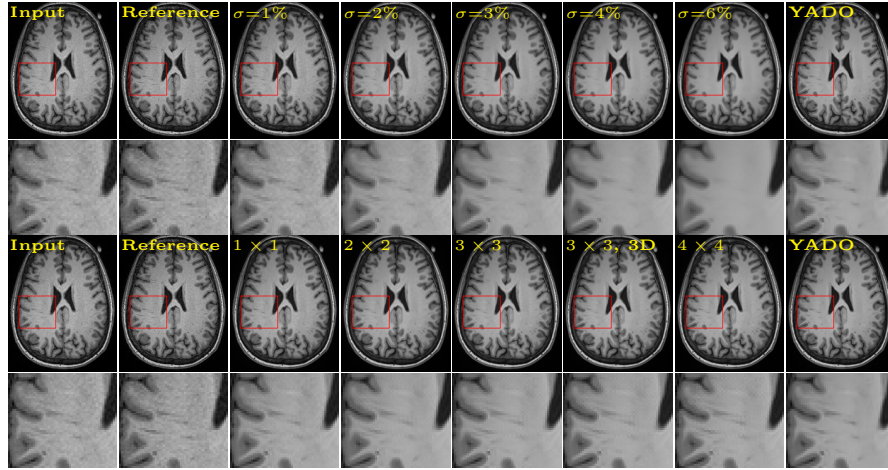


Fig. 8: Input noise concealment vs. information depletion trade-off for degradation-based N2N-derived methods. Top: Ner2N with varying amount of Rician noise, Bottom: (S-)N2Void with varying blind-spot sizes. (YADO: u - ReN2N)

Table 4: Effect of degradation severity on denoiser properties.

We compared the effect of blindspot size for (S-)N2Void [4, 8, 37]) and noise magnitude for Ner2N [53] in the OASIS-3 validation set ($n = 101$). We provide the similarity to either the input source images or independent reference target images.

method	degradation	SSIM		PSNR	
		input	reference	input	reference
uYADO (pN2N)	-	96.15 ± 1.04	90.93 ± 3.16	33.95 ± 1.44	29.66 ± 1.77
uYADO (ReN2N)	-	96.83 ± 0.50	91.00 ± 3.35	35.33 ± 0.96	29.60 ± 1.88
Identity	size=0 / $\sigma = 0\%$	100 ± 0	88.69 ± 3.89	100 ± 0	28.38 ± 1.81
(S-)N2Void [8, 37]	size=1 (1 $\times$ 1)	97.37 ± 0.36	90.27 ± 3.54	35.32 ± 0.96	29.18 ± 1.85
	size=4 (2 $\times$ 2)	95.94 ± 0.53	90.33 ± 3.44	33.78 ± 0.94	29.25 ± 1.81
	size=9 (3 $\times$ 3)	95.85 ± 0.56	90.35 ± 3.45	33.65 ± 0.93	29.21 ± 1.81
	size=11 (3 $\times$ 3, 3D)	95.20 ± 0.63	90.28 ± 3.38	33.04 ± 0.92	29.18 ± 1.77
	size=16 (4 $\times$ 4)	95.66 ± 0.59	90.28 ± 3.48	33.37 ± 0.95	29.19 ± 1.82
Ner2N [53]	$\sigma = 1\%$	99.77 ± 0.06	89.38 ± 3.73	45.55 ± 0.35	28.76 ± 1.83
	$\sigma = 2\%$	98.42 ± 0.38	90.13 ± 3.45	38.18 ± 0.26	29.18 ± 1.81
	$\sigma = 3\%$	97.13 ± 0.50	90.21 ± 3.28	35.59 ± 0.35	29.28 ± 1.74
	$\sigma = 4\%$	95.79 ± 0.59	89.83 ± 3.06	33.83 ± 0.48	29.20 ± 1.62
	$\sigma = 6\%$	92.47 ± 0.90	88.03 ± 2.67	31.12 ± 0.55	28.53 ± 1.28

12.2 Isolating the domain-shift

To isolate the domain-shift from noise-correlation problem, we trained ablated Ner2N ($\sigma = 2\%$) and S-N2Void (3×3 blindspot) using *physical* repeats as references. This guarantees the unpredictability of noise and, thus, corresponds to YADO’s pN2N training, only with the additional blind-spot masking or noising. We present the performance metrics for such models in Tab. 5.

Consistent with our previous observations, we find that for both Ner2N and S-N2V, physical decorrelation in a pN2N-like training improves denoising performance, narrowing the performance gap to plain, un-degraded pN2N. As pNer2N and pS-N2Void still underperformed pN2N, evidently degradation-induced domains-shifts (see also Sec. 13.2) still impaired denoiser training. This

Table 5: Ner2N and S-N2Void with physical repeats We assessed the effect of replacing the re-used degraded input image with an un-degraded re-acquisition, thus, physically de-correlating input and reference images during training. Results for physical (p) (S-)N2Void [8] (3×3 blindspot) and Ner2N [53] ($\sigma = 2\%$) are provided on the OASIS-3 validation set ($n = 101$).

method	SSIM	PSNR
uYADO (pN2N)	90.93 ± 3.16	29.66 ± 1.77
uYADO (ReN2N)	91.00 ± 3.35	29.60 ± 1.88
Ner2N	90.13 ± 3.45	29.18 ± 1.81
pNer2N	90.25 ± 2.90	29.24 ± 1.54
S-N2Void	90.35 ± 3.45	29.21 ± 1.81
pS-N2Void	90.73 ± 3.06	29.53 ± 1.70
Identity	88.69 ± 3.89	28.38 ± 1.81

highlights that both shortcomings of the degradation-based training – correlation remnants *and* domain-shifts – inhibit the training of well-performing denoisers.

12.3 ReN as augmentation

To account for ReN acting as data augmentation, effectively amplifying the training set size and giving indirect access to the RS dataset via the pre-trained DM, we supplemented S-N2V, Ner2N, and CNN3D with the ReN images. We present the results in Tab. 6.

We observed that ReN improves Ner2N (median performance gain of 0.20% SSIM), although to a lesser extend than for YADO (0.37%) and without competitive performance to ReN2N YADO. S-N2Void and CNN3D did not systematically profit from the ReN augmentation, which we attribute to the inferior training supervision.

Table 6: ReN augmentation for single-noisy baseline methods. Baseline methods were trained on the OASIS training data supplemented with the ReN images as augmentation, granting baseline the same access to RS-derived augmented images as ReN2N YADO training.

Method	OAS3	CHDI	Kirby	HCP-E	-YA	RS
uYADO (pN2N)	89.39 ±4.69	94.25 ±1.54	95.69 ±0.85	89.65 ±3.82	94.06 ±0.69	93.83 ±2.23
uYADO (ReN)	89.41 ±4.87	94.33 ±1.59	95.99 ±0.90	90.60 ±3.38	94.50 ±0.64	94.37 ±2.39
Ner2N	88.59 ±4.82	93.53 ±1.71	95.58 ±0.90	89.47 ±3.28	93.91 ±0.72	93.42 ±2.60
Ner2N (ReN)	88.79 ±4.79	93.71 ±1.66	95.63 ±0.86	89.80 ±3.31	94.13 ±0.73	93.63 ±2.55
S-N2V	88.69 ±5.05	93.67 ±1.74	95.59 ±1.01	88.43 ±3.73	91.38 ±0.70	92.62 ±2.51
S-N2V (ReN)	88.65 ±4.94	93.84 ±1.64	95.63 ±0.89	88.28 ±4.05	89.59 ±0.86	93.29 ±2.59
CNN3D	87.79 ±5.05	93.00 ±1.82	95.11 ±1.05	88.86 ±3.61	92.62 ±0.93	92.62 ±2.87
CNN3D (ReN)	87.84 ±5.01	92.99 ±1.81	95.09 ±1.02	88.85 ±3.60	92.71 ±0.87	92.66 ±2.83
Identity	86.91 ±5.65	92.66 ±2.09	95.03 ±1.21	87.09 ±4.46	89.50 ±1.04	91.60 ±3.25

12.4 Trivial re-noising

We further assessed whether simplifying the ReN augmentation to adding trivial Rician noise to a clean, regression- (*i.e.* single-step [63]) denoised image suffices,

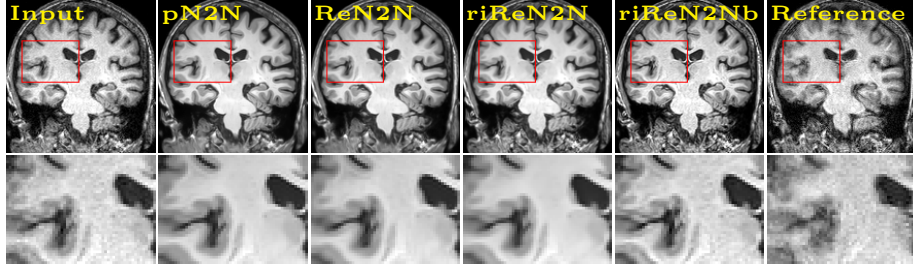


Fig. 9: Denoising results for a high-motion case in OASIS-3.

riReN2N: ReN2N training with explicit de- and trivial, Rician renoising.

riReN2Nb: riReN2N with spatial correlation due to a Gaussian blurr after noising.

i.e. whether using a direct estimate of $\hat{X}'_{\text{YADO}} \approx X'$ followed by explicitly adding $\hat{\mathbf{n}}_{\text{Rice}} \sim \mathcal{D}_{\text{Rice}}$ (cf. (6)) constitutes a computationally cheap alternative to DM-based ReN.

To assert a fair comparison between DM and regression paradigms [63], we trained a dedicated YADO (guided pN2N) denoiser equivalently to the ReN DM on the RS training set using this RS-YADO to denoise the OASIS-3 training set. We then trained with these clean images $\hat{X}'_{\text{YADO}}$, adding Rician noise ($\sigma = 3\%$) and predicting noisy re-acquisition. We compare the obtained Rician ReN (riReN2N) YADO against DM-based ReN2N and pN2N YADO versions in Tab. 7.

Although riReN2N resulted in a functional denoiser performing better than an Identity function, the performance was considerably worse than DM-based ReN2N and pN2N. To also simulate spatially correlated noise (Sec. 11), we additionally considered smoothing the riReN images with a Gaussian blur. While this blurred riReN2N YADO still did not match the DM-based ReN2N and pN2N versions, smoothing improved denoising notably, demonstrating that this simple smoothing operation already improves realism over plain white noise. Nevertheless, we observed that the blurred riReN2N YADO was particularly

Table 7: Trivial renoise (ReN) ablation

We simplified the compute-intensive DM-based ReN augmentation (ReN2N) to explicit de-noising (estimating $\hat{X}' \approx X'$ in Eq. (6)) followed by adding simple, Rician noise (riReN2N). In addition, we induced spatial noise correlation via smoothing the image with a Gaussian blur ($\sigma_{\text{blur}}=0.5$). Results are on the OASIS-3 validation set ($n=101$).

	SSIM	PSNR
pN2N	90.93 $\pm$ 3.16	29.66 $\pm$ 1.77
ReN2N	91.00 $\pm$ 3.35	29.60 $\pm$ 1.88
riReN2N ($\sigma = 3\%$)	89.52 $\pm$ 3.94	28.36 $\pm$ 1.84
riReN2N ($\sigma = 3\%$, blurred)	90.35 $\pm$ 3.13	29.35 $\pm$ 1.70
Identity	88.69 $\pm$ 3.89	28.38 $\pm$ 1.81

prone to high-motion and -noise images, where noise removal was insufficient (Fig. 9), *i.e.* the denoiser showed partial identity decay for low-quality inputs.

In general, this indicates that the power of ReN2N lies not only in implicit denoising (*i.e.* predicting $\hat{X}' \approx X'$ in (6)), but also in simulating realistic, complex noise ($\hat{\mathbf{n}}_{\text{DM}}$). This not only improves performance due to better realism than trivial re-noising, but also supports robustness to low-quality images.

