N5: A Unified FiLM-Conditioned Output-Stack Cascade for Cross-Field Brain MRI Translation

Yuehan Wu[1], Lu Wang, Siyuan Ge[1], Bo Liu[1], Jenq-Neng Hwang[2], and Cheng Cheng Zhu[1]

[1] Department of Radiology, University of Washington, Seattle, WA, USA
zhucheng@uw.edu
[2] Department of Electrical and Computer Engineering, University of Washington, Seattle, WA, USA

Abstract. Brain MRI appearance varies substantially across magnetic field strengths, which complicates cross-field harmonization. This task requires same-contrast any-to-any translation with a single conditional model rather than separate pairwise generators. We propose N5, a five-stage, nine-slice output-stack residual cascade for cross-field brain MRI translation. The model uses Feature-wise Linear Modulation (FiLM) [6] to condition a shared residual backbone on source field, target field, contrast, ordered field pair, and slice position. Direction- and contrast-specific synthetic pairs are used as offline training augmentation, while inference uses one N5 checkpoint. A training-derived brightness prior is used for synthetic-volume calibration and selective correction of intensity outliers. On 180 development validation predictions, N5 achieves SSIM 0.900799, nRMSE 0.243617, and LPIPS 0.113774, improving over source copy and the provided StarGAN v2 baseline. Targeted experiments suggest associations with synthetic pretraining and nine-slice context, whereas applying the same two-stage training schedule to StarGAN v2 does not improve its baseline. Independent external-cohort and clinical validation remain necessary.

Keywords: Cross-field MRI synthesis · MRI harmonization · FiLM conditioning · Low-field MRI

1 Introduction

Magnetic resonance imaging (MRI) appearance varies substantially across magnetic field strengths. Ultra-low-field systems improve accessibility and portability, while clinical 1.5T/3T and high-field 5T/7T scanners provide different trade-offs in signal-to-noise ratio, tissue contrast, spatial detail, and noise characteristics. Cross-field translation may support research-use harmonization, retrospective scanner comparison, and simulation of field-dependent image appearance. However, it is also a medically sensitive synthesis problem: a visually plausible output may suppress lesions, alter anatomical boundaries, or hallucinate high-frequency structures. Therefore, a successful model should modify field-dependent appearance while preserving the underlying anatomical and pathological content.

We study a unified same-contrast cross-field translation task covering five field strengths and three MRI contrasts. Unlike pairwise translation, all directions are handled by one shared model. This requires direction-aware conditioning, local through-plane context, and a training strategy that can use unpaired data without adding inference-time generators.

We propose N5, a unified conditional 2.5D residual cascade for cross-field MRI synthesis. It combines FiLM conditioning, output-stack refinement, pseudo-pair generation, and a training-derived brightness prior. Under the development validation protocol, N5 improves SSIM, nRMSE, and LPIPS relative to source copy and StarGAN v2.

Our main contributions are summarized as follows:

1. We propose a unified 2.5D output-stack cascade for same-contrast cross-field MRI translation.
2. We introduce a synthetic-to-paired curriculum based on direction-specific pseudo-pairs.
3. We use a training-derived brightness prior for synthetic calibration and selective correction of intensity outliers.

2 Related Work

Unpaired image-to-image translation methods such as CycleGAN [1], CUT [2], and StarGAN v2 [3] provide frameworks for mapping between visual domains. In medical imaging, however, domain translation must be interpreted cautiously because visually plausible synthesis can alter anatomy or pathology [4, 5]. This limits the direct use of generic multi-domain translation when the desired output must preserve registered brain structure across field strengths.

MRI field-strength synthesis has also been studied with cascaded and residual convolutional networks. Bahrami et al. [9] used cascaded CNNs for 7T-like MRI synthesis from 3T images, and Qu et al. [10] combined spatial and wavelet-domain learning for 7T synthesis. These studies motivate cascaded refinement, while residual learning [8] is suitable for registered MRI because the network can focus on appearance changes around shared anatomy. The present setting additionally requires one model for all ordered field directions and contrasts.

Conditional normalization and Feature-wise Linear Modulation (FiLM) provide a compact way to inject task information into a shared network [6, 7]. In parallel, MRI intensity standardization methods, including landmark standardization, WhiteStripe, RAVEL, and ComBat [11, 13–15], show that intensity statistics can affect cross-scanner analysis. Our method uses FiLM for field/contrast conditioning and a training-derived brightness prior for synthetic-data calibration.