12.5 Unguided Re-noising

Per default, we use cross-contrast conditioned DMs for ReN. Hence, ReN2N relies on the assumption that the training data provide co-acquired conditioning images. While this assumption is most often fulfilled (fnt. 1) and training sets can be selected accordingly, we still assessed ReN2N using an unguided DM to lift this constraint. To this end, we trained and sampled a unguided, T1w-only DM for unguided ReN2N (uReN2N), benchmarking the resulting ablated YADO version on the 1 mm T1w denoising test sets. We present the results in Tab. 8. While disregarding the T2w conditioning for the ReN augmentation resulted in slightly worse ranking of uReN2N compared to similarly T1w-only pN2N, uReN2N YADO still outperformed all equally single-contrast, single-noisy baselines. This demonstrates that even under the un-realistic requirement that training has to happen in non-repeated, single-contrast scenarios, YADO still constitutes a SOTA denoiser.

Note that, uReN2N generalizes better to unseen data sources (tie in Kirby) than

Table 8: Quantitative 1 mm T1w denoising results for training without T2w images (% SSIM; except PVS: abs. CNR [63], OASIS-3 trained). uReN2N used an ablated T1w-only DM for re-noising, lifting the cross-modality conditioning requirement. Formatting: **Bold**: best or statistically tied ($p \geq 5\%$, Wilcoxon vs. best, Bonferroni-corrected). **Gray background**: not significantly better than Identity. (Parentheses): guidance with anisotropic/low quality T2w.

Test dataset	OAS3	CHDI	Kirby	HCP-E	-YA	RS	PVS	rank
uYADO (pN2N)	89.39 ± 4.69	94.25 ± 1.54	95.69 ± 0.85	89.65 ± 3.82	94.06 ± 0.69	93.83 ± 2.23	2.63 ± 0.81	1.48
uYADO (uReN)	89.23 ± 5.02	94.24 ± 1.63	95.98 ± 0.93	90.47 ± 3.45	94.46 ± 0.65	94.30 ± 2.47	2.21 ± 0.40	1.97
Noisier2N	88.59 ± 4.82	93.53 ± 1.71	95.58 ± 0.90	89.47 ± 3.28	93.91 ± 0.72	93.42 ± 2.60	2.28 ± 0.83	3.59
Struc N2Void	88.69 ± 5.05	93.67 ± 1.74	95.59 ± 1.01	88.43 ± 3.73	91.38 ± 0.70	92.62 ± 2.51	2.17 ± 0.65	4.68
DIP	87.87 ± 5.34	93.27 ± 1.80	95.31 ± 1.16	89.04 ± 4.09	92.23 ± 1.04	92.88 ± 3.07	2.33 ± 0.60	4.99
BM4D	88.28 ± 4.52	93.40 ± 1.67	95.33 ± 0.90	89.37 ± 3.66	93.71 ± 0.73	92.93 ± 2.43	2.13 ± 0.83	4.89
Identity	86.91 ± 5.65	92.66 ± 2.09	95.03 ± 1.21	87.09 ± 4.46	89.50 ± 1.04	91.60 ± 3.25	2.33 ± 0.46	6.40

pN2N, *i.e.* retained this advantage from regular gReN2N. The only evaluation in which uReN2N underperformed pN2N training was PVS denoising. We hypothesize that T2w-agnostic re-noising hampers PVS generation, thus, providing improper training data for subsequent YADO training.

12.6 ReN2N to leverage scaling laws

In all denoising training settings, we constraint ReN to paired images matching the training corpus of pN2N as gold-standard training. However, inherently, ReN2N expands pN2N training to single-repeat corpora, thus, can potentially learn from more data and thereby exploit scaling laws.

To explore scaling for T2w denoising, we supplemented the 154 participants (308 images) for default, pN2N-matched training with 300 non-repeat pre-training participants and present the results in Tab. 9.

We observe that expanding the training set improves performance on the internal RS test set, however, at the cost of slightly reduced generalization. We hypothesize that further scaling of the unpaired ReN2N training set would likely close the performance gap to pN2N. In turn, without the requirements of repeated acquisitions, generalization can easily be improved by including more data sources for training.

Table 9: Quantitative T2w denoising results (% SSIM; except PVS CNR [63]). Formatting: **Bold**: best or statistically tied ($p \geq 5\%$, Wilcoxon vs. best, Bonferroni-corrected). All metrics were significantly better than Identity. ReN2N+unpaired: ReN2N-trained model with additional 300 unpaired images from the single-image dataset.

Test dataset n	RS 39	HCP-E 100	-YA 100	PVS 79	rank 318
uYADO (pN2N)	88.80 ± 10.44	90.21 ± 3.80	89.50 ± 1.48	2.69 ± 0.95	4.28
uYADO (ReN2N)	88.58 ± 10.43	91.10 ± 3.46	89.76 ± 1.51	2.59 ± 0.99	3.86
gYADO (pN2N)	89.65 ± 8.60	90.38 ± 3.89	89.24 ± 1.53	3.12 ± 0.87	2.99
gYADO (ReN2N)	88.98 ± 9.69	91.43 ± 3.39	90.28 ± 1.43	2.91 ± 0.85	1.86
gYADO (ReN2N+unpaired)	89.22 ± 9.28	91.16 ± 3.56	89.95 ± 1.54	3.04 ± 0.83	2.00

12.7 Effect of guidance on denoising

To further study the effect of guidance for denoising, we also trained T2w-guided (g) versions of the single-image denoisers Ner2N and S-N2V on the OASIS T1w dataset, presenting the results in Tab. 10. We observed that, akin to YADO, Ner2N profits in all test datasets except Kirby (low-resolution T2w). In CHDI and HCP-E/-YA gNer2N could thus match or slightly surpass pN2N-gYADO's performance, while gNer2N remained below ReN2N-gYADO in all tests. Guidance also improved S-N2Void, yet, with the exception of HCP-YA and the RS. Overall, these findings support the theoretical information enrichment due to denoiser guidance, independent of the training formulation and objective.

Table 10: Guided single-noisy training. Ablated versions with T2w as additional cross-modality guidance (g) were trained on the OASIS dataset. (Parentheses): guidance with anisotropic/low quality T2w.

Method	OASIS-3	CHDI	Kirby	HCP-E	-YA	RS
uYADO (pN2N)	89.39 ±4.69	94.25 ±1.54	95.69 ±0.85	89.65 ±3.82	94.06 ±0.69	93.83 ±2.23
gYADO (pN2N)	90.03 ± 4.35	94.11 ±1.42	(94.63) ±0.76	89.96 ±3.31	94.11 ±0.66	94.31 ±1.59
uYADO (ReN)	89.41 ±4.87	94.33 ±1.59	95.99 ± 0.90	90.60 ±3.38	94.50 ±0.64	94.37 ±2.39
gYADO (ReN)	89.77 ± 4.82	94.59 ± 1.47	(95.63) ±0.78	90.91 ± 3.24	94.59 ± 0.60	94.59 ± 2.25
Ner2N	88.59 ±4.82	93.53 ±1.71	95.58 ±0.90	89.47 ±3.28	93.91 ±0.72	93.42 ±2.60
gNer2N	89.07 ±4.87	94.19 ±1.53	(95.43) ±0.82	90.18 ±3.29	94.12 ±0.72	93.89 ±2.43
S-N2Void	88.69 ±5.05	93.67 ±1.74	95.59 ±1.01	88.43 ±3.73	91.38 ±0.70	92.62 ±2.51
gS-N2Void	88.89 ±5.14	93.79 ±1.72	(95.62) ±0.98	88.49 ±3.68	91.10 ±0.77	92.56 ±2.53
Identity	86.91 ±5.65	92.66 ±2.09	95.03 ±1.21	87.09 ±4.46	89.50 ±1.04	91.60 ±3.25

13 Model optimization and experiment details

13.1 YADO

Per default, we followed the hyperparameter choices previously determined for YODA [63], adding weight decay of 10^{-3} . Yet, when training solely with physical repeats (pN2N), we observed severe numerical instabilities as second-order normalization terms (v , Alg. 1 in [32]) of ADAM became very small ($v_{t \gg 1} < 10^{-20}$), which we attribute to strong training determinism. We stabilized the training, by setting the normalization-stabilizing $\varepsilon = 10^{-4}$ (Tab. 11). This reduces the maximum effective learning rate for parameters with small v_t and thus avoids chained gradient explosion in the case of individual, random, large gradients. We also attempted to remove weight decay and decrease the momentum term on v , β_2 , to allow for more rapid LR reduction in such cases, yet, this alone did not remedy training failure.

For training with DM-generated ReN images, we observed that using $N_{\text{Ex}} = 4$ rather than $N_{\text{Ex}} = 1$ synthetic images improves performance. Additionally, using ReN images as both input and reference, not only further improves the performance but also induces training entropy, avoiding overfitting as the target becomes unpredictable. This enables T1w training with default $\varepsilon = 10^{-8}$ (setting $\beta_2 = 0.9$ and omitting weight decay), which improves performance over training with higher $\varepsilon = 10^{-4}$. Furthermore, additionally providing ReN2N input along with the outputs, *i.e.* mixing the physical acquisition X_t with the synthetic repeats $\hat{X}_t^{(i)} \sim \text{DM}_\phi(X_t)$ generated thereof, further improve performance at no extra cost (Tab. 11).

Note that reducing ε caps the effective LR at a higher level, allowing better adapting the parameter-specific LR and, thus, exploiting the strength of second-order optimizers such as ADAM. While noising/blind-spot degradations similarly allowed for the default $\varepsilon = 10^{-8}$, this choice did not benefit the performance of any of the assessed N2N-derived baseline methods (Tab. 11).

For training, we sampled each image 10k times (*i.e.* each repeat subject/session a total of 20k times). To match training budgets between YADO and competing methods, we re-distributed the compute between pretraining and repeat fine-tuning for YADO. For HCP-A we doubled the training effort for all methods by sampling each image 20k time to account for non-repeated acquisitions.

During training, we randomly select a slice containing the brain (based on *FastSurfer* [22,23]) or the 10 adjacent voxels. Hereby, YADO ingests slabs of 5 slices and predicts the center slice of the reference image. We unify all slicing planes in a shared model by randomly selecting a plane during training. We predict all 3 slicing planes and average the results using the RMS. We concatenate guidance as the same multi-slice image as additional input channels.

For single-image pretraining we formulated $0.8\text{mm T2w} \leftrightarrow \text{T1w}$ (random inversion of the translation direction) or $\text{T1w, T2w} \rightarrow \text{FLAIR}$ translation for T1w/T2w or FLAIR denoising, respectively. We transferred the weights from translation pretraining to guided/unguided denoising by appending/replacing the respective input channels with randomly initialized weights.

Table 11: Ablation of (guided) YADO hyperparameters and training scenarios on the OASIS-3 validation set ($n = 101$).

We considered ADAM’s ε (setting $\beta_2 = 0.9$ when $\varepsilon = 10^{-8}$) and physical vs. DM-based synthetic repeats (ReN) as inputs and reference images. Per default, we used $N_{\text{Ex}} = 4$ DM-generated images for ReN2N training, which was ablated to $N_{\text{Ex}} = 1$, additionally combining with physical repeats (ReN + phys.). Note that ablations partially used RS-pretrained (pN2N) denoisers to reduce compute.

Bold: best performing, Underline: selected setting for pN2N or ReN2N.

input	reference	pretrain	ADAM ε	SSIM	PSNR
physical	physical	n	10^{-4}	<u>91.408 $\pm$ 2.786</u>	<u>29.883 $\pm$ 1.706</u>
physical	physical	n	10^{-8}	fail	
ReN	physical	y	10^{-4}	90.867 $\pm$ 3.489	29.157 $\pm$ 1.865
ReN	physical	y	10^{-4}	90.884 $\pm$ 3.529	29.176 $\pm$ 1.878
ReN	ReN	y	10^{-4}	<u>91.083 $\pm$ 3.275</u>	29.616 $\pm$ 1.858
ReN	ReN	n	10^{-4}	91.034 $\pm$ 3.287	29.583 $\pm$ 1.855
ReN	ReN	n	10^{-8}	<u>91.301 $\pm$ 3.264</u>	29.752 $\pm$ 1.883
ReN + phys.	ReN + phys.	n	10^{-8}	<u>91.313 $\pm$ 3.247</u>	<u>29.759 $\pm$ 1.877</u>

13.2 Renoise (ReN) DM

The re-noising DM used the same backbone as YADO with an additional time-step embedding following [24]. Similarly to YADO, the model was pretrained for T2w $\rightarrow$ T1w translation (0.8 mm) on the RS pretraining dataset and subsequently fine-tuned on the RS T1w-repeat training set (0.8 mm or 1 mm). We train the renoising DM with ADAM (LR= 10^{-4} , no weight decay, $\varepsilon = 10^{-6}$) with an L_2 loss and the same brain weighting scheme as YADO. We use a 2.5D inference scheme as previously described [63]. The DM employs a v -prediction target. FLAIR ReN DMs were fine-tuned on repeated data from 1 mm translation pretraining [63]. To spare compute when sampling multiple images for the re-noising augmentation, we share the late ($T > 250$) time steps and branch the sampling only for the early, low-SNR timesteps ($T \leq 250$).

As we assumed image realism to be key for successful re-noising, we applied a cosine² (cos) [55] noise schedule, which shifts training and inference to higher SNRs and, thus, results in stronger noise imitation [63]. To further enhance this effect, we modified the noise schedule function to $\sqrt{\cos}$ rather than $\cos^2$. This further improves visual realism (Tab. 12, upper), measured as FID and paired cosine similarity (Cos. Sim.) of the extracted features as described in [63].

To further investigate the advantage of re-noising over degradation (blindspot masking or noising), we also compared the visual realism of ReN training images to S-N2Void and Ner2N. We observed that ReN images more closely resembled gold-standard (noisy) physical acquisition (Tab. 12, lower), further underlining enhanced visual realism of YADO’s ReN2N training over existing single-image training schemes and, thus, reduction of domain-shift. Note that reduced SSIM/PSNR can be attributed to increased noise replication and, therefore, does not indicate reduced translation accuracy [63].

Table 12: Visual realism of input images for single-noisy training (proposed ReN2N, Ner2N, S-N2V) on a subset ($n = 30$) of the external Oasis-3 set. Values for pN2N are provided as a reference.

method	noise schedule	SSIM (% $\uparrow$)	PSNR (dB $\uparrow$)	FID ($\cdot 10^2$, $\downarrow$)	Cos. Sim. (% $\uparrow$)
ReN2N	$\sqrt{\cos}$	89.63 ± 2.80	28.46 ± 1.11	3.90	98.70 ± 0.55
	$\cos^2$	89.72 ± 2.69	28.55 ± 1.09	4.23	98.67 ± 0.49
Ner2N	-	72.16 ± 4.53	22.96 ± 1.58	59.8	92.18 ± 1.16
S-N2Void	-	83.84 ± 3.16	25.55 ± 0.89	9.34	97.08 ± 0.52
pN2N	-	89.50 ± 3.39	28.53 ± 1.29	2.94	98.86 ± 0.81

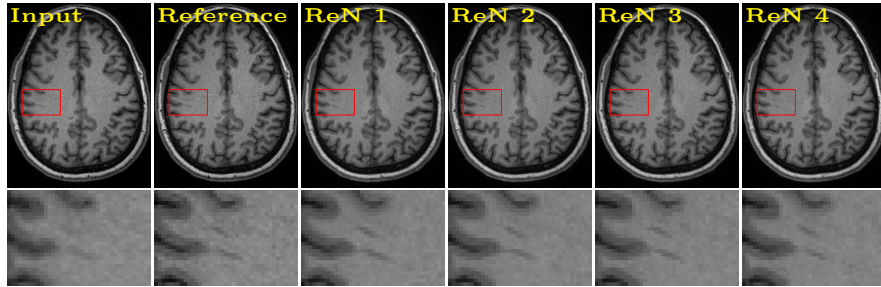


Fig. 10: Exemplary result for sampling a fixed DM-based renoising (ReN) model (T2w-guided).

13.3 Reference methods

We equally optimized the principled N2N-derived methods (Ner2N, N2Score, (S-)N2Void, Neigh2Neigh) with respect to the degradation strength (Ner2N, N2Score: σ of added noise, (S-)N2V: blind-spot size) and the same optimizer hyperparameters as considered for ablating YADO (Tab. 13). The remaining hyperparameters (network topology and size, L_2 loss, batch size, EMA β , etc.) matched YADO. We found that the same optimizer setting as for pN2N YADO is optimal, except for N2Score. Here, we selected the higher LR of $\eta = 10^{-3}$, however, this was not stable for training on HCP-A so we resorted to YADO-like settings, choosing $\eta = 10^{-4}$. To efficiently optimize the image-specific amount of noise removal σ^* during N2Score inference, we minimize the total variation [31] via an iterative bracketing search over the noise-scaling parameter.

We trained RED-WGAN [62] and the derived regression CNN (CNN3D) with Rician noise ($\sigma = 3\%$) based on Ner2N results. We re-implemented BME-X in PyTorch (3D Res-UNet [61], 3 layers with [64, 128, 256] channels), using the open-source tool FastSurfer [22, 23] (v2.2, *seg-only*) as segmentation prior, reducing tissue labels to proposed $c = 4$ classes, and included the Cerebellum and omitted skull-stripping for a fair comparison. We simulated motion as suggested by Sun et al. [71], applied a volume-wise phase shift as $\sigma \sim [1, 2]$. For inference, we used 25 % patch overlap with Gaussian smoothing.

For ZSDs (DIP [75] and Pixel2Pixel [47]), we leverage the volumetric nature of

Table 13: Ablation Results for principled Noise2Noise (N2N)-derived baseline methods [8, 26, 31, 37, 53], over degradation strength (Rician noise σ or blind-spot size), AdamW optimizer settings (learning rate η , weight decay λ , ε for num. stability; for $\varepsilon > 10^{-8}$, we also set $\beta_2 = 0.9$ from default 0.999).

Bold: best performing setting for each methods, underline: selected settings.

method	setting	η	λ	ε	SSIM	PSNR
N2Void	-	10^{-4}	10^{-3}	10^{-4}	<u>90.273 ± 3.542</u>	<u>29.177 ± 1.848</u>
N2Void	-	10^{-4}	10^{-3}	10^{-8}	89.727 ± 3.759	28.923 ± 1.872
N2Void	-	10^{-3}	10^{-3}	10^{-4}	90.200 ± 3.639	29.150 ± 1.873
N2Void	-	10^{-5}	10^{-3}	10^{-4}	90.192 ± 3.521	29.139 ± 1.831
N2Void	-	10^{-4}	0	10^{-4}	90.268 ± 3.554	29.174 ± 1.853
N2Void	-	10^{-4}	10^{-1}	10^{-4}	90.144 ± 3.644	29.080 ± 1.876
S-N2Void	2×2 mask	10^{-4}	10^{-3}	10^{-4}	90.334 ± 3.437	<u>29.250 ± 1.810</u>
S-N2Void	3×3 mask	10^{-4}	10^{-3}	10^{-4}	<u>90.345 ± 3.449</u>	29.212 ± 1.812
S-N2Void	3×3 mask, 3D	10^{-4}	10^{-3}	10^{-4}	90.278 ± 3.381	29.179 ± 1.774
S-N2Void	4×4 mask	10^{-4}	10^{-3}	10^{-4}	90.275 ± 3.481	29.194 ± 1.819
Neigh2Neigh	-	10^{-4}	10^{-3}	10^{-4}	<u>89.395 ± 2.900</u>	28.650 ± 1.463
Neigh2Neigh	-	10^{-4}	10^{-3}	10^{-8}	89.263 ± 2.884	28.514 ± 1.439
Neigh2Neigh	-	10^{-3}	10^{-3}	10^{-4}	89.131 ± 2.840	28.541 ± 1.417
Neigh2Neigh	-	10^{-5}	10^{-3}	10^{-4}	89.335 ± 2.872	<u>28.663 ± 1.467</u>
Neigh2Neigh	-	10^{-4}	0	10^{-4}	88.703 ± 3.895	28.427 ± 1.815
Neigh2Neigh	-	10^{-4}	10^{-1}	10^{-4}	89.227 ± 2.856	28.557 ± 1.447
Ner2N	$\sigma \sim \mathbf{U}[2\%, 4\%]$	10^{-4}	10^{-3}	10^{-4}	89.102 ± 3.840	28.633 ± 1.827
Ner2N	$\sigma = 2\%$	10^{-4}	10^{-3}	10^{-4}	90.128 ± 3.448	29.178 ± 1.814
Ner2N	$\sigma = 3\%$	10^{-4}	10^{-3}	10^{-4}	<u>90.207 ± 3.282</u>	<u>29.282 ± 1.744</u>
Ner2N	$\sigma = 4\%$	10^{-4}	10^{-3}	10^{-4}	89.828 ± 3.062	29.195 ± 1.621
Ner2N	$\sigma = 6\%$	10^{-4}	10^{-3}	10^{-4}	88.033 ± 2.669	28.530 ± 1.283
Ner2N	$\sigma = 3\%$	10^{-3}	10^{-3}	10^{-4}	88.723 ± 3.879	28.463 ± 1.809
Ner2N	$\sigma = 3\%$	10^{-5}	10^{-3}	10^{-4}	88.658 ± 3.902	28.379 ± 1.807
Ner2N	$\sigma = 3\%$	10^{-4}	0	10^{-4}	90.137 ± 3.514	29.170 ± 1.849
Ner2N	$\sigma = 3\%$	10^{-4}	10^{-1}	10^{-4}	89.755 ± 3.483	28.995 ± 1.775
Ner2N	$\sigma = 3\%$	10^{-4}	10^{-3}	10^{-8}	88.612 ± 3.915	28.389 ± 1.811
N2Score	$\sigma \sim \mathbf{U}[2\%, 4\%]$	10^{-4}	10^{-3}	10^{-4}	89.391 ± 3.766	28.732 ± 1.857
N2Score	$\sigma = 2\%$	10^{-4}	10^{-3}	10^{-4}	89.938 ± 3.694	28.991 ± 1.912
N2Score	$\sigma = 3\%$	10^{-4}	10^{-3}	10^{-4}	90.485 ± 3.584	29.272 ± 1.940
N2Score	$\sigma = 4\%$	10^{-4}	10^{-3}	10^{-4}	<u>90.662 ± 3.481</u>	29.413 ± 1.900
N2Score	$\sigma = 6\%$	10^{-4}	10^{-3}	10^{-4}	90.245 ± 3.264	29.348 ± 1.752
N2Score	$\sigma = 4\%$	10^{-3}	10^{-3}	10^{-4}	<u>90.716 ± 3.472</u>	<u>29.448 ± 1.916</u>
N2Score	$\sigma = 4\%$	10^{-5}	10^{-3}	10^{-4}	90.649 ± 3.465	29.419 ± 1.893
N2Score	$\sigma = 4\%$	10^{-4}	0	10^{-4}	90.705 ± 3.464	29.438 ± 1.903
N2Score	$\sigma = 4\%$	10^{-4}	10^{-1}	10^{-4}	90.645 ± 3.488	29.400 ± 1.897
N2Score	$\sigma = 4\%$	10^{-3}	0	10^{-4}	90.677 ± 3.471	29.422 ± 1.899
N2Score	$\sigma = 4\%$	10^{-4}	10^{-3}	10^{-8}	90.439 ± 3.630	29.242 ± 1.957

MRI and fit the models on all contained axial slices, combating overfitting noise instances, and therefore improved performance over fitting individual slice-wise models. For DIP input, a noised version of the input slice (3% Rician noise, selected from the same choices as default Ner2N training) to guide generation and trained for 2000 epochs.

We implemented N2Contrast [78] using the provided implementation of the trainable bilateral filters [58] and custom a 3D U-Net [61] with 16 input channels as reported [78], as no code was released (full U-Net [61] configuration: 3 layers with [16, 32, 64] channels). We did not consider R2D2+ [12] and Neighboring Slice N2N [85] due to the unavailability of publicly-shared code and implementation details.

13.4 Image pre-processing and testing

Image registration for T1w/T2w denoising was performed in FreeSurfer (v7.4.1) [16] using ‘mri_robust_register’ [64] (12 DOFs), using SynthStrip brain-masking [25] for intra-modality and ‘bbregister’ [20] (6 DOFs, based on FastSurfer v2.2 [22, 23]) for cross-modality registration. For FLAIR registration and the large pretraining set of the RS with >2.5k images, we omitted costly surface registration and used ‘mri_coreg’ with (NMI loss) instead. Resampling was performed using cubic interpolation.

For training and SSIM/PSNR testing, we resampled the input and guidance conditioning into the native space of the target image to avoid blurring the reference due to interpolation. For $N_{\text{Ex}} > 2$ we resampled in the joint template space [64]. HPC-Epilepsy was resampled from the provided $0.5 \times 0.5 \times 0.8$ mm resolution back to the native 0.8 mm isotropic acquisition resolution. For HCP-A, we used the provided, bias-field corrected images, whereas we performed a custom bias-field correction for HCP-YA with n4 [74]. For the remaining images we used the image as provided, *i.e.* typically with vendor-provided pre-scan normalization. For all training and test images, we performed a robust intensity normalization following the FreeSurfer preprocessing protocol [22].