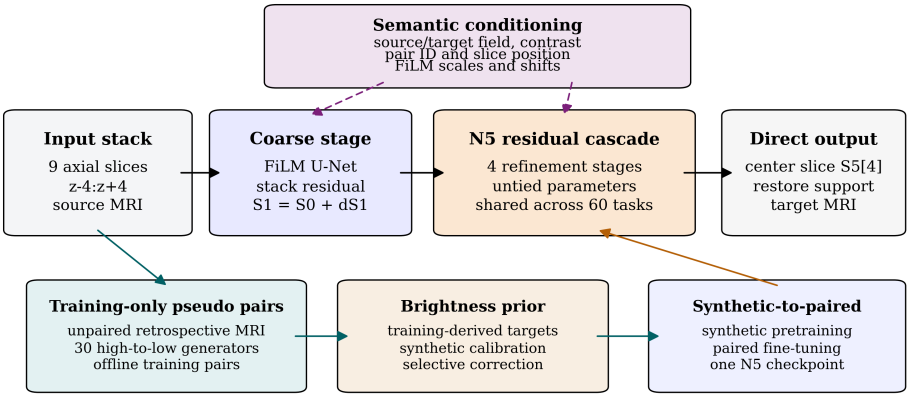

Fig. 1. Overview of the synthetic-to-paired curriculum and conditional residual cascade. N5 follows this principle with a 9-slice output-stack cascade: every stage translates a 2.5D stack, and the center slice is scored.

3 Method

3.1 Problem Formulation

Let

$$\mathcal{F} = \{0.1\text{T}, 1.5\text{T}, 3\text{T}, 5\text{T}, 7\text{T}\}, \quad \mathcal{M} = \{\text{T1W}, \text{T2W}, \text{T2FLAIR}\}$$

denote the field-strength and contrast sets. For each contrast, all non-identity ordered source-target field pairs are considered, giving $5 \times 4 = 20$ field directions and 60 same-contrast subtasks. Cross-contrast synthesis is not modeled.

For axial slice z, the model receives a 2.5D source stack

$$X_z = [x_{z-4}, x_{z-3}, \ldots, x_z, \ldots, x_{z+4}] \in \mathbb{R}^{9\times H\times W}, \tag{1}$$

and predicts the same-contrast target-field center slice conditioned on source field f_s, target field f_t, contrast m, and slice index z:

$$\hat{x}_z^{(f_t,m)} = G_\theta(X_z^{(f_s,m)}, f_s, f_t, m, z). \tag{2}$$

The task variables are used only as conditioning inputs; one generator checkpoint is shared across field directions and contrasts.

3.2 FiLM-Based Unified Task Conditioning

N5 treats cross-field translation as a conditional network with Feature-wise Linear Modulation (FiLM) [6]. A shared backbone is modulated by task variables rather than trained separately for each field direction. Source field, target field, contrast, task identity, and slice position are embedded and concatenated:

$$h_c = \text{MLP}([E_s(f_s); E_t(f_t); E_m(m); E_{\text{task}}(f_s, f_t, m); E_z(z)]). \tag{3}$$

For cascade stage k and residual block ℓ, h_c produces channel-wise scale and shift parameters

$$(\gamma_{k,\ell}, \beta_{k,\ell}) = W_{k,\ell} h_c + b_{k,\ell}, \tag{4}$$

and modulates feature tensor $F_{k,\ell}$ as

$$\mathrm{FiLM}(F_{k,\ell}; h_c) = F_{k,\ell} \odot (1 + \gamma_{k,\ell}) + \beta_{k,\ell}. \tag{5}$$

The task embedding is a lightweight conditioning token for the ordered field-pair-and-contrast identity. FiLM allows the cascade to adapt across directions and contrasts while retaining one checkpoint.

3.3 N5 Output-Stack Cascade

N5 denotes a five-stage cascade ($N = 5$) whose state is a nine-slice output stack. It is not five separate networks for five field strengths; it is one conditional residual cascade applied sequentially five times. The five stages use untied convolutional parameters, while the complete cascade is shared across all field directions and contrasts. Let $S_0 = X_z$ be the input stack. The coarse stage predicts a stack residual,

$$\Delta S_1 = C_{\theta_c}(S_0, c), \qquad S_1 = S_0 + \alpha_1 \Delta S_1, \tag{6}$$

where c denotes the conditioning variables. The residual scales α_k are fixed by the model configuration rather than learned; in the reported N5 model, $\alpha_k = 1$ for all stages. For refinement stages $k = 2, \ldots, 5$,

$$\Delta S_k = R_{\theta_k}(S_{k-1}, c), \qquad S_k = S_{k-1} + \alpha_k \Delta S_k. \tag{7}$$

The scored output is the center slice $S_5[4, :, :]$. Each stack generator has a U-Net-like residual backbone with two downsampling layers, nine FiLM ResNet blocks, and two upsampling layers. The N5 model has 29.47M parameters.