For testing, we cropped to a brain bounding-box dilated by 10 voxels in each direction. To constrain SSIM and PSNR to only brain tissue, we only evaluated voxels within the dilated FastSurfer (*seg-only*) brain mask. *SHIVA*-WMH was run after bias-field removal with n4 [74]. As BME-X (as-is pipeline) entails a custom skull-stripping for preprocessing which removes *e.g.* the Cerebellum [71], we report SSIM/PSNR only within BME-X’s brain mask. While we consider not translating defined brain parts a inherent shortcoming, the performance was anyways not competitive.

13.5 High-resolution FLAIR acquisition

The high-resolution (0.8 mm) RS FLAIR images were acquired as a custom sequence (TR/TI/TE=5s/1.8s/0.382s, PI factor 4, FOV = 256 x 256 x 179 mm³ (sagittal orientation), TA=6:47 min) in the free-protocol of the RS [34]. The remaining MRI parameters are described in the respective references.

14 Additional results

14.1 PSNR

Table 14: Quantitative T1w denoising results (dB PSNR).

Formatting: **Bold:** best or statistically tied ($p \geq 5\%$, Wilcoxon vs. best, Bonferroni-corrected). **Gray background:** not significantly better than Identity. (Parentheses): guidance with anisotropic/low quality T2w. [†]: synthesis and evaluation at 1 mm.

	OASIS-3 trained (ca. 1.0 mm)						HCP-A trained (0.8 mm) rank			
Test dataset	OAS3	CHDI	Kirby	HCP-E	-YA	RS	HCP-E	-YA	RS	all
Resolution (mm)	1.0	1.0	1.0	0.8 [†]	0.7 [†]	0.8 [†]	0.8	0.7	0.8	0.7-1
Vendor	S	S/P	P	GE	S	S	GE	S	S	-
n	169	100	21	100	100	33	100	100	33	423
uYADO (pN2N)	28.97 ±2.20	30.32 ±3.16	29.83 ±1.58	30.10 ±1.91	29.50 ±2.09	30.12 ±2.00	28.53 ±2.03	28.32 ±1.40	28.65 ±1.73	8.79
uYADO (ReN2N)	28.92 ±2.31	30.04 ±3.48	29.70 ±1.88	30.57 ±2.03	30.93 ±1.86	31.60 ±2.06	29.98 ±2.02	29.42 ±1.44	30.75 ±1.84	3.02
gYADO (pN2N)	29.27 ±2.14	30.01 ±3.41	29.85 ±1.05	30.04 ±1.86	29.48 ±2.07	30.25 ±2.01	28.25 ±2.08	29.02 ±1.39	27.55 ±1.27	8.45
gYADO (ReN2N)	29.09 ±2.32	30.18 ±3.55	29.48 ±1.81	30.63 ±2.02	30.91 ±1.85	31.73 ±2.06	30.08 ±2.09	29.37 ±1.45	30.94 ±1.76	2.28
Ner2N [53]	28.61 ±2.17	29.66 ±3.10	29.62 ±1.77	30.22 ±1.86	30.55 ±2.02	30.85 ±1.99	29.16 ±1.98	28.47 ±1.35	29.46 ±1.86	7.44
N2Score [31]	28.74 ±2.36	29.77 ±3.40	29.24 ±1.67	29.71 ±1.77	30.42 ±1.74	30.64 ±1.84	28.33 ±5.78	24.98 ±8.27	30.26 ±1.82	9.08
N2Void [4, 37]	28.50 ±2.30	29.70 ±3.38	29.48 ±1.79	29.70 ±1.83	29.36 ±1.55	30.42 ±1.85	29.87 ±1.98	29.28 ±1.46	30.13 ±1.94	8.44
S-N2Void [8]	28.55 ±2.26	29.49 ±3.44	29.17 ±1.83	29.80 ±1.81	29.64 ±1.58	30.47 ±1.80	30.03 ±1.98	29.59 ±1.46	30.30 ±1.92	6.56
Neigh2Neigh [26]	28.09 ±1.84	29.24 ±3.06	28.62 ±1.43	29.75 ±1.64	30.06 ±1.84	29.89 ±1.65	29.86 ±1.78	29.29 ±1.36	29.61 ±1.56	10.35
RED-WGAN [62]	27.39 ±2.00	28.48 ±2.72	28.27 ±1.32	29.18 ±1.72	28.70 ±2.08	28.84 ±2.00	28.84 ±1.79	28.66 ±1.18	28.72 ±1.59	15.42
CNN3D [62]	28.14 ±2.15	29.34 ±3.16	29.19 ±1.63	29.88 ±1.88	30.13 ±1.77	30.44 ±1.94	29.63 ±1.93	29.11 ±1.40	29.73 ±1.79	8.75
DDM ² [83]	17.38 ±1.10	16.60 ±1.30	15.17 ±1.01	17.59 ±0.71	14.93 ±1.42	16.44 ±1.40	15.82 ±0.88	14.60 ±0.72	15.71 ±0.50	20.87
DDM ² Reg. [63, 83]	26.93 ±1.60	26.61 ±2.43	26.84 ±1.11	27.07 ±1.54	26.73 ±1.61	27.34 ±1.78	25.24 ±1.05	25.06 ±0.66	26.40 ±0.75	18.16
BME-X [71]	25.69 ±1.07	25.65 ±1.52	23.93 ±1.07	24.60 ±1.80	23.13 ±1.19	24.24 ±1.07	25.31 ±1.27	24.66 ±1.33	25.96 ±1.29	19.04
N2Contrast [78]	27.94 ±2.30	29.34 ±3.28	29.27 ±1.77	29.57 ±2.00	29.04 ±1.58	30.14 ±2.01	29.40 ±2.03	28.67 ±1.44	29.59 ±1.90	12.35
YODA [63]	26.95 ±2.35	23.17 ±2.32	20.35 ±1.00	20.59 ±3.47	20.23 ±2.37	22.60 ±3.75	13.42 ±2.63	18.57 ±2.56	18.58 ±2.76	19.80
DIP [75]	28.09 ±2.20	29.21 ±3.22	28.94 ±1.86	29.81 ±1.93	29.99 ±1.70	30.46 ±1.99	29.59 ±2.02	29.06 ±1.55	29.73 ±1.90	10.30
Pixel2Pixel [47]	27.27 ±1.50	28.22 ±2.78	27.48 ±1.11	29.12 ±1.48	29.13 ±1.64	28.50 ±1.37	29.53 ±1.70	28.75 ±1.32	28.75 ±1.42	14.76
ANTs NLM [49]	28.25 ±1.96	29.24 ±3.19	28.87 ±1.62	29.99 ±1.81	30.50 ±1.79	30.40 ±1.78	30.00 ±1.93	29.44 ±1.45	29.99 ±1.73	7.61
BM4D [48]	28.36 ±2.03	29.45 ±3.28	29.14 ±1.67	30.11 ±1.85	30.57 ±1.81	30.60 ±1.82	30.13 ±1.97	29.60 ±1.47	30.24 ±1.82	5.12
BME-X (as-is) [71]	9.18 ±1.14	9.16 ±1.38	10.12 ±0.82	9.40 ±1.26	11.62 ±1.81	9.91 ±0.80	10.10 ±2.74	11.49 ±2.15	11.11 ±1.50	22.00
Identity	27.74 ±2.24	29.06 ±3.25	28.98 ±1.77	29.30 ±1.95	28.91 ±1.48	30.00 ±1.95	29.15 ±1.97	28.40 ±1.40	29.46 ±1.86	14.41

Table 15: Quantitative T2w/FLAIR denoising results (dB PSNR).
 Formatting: **Bold**: best or statistically tied ($p \geq 5\%$, Wilcoxon vs. best, Bonferroni-corrected). Gray background: not significantly better than Identity. (Parentheses): synthesis/guidance with anisotropic/low quality T2w. [†]: synthesis and evaluation at 1 mm.

	T2w			FLAIR			rank
Test dataset	RS	HCP-E	-YA	RS	still	nodd	all
Resolution (mm)	0.8	0.8	0.7	0.8	1.0 [†]	1.0 [†]	0.8-1
Vendor	S	GE	S	S	S	S	-
n	39	100	100	50	10	10	304
uYADO (pN2N)	28.20 ±3.29	25.83 ±2.76	26.25 ±0.94	28.97 ±2.53	25.40 ± 0.79	21.80 ± 1.91	6.29
uYADO (ReN2N)	28.05 ±3.30	27.88 ± 2.60	27.88 ±1.17	28.64 ±2.62	25.26 ± 0.80	21.94 ± 1.82	2.84
gYADO (pN2N)	28.39 ± 3.08	26.39 ±2.56	26.26 ±1.00	29.40 ± 2.34	17.07 ±0.98	15.92 ±1.42	6.31
gYADO (ReN2N)	28.12 ±3.27	27.95 ± 2.55	28.24 ± 1.30	29.10 ±2.43	13.63 ±1.07	13.20 ±1.14	2.93
Ner2N [53]	27.49 ±3.44	27.61 ±2.56	27.71 ±1.37	27.85 ±2.56	24.53 ±0.77	21.16 ±1.69	6.35
S-N2Void [8]	27.65 ±3.40	27.69 ±2.46	27.76 ±1.27	28.08 ±2.62	24.64 ±0.75	21.44 ±1.70	4.86
DIP [75]	27.53 ±3.34	27.66 ±2.48	27.89 ±1.36	28.09 ±2.56	24.68 ±0.77	21.39 ±1.64	4.59
ANTs NLM [49]	26.53 ±2.79	27.00 ±2.01	27.26 ±1.09	28.14 ±2.47	24.79 ±0.66	21.84 ± 1.72	7.24
BM4D [48]	27.14 ±3.08	27.42 ±2.22	27.70 ±1.21	28.36 ±2.63	24.89 ±0.89	21.85 ± 1.78	5.99
Identity	27.47 ±3.44	27.58 ±2.56	27.68 ±1.38	27.77 ±2.54	24.41 ±0.77	21.15 ±1.67	7.60

14.2 Additional examples

Used Abbreviations:

g-/u- pN2N/ReN2N - guided/unguided YADO trained with pN2N/ReN2N

Ner2N - Noisier2N [53]

N2Sc - Noise2Score [31]

S-(N2V) - (Structure-)Noise2Void [4, 8, 37]

Neighb2N - Neighbor2Neighbor [26]

WGAN - RED-WGAN [62]

CNN3D - CNN3D (regression-only RED-WGAN) [62]

DDM2 (R) - (regression-sampled [63]) DDM² [83]

BME-X - BME-X [71] (re-implemented)

N2Contr - Noise2Contrast [78]

YODA - YODA [63]

DIP - Deep Image Prior [75]

P2P - Pixel2Pixel [47]

stBME-X - static as-is BME-X [71]

BM4D - BM4D [48]

ANTs - ANTs NLM [49]

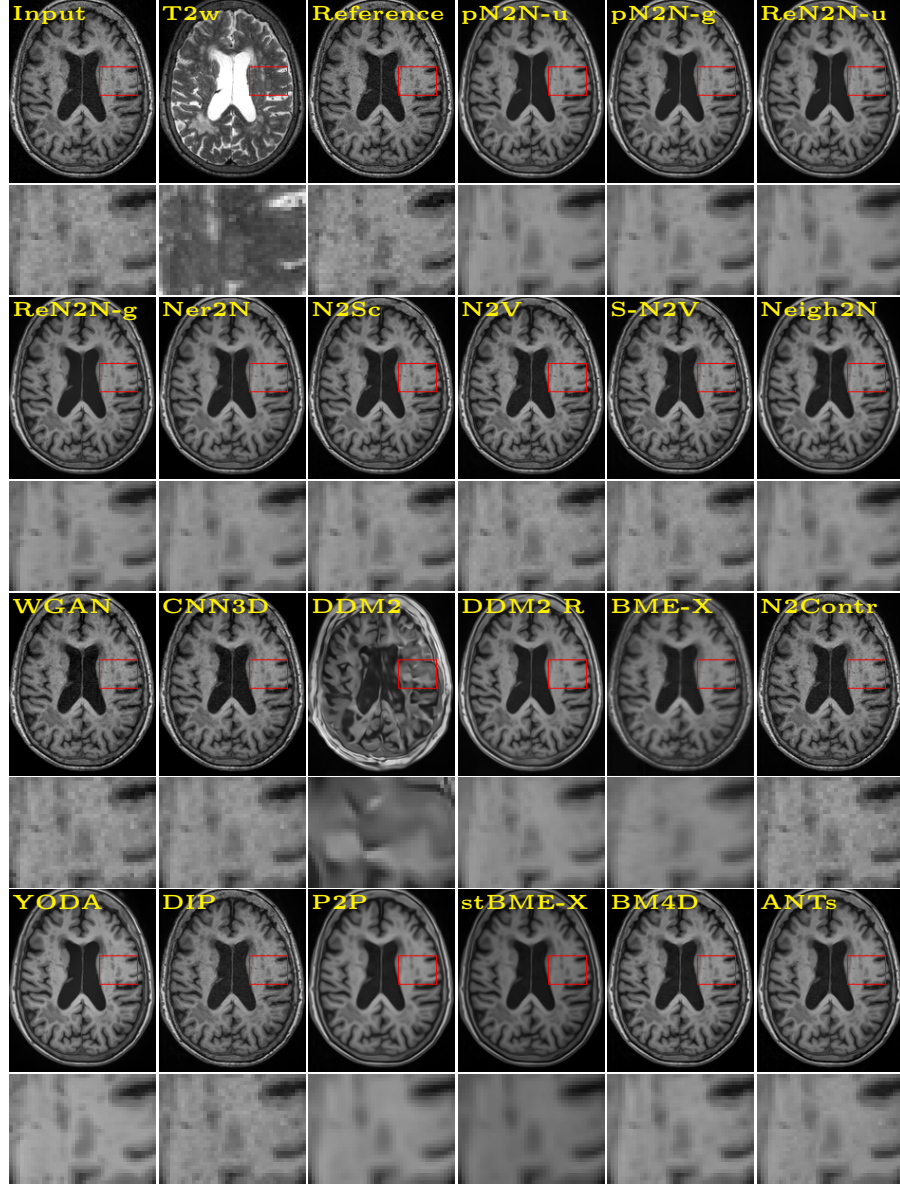


Fig. 11: T1w (1 mm) denoising performance in the OASIS test set.

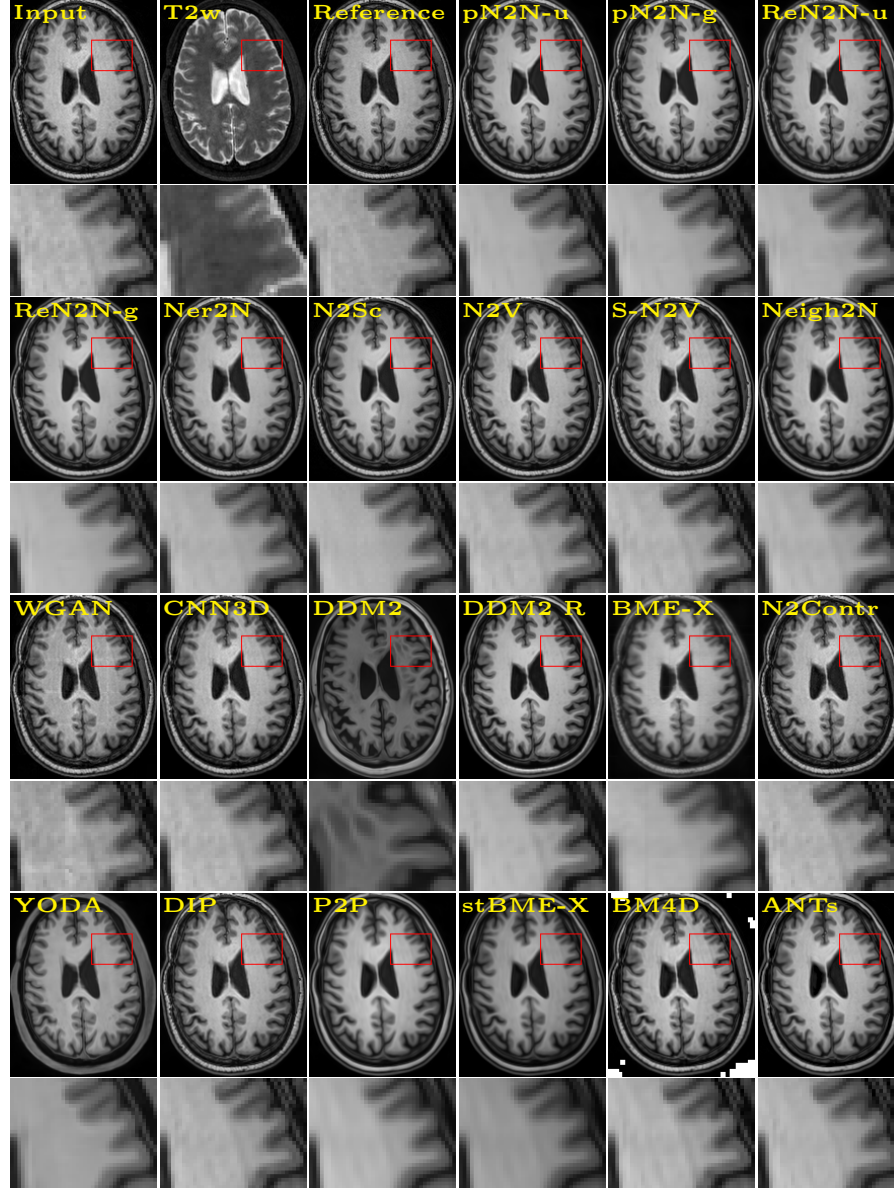


Fig. 12: T1w (1 mm) denoising performance in the CHDI test set.

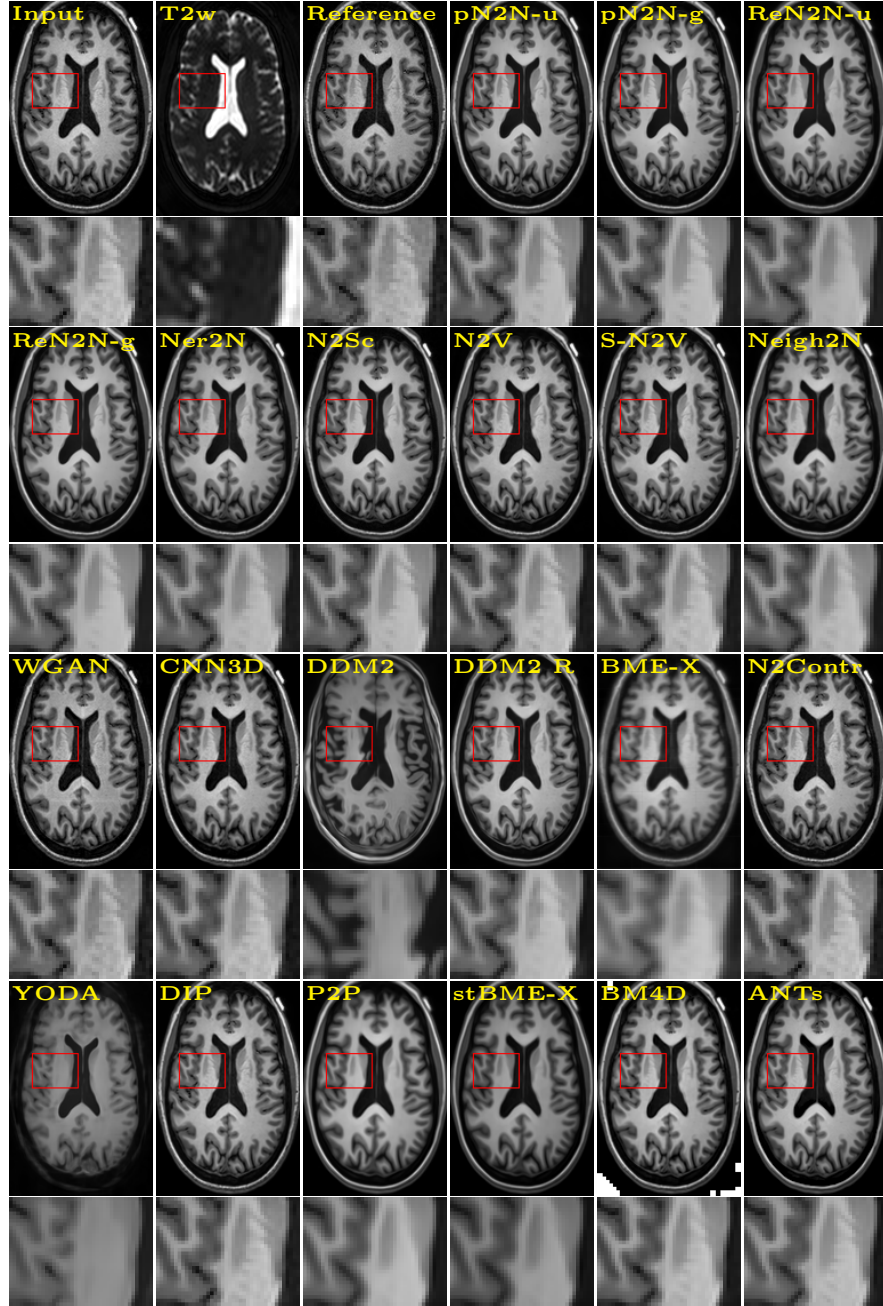


Fig. 13: T1w (1 mm) denoising performance in the Kirby test set.

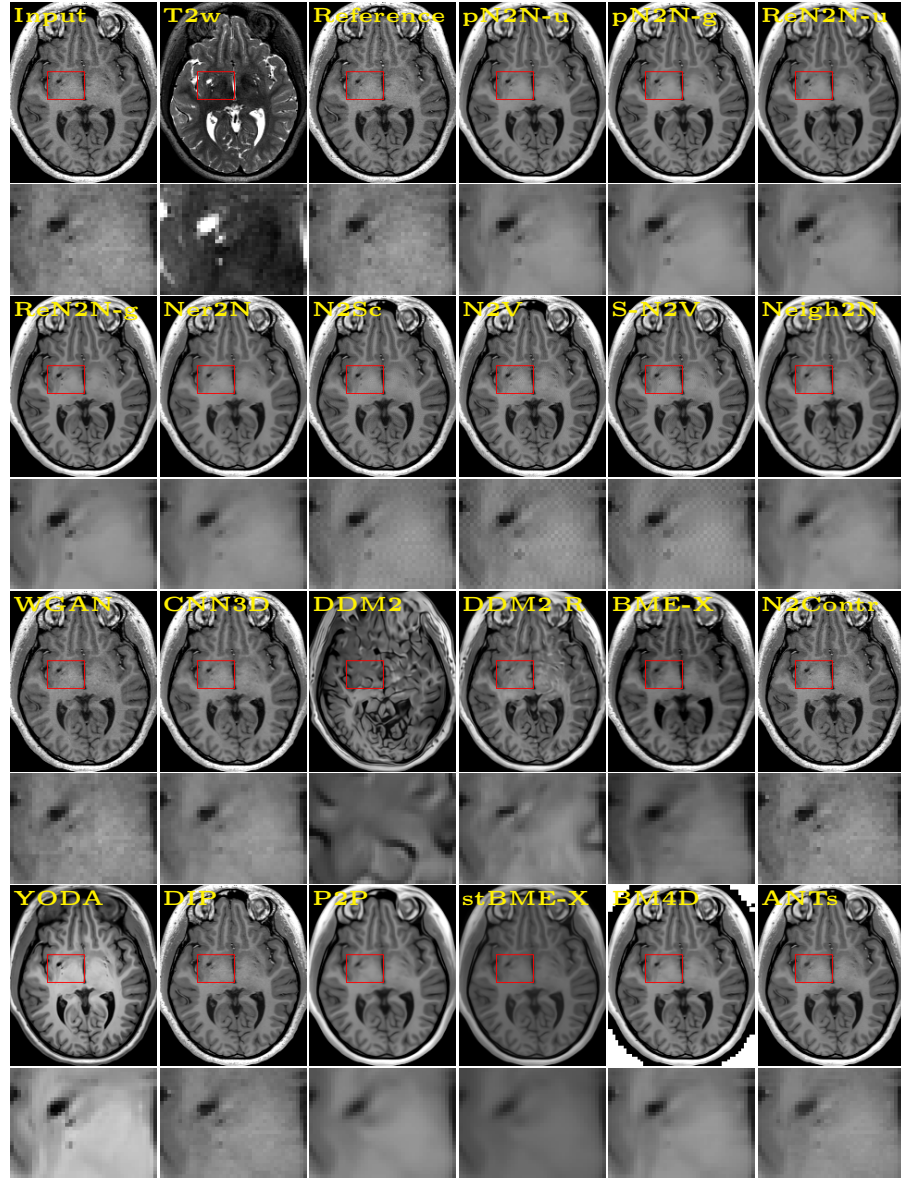


Fig. 14: T1w (1 mm) denoising performance in the HCP-E test set.