3.4 Loss and Training Curriculum

For center-slice prediction $\hat{x}$ and target x, the main loss is

$$\mathcal{L}_{\mathrm{main}} = \mathcal{L}_1(\hat{x}, x) + \lambda_{\mathrm{ssim}} \mathcal{L}_{\mathrm{ssim}}(\hat{x}, x) + \lambda_{\mathrm{freq}} \mathcal{L}_{\mathrm{freq}}(\hat{x}, x), \tag{8}$$

with $\lambda_{\mathrm{ssim}} = 1.0$ and $\lambda_{\mathrm{freq}} = 0.10$. The frequency term uses log-compressed orthonormal FFT magnitudes with radial high-frequency weighting. Intermediate center predictions are supervised with the same objective, and the output stack receives slice-wise supervision over all nine slices:

$$\mathcal{L} = \mathcal{L}_{\mathrm{main}} + 0.2\, \mathcal{L}_{\mathrm{aux}} + 0.5\, \mathcal{L}_{\mathrm{stack}}. \tag{9}$$

Training follows a synthetic-to-paired curriculum. Synthetic pretraining is followed by paired fine-tuning using prospective paired training data.

3.5 Pseudo/Synthetic Data Generation

Pseudo-paired data are generated from the unpaired retrospective training pool of 1,939 real volumes. Rather than using one mixed generator for all fields and contrasts, pseudo-pairs are produced offline by lightweight direction- and contrast-specific degradation generators over three contrasts and ten high-to-low field mappings, giving 30 generation tasks. The pseudo generators are used only to create training pairs and are not used during N5 inference.

The training-only synthetic manifest contains 3,800 volumes across T1W, T2W, and T2FLAIR. These generators are training-data tools only; they are discarded after pseudo-pair construction, and the inference system remains a single conditional N5 checkpoint. After synthetic pretraining, N5 is fine-tuned on paired prospective training data over the ordered field directions and contrasts.

3.6 Synthetic Brightness Calibration and Selective Output Calibration

FiLM remains the main field-aware mechanism. The brightness prior is auxiliary, motivated by MRI intensity standardization work [11–14] and by evidence that normalization affects MR synthesis [16]. We define brightness as the foreground mean in the $[0, 1]$ image space over slices $z \in [150, 180)$, using $x > 0.02$ as the foreground mask. The cached images retain domain-specific brightness because preprocessing applies only the linear mapping $x_{\text{model}} = 2x_{\text{original}} - 1$. Training data define bright and dark contrast-field groups with fixed target means 0.50 and 0.25. The 0.50 group consists of T1W at 0.1T, 1.5T, and 3T; T2W at 0.1T; and T2FLAIR at 0.1T, 1.5T, 3T, and 5T. All remaining contrast-field combinations use 0.25. The fixed target map was derived from training data. The reported calibration configuration and its observed benefit were assessed through returned leaderboard scores; no robustness claim is made.

Synthetic volumes are calibrated before N5 training with a volume-level additive shift. For real volumes, the foreground is $x > 0.02$; for synthetic volumes, the synthetic foreground mask is intersected with the corresponding real high-field foreground mask. Bisection finds the shift whose clipped foreground mean matches the contrast-field target. The resulting offset manifest is used during data preparation.

After prediction, the same prior is applied selectively. Let μ_p be the prediction mean over source foreground $x_s > 0.02$, and let $\mu_t \in \{0.50, 0.25\}$ be the fixed target mean. If $|\mu_p - \mu_t| > 0.1$, bisection shifts the clipped output toward μ_t; otherwise it is unchanged. The rule uses only source information and the training-derived target map.

4 Experiments

4.1 Dataset and Evaluation Protocol

We use de-identified MRIxFields2026 challenge data released under the challenge data-use terms [22]. The data used in this study include unpaired retrospective

training volumes for pseudo-pair generation and intensity statistics, prospective paired training volumes for paired fine-tuning, and