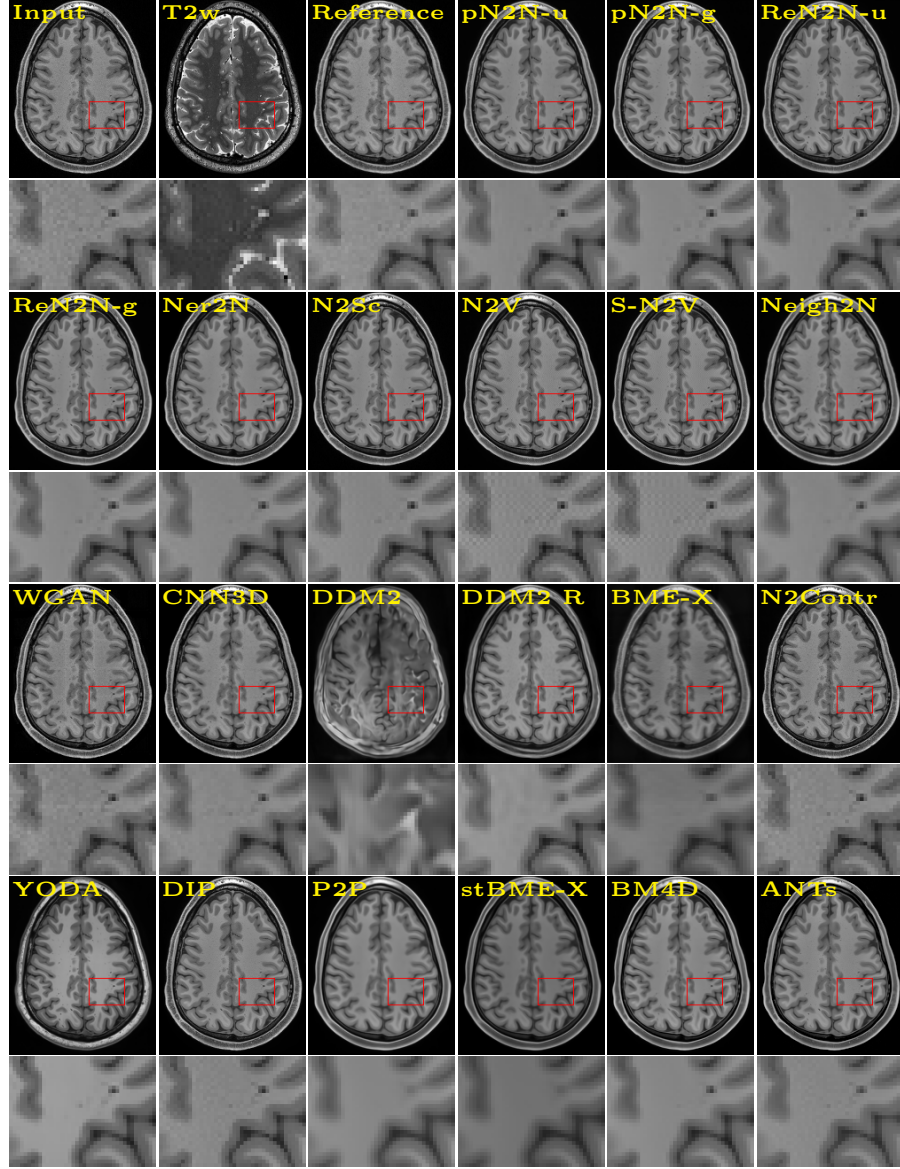


Fig. 15: T1w (1mm) denoising performance in the HCP-YA test set.

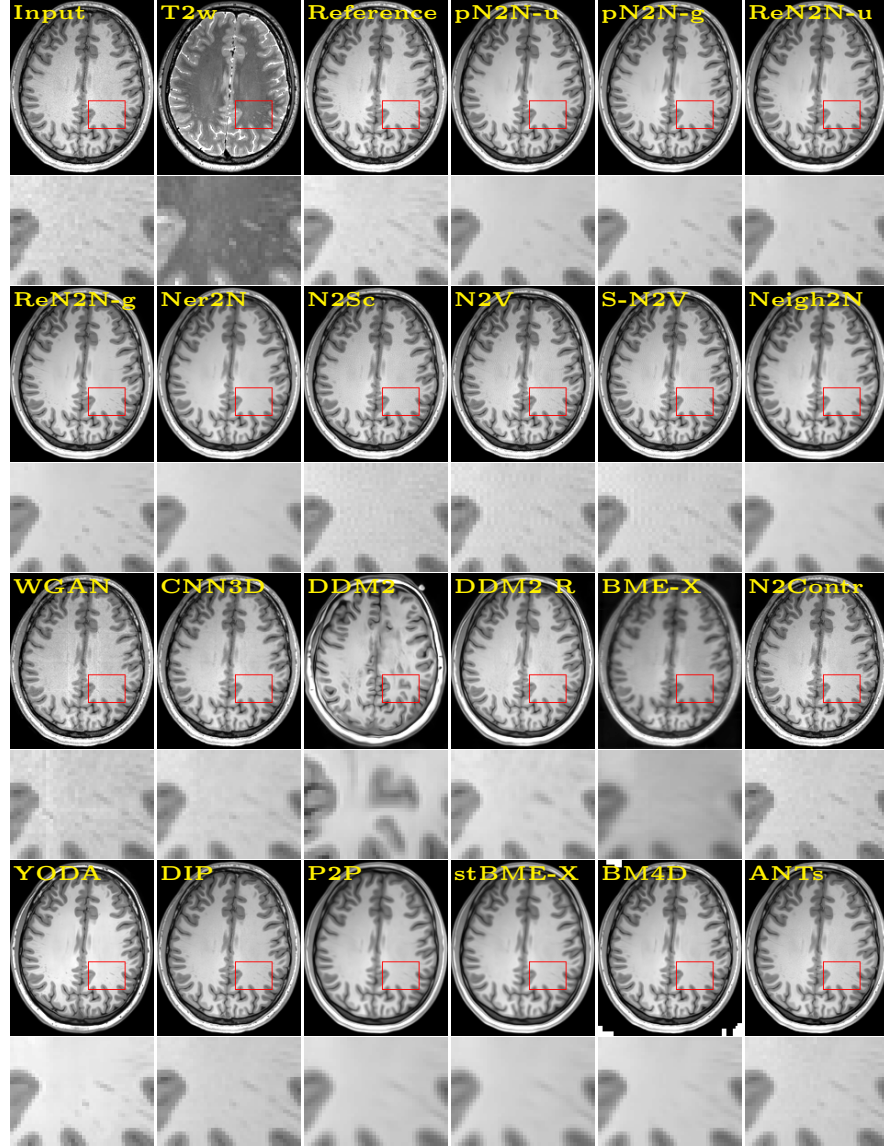


Fig. 16: RS T1w denoising performance (1 mm, from routine images $N_{\text{Ex}}=1$) with physically noise-suppressed reference images ($N_{\text{Ex}}=7$, ca. 45 min acquisition time).

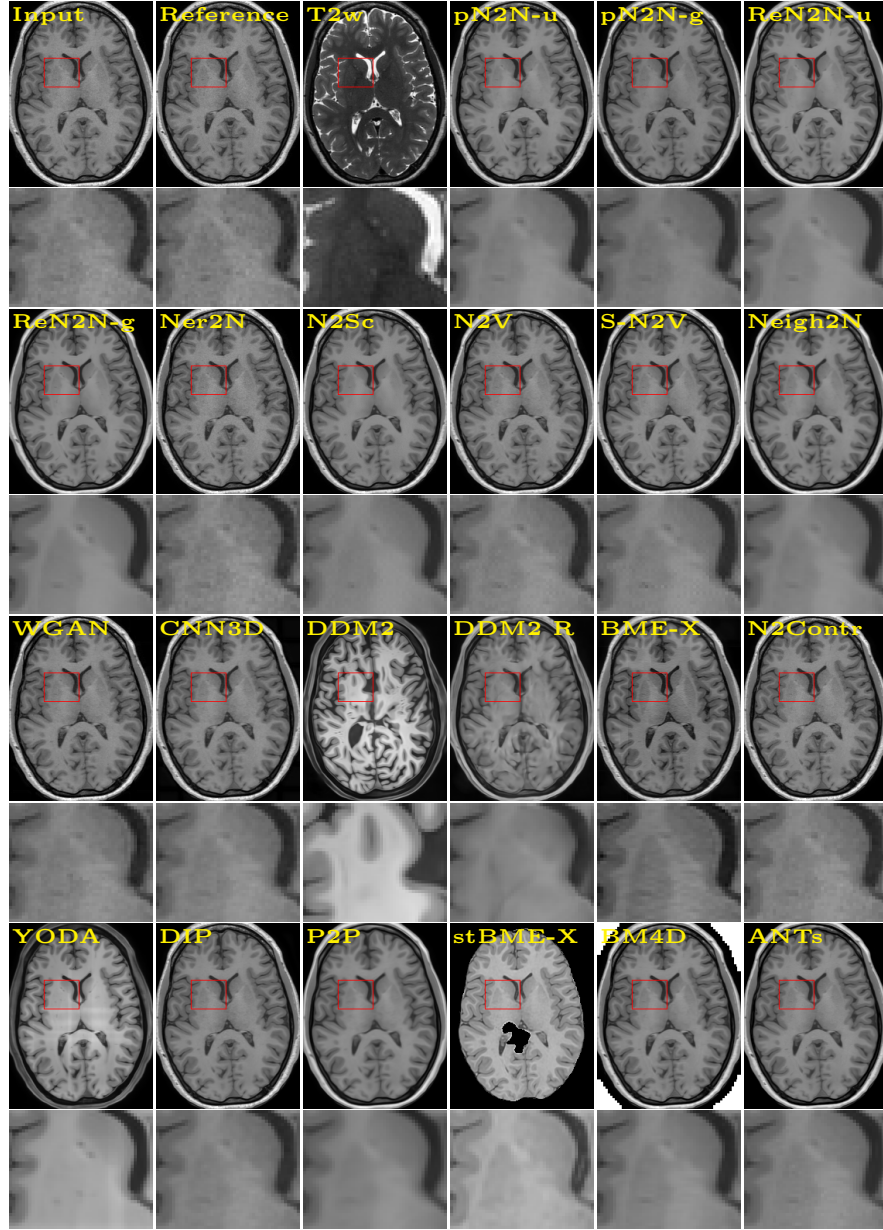


Fig. 17: T1w (0.8 mm) denoising performance in the HCP-E test set.

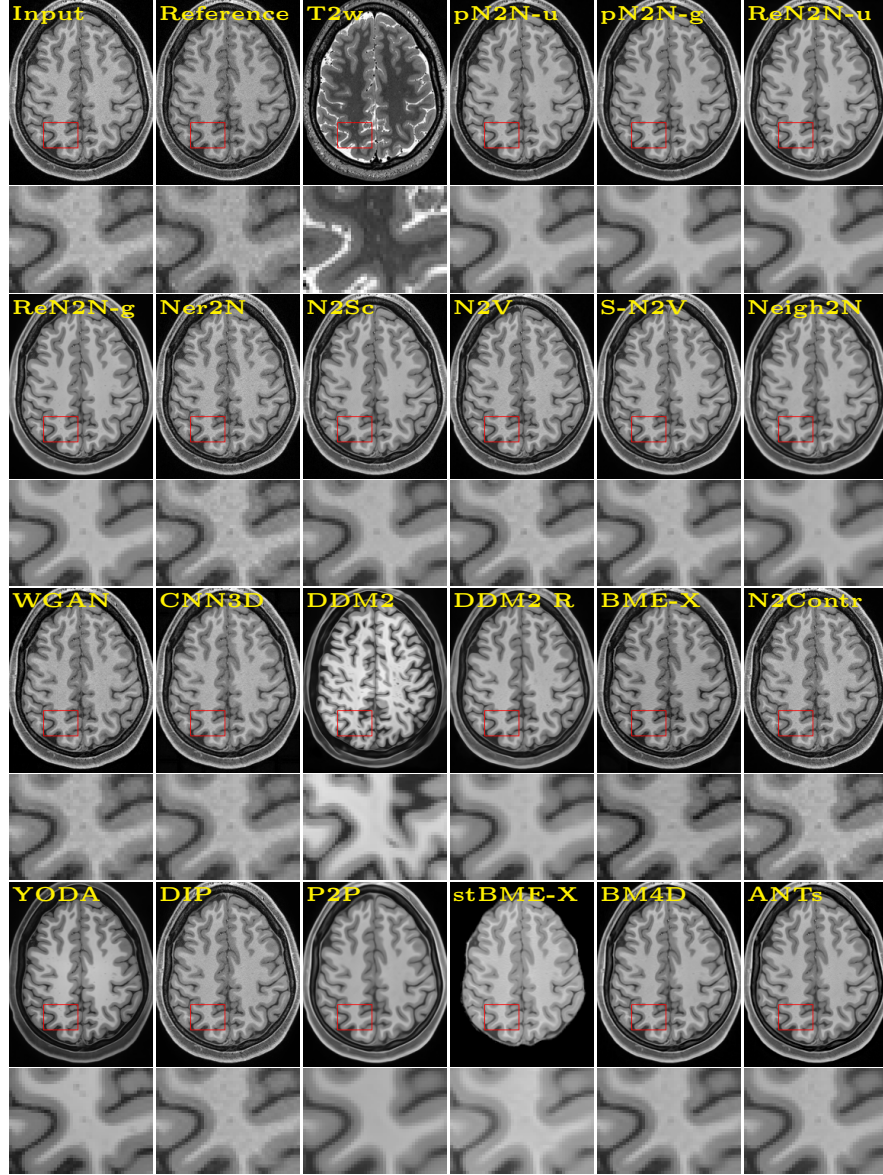


Fig. 18: T1w (0.8 mm) denoising performance in the HCP-YA test set.

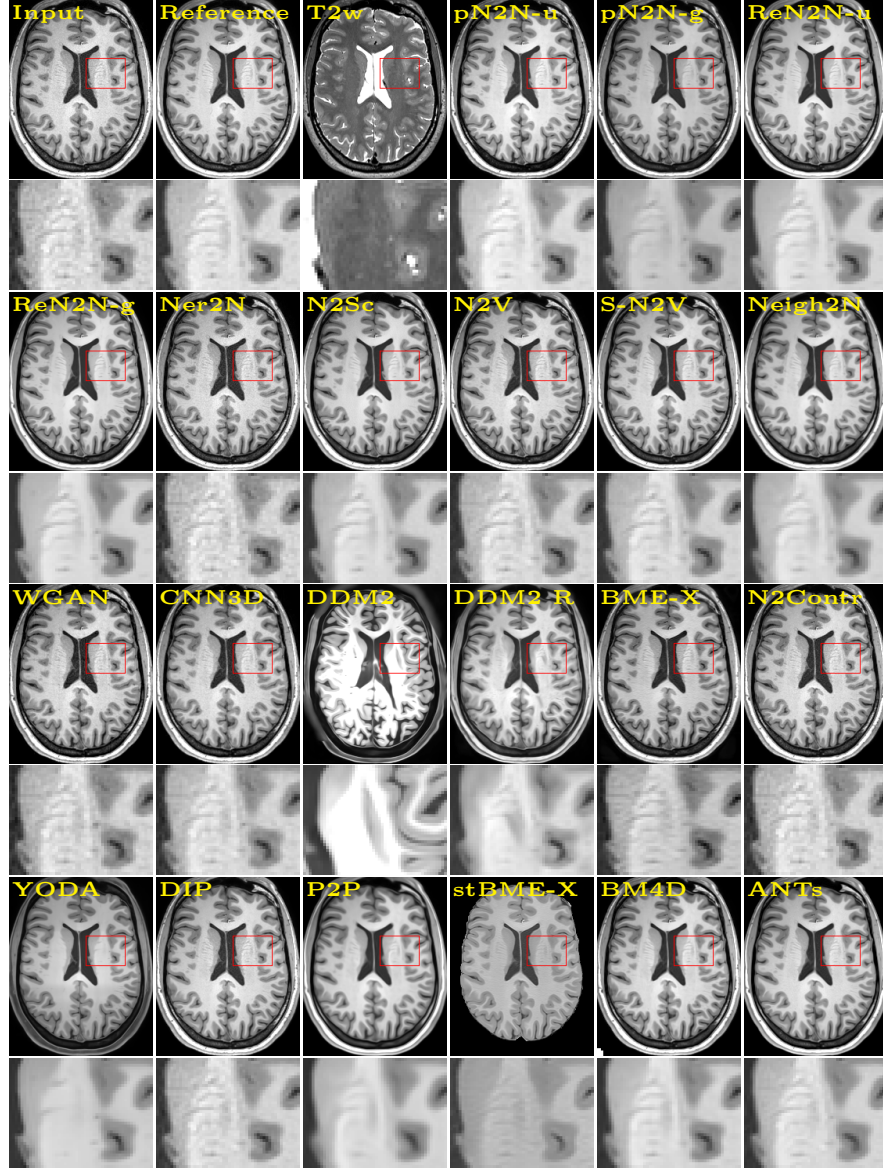


Fig. 19: RS T1w denoising performance (0.8 mm, from routine images $N_{Ex}=1$) with physically noise-suppressed reference images ($N_{Ex}=7$, ca. 45 min acquisition time).

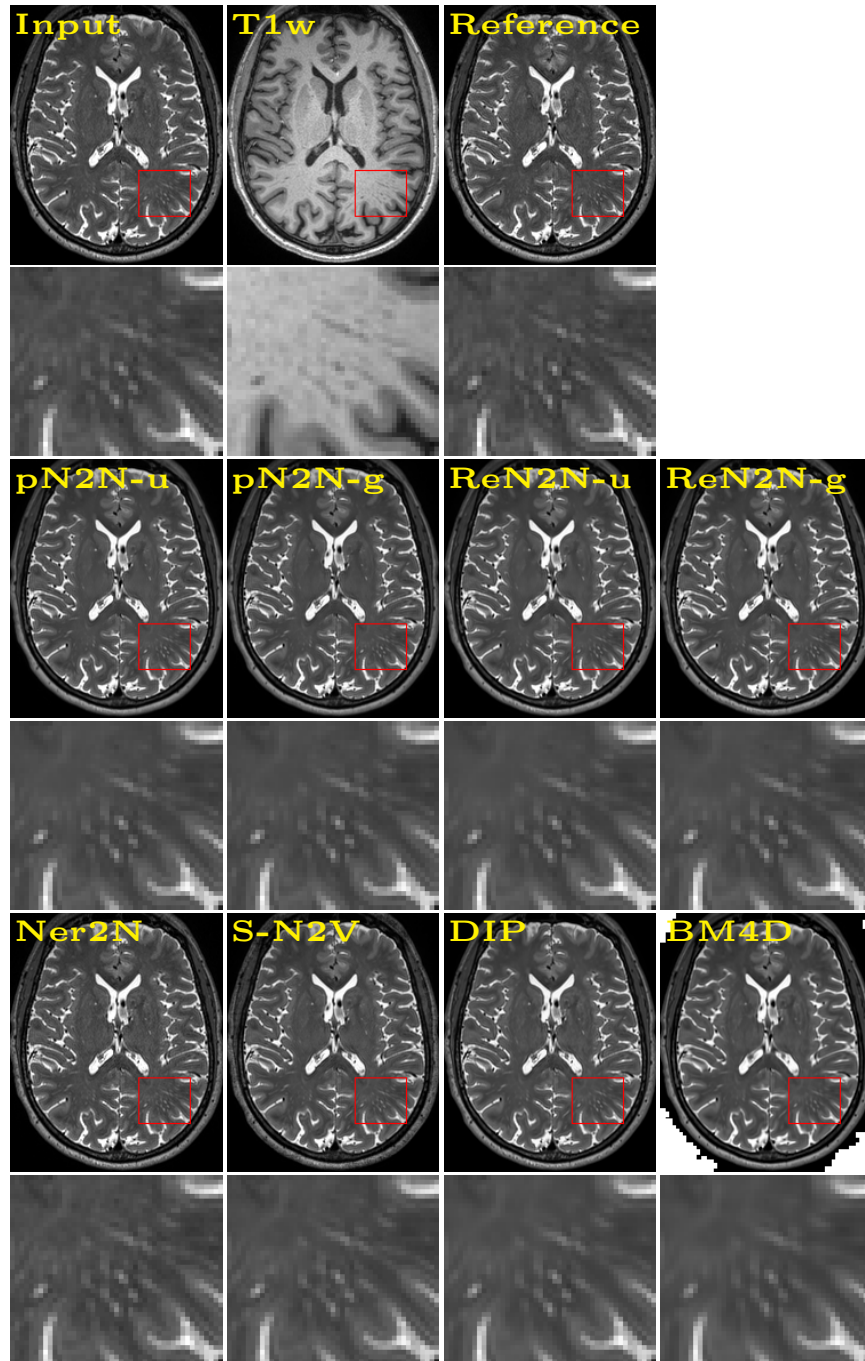


Fig. 20: T2w denoising results in the RS test set.

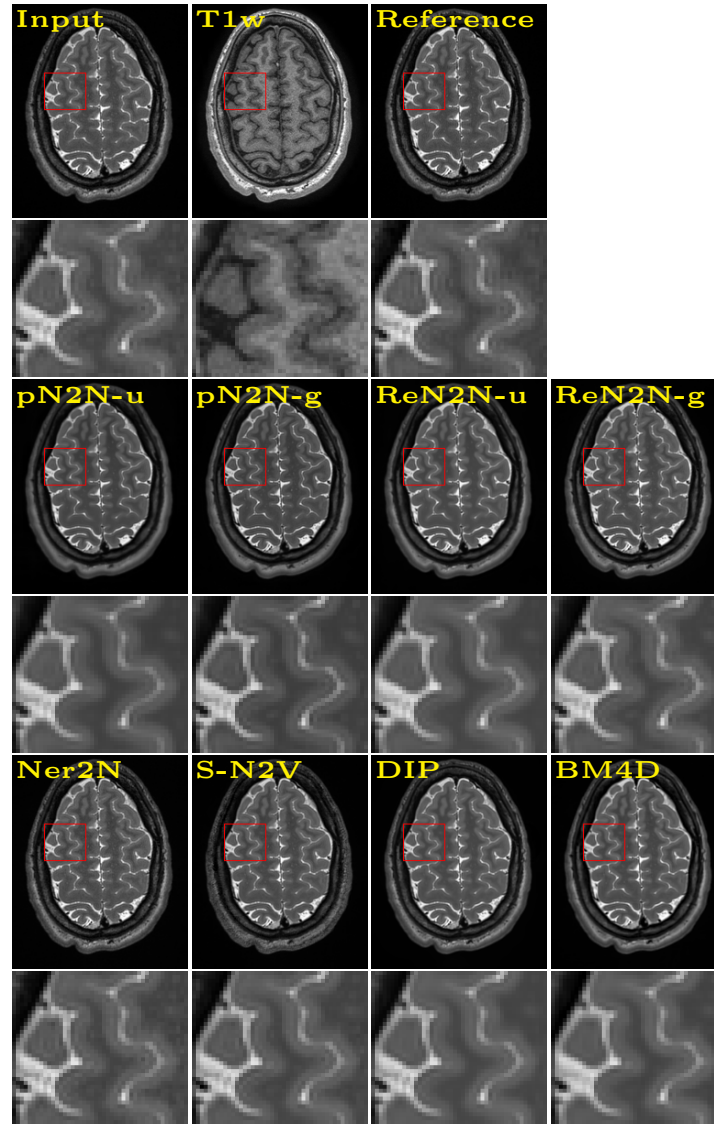


Fig. 21: T2w denoising results in HCP-E.

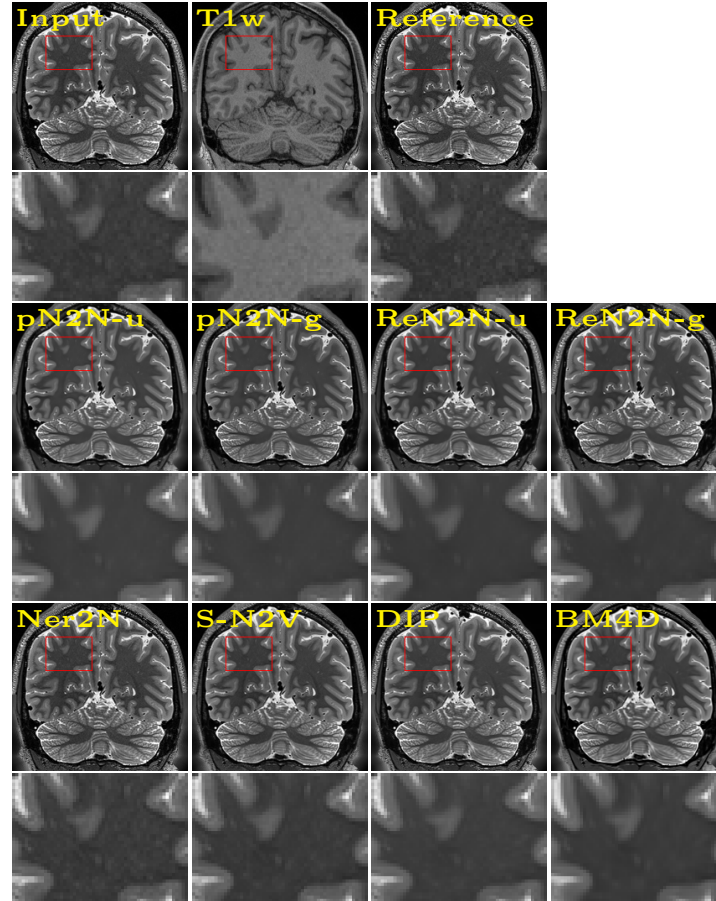


Fig. 22: T2w denoising results in HCP-YA.

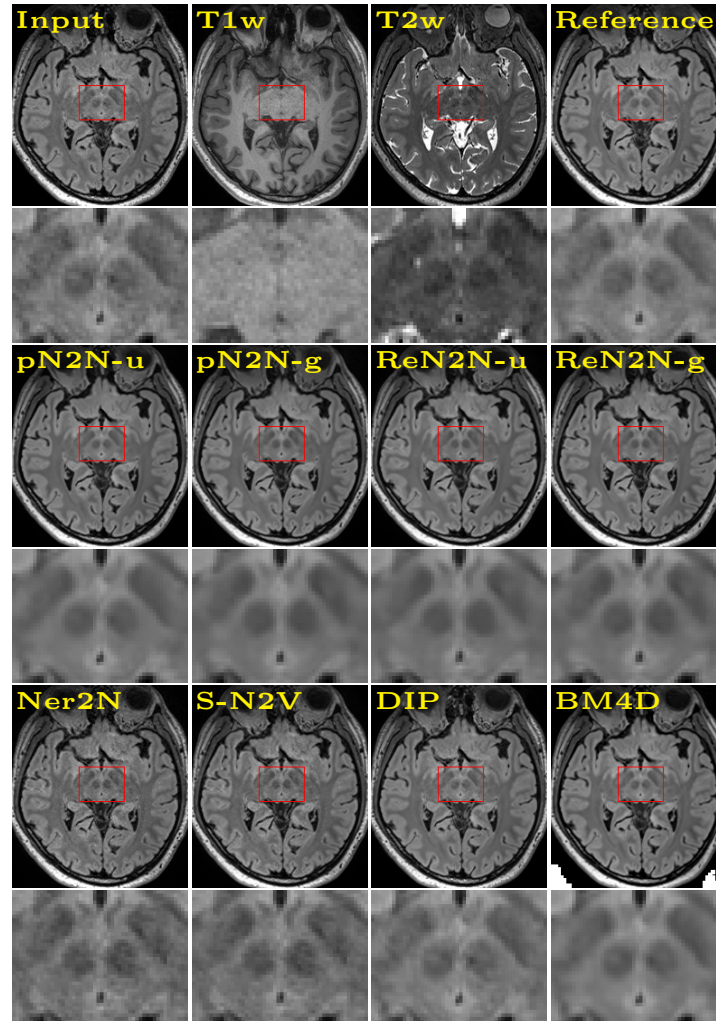


Fig. 23: FLAIR denoising results in the RS (0.8 mm).

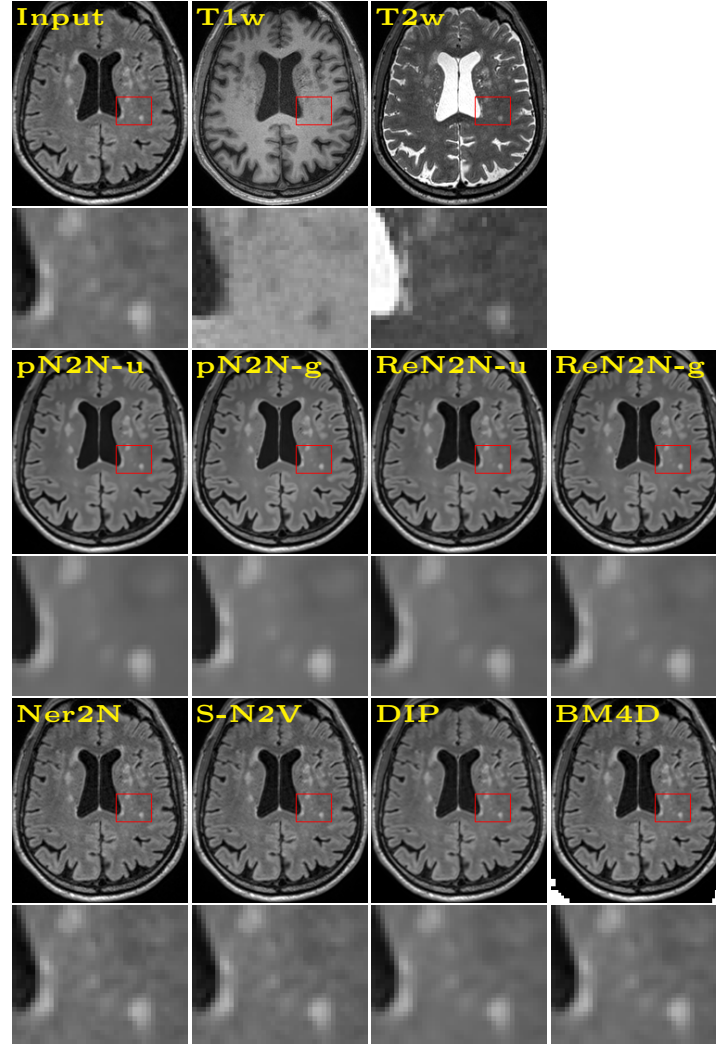


Fig. 24: FLAIR denoising results for *routine* RS (1.0 mm, as used for WMH analysis). Note that this test dataset does not entail independent re-acquisitions such that evaluation was constraint to annotated WMHs.

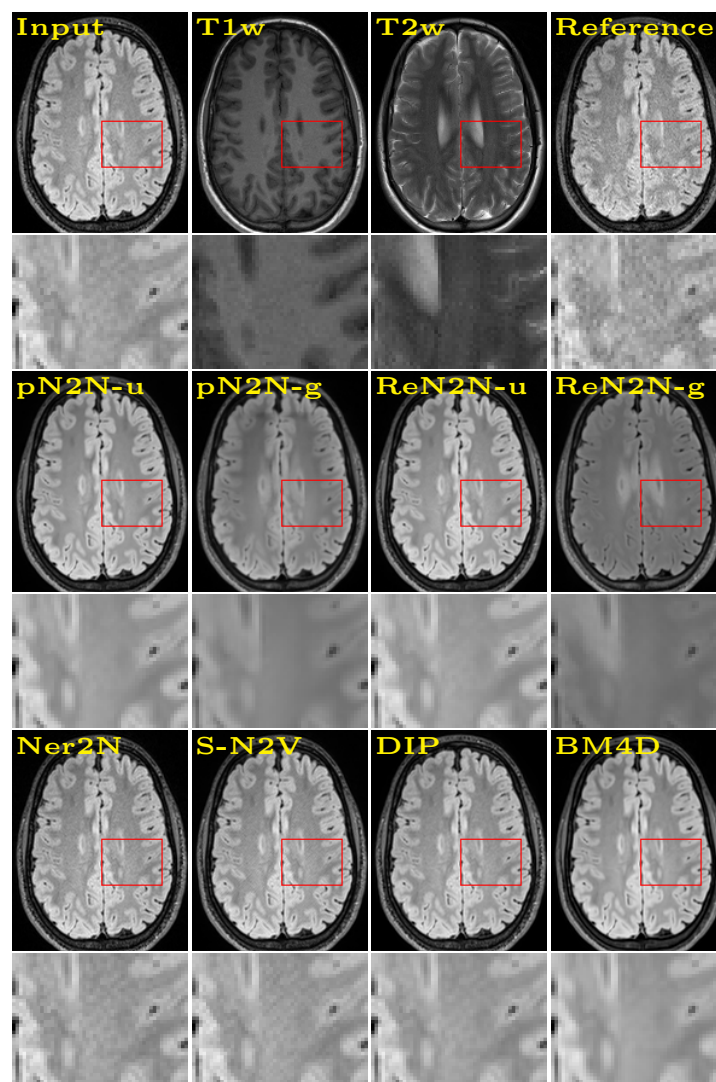


Fig. 25: FLAIR denoising results in motion dataset (still, 1.0 mm) [17].
 Note that T2w acquisition is highly anisotropic along the inferior-superior axis.

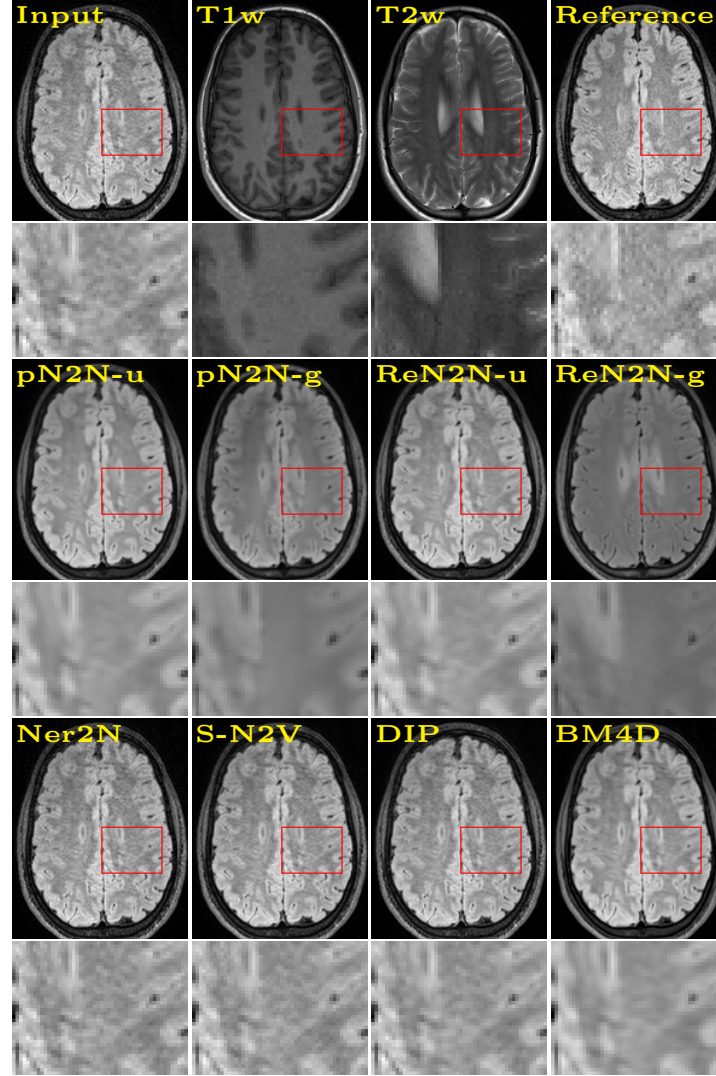


Fig. 26: FLAIR denoising results in motion dataset (nodding, 1.0 mm) [17]. Note that T2w acquisition is highly anisotropic along the inferior-superior axis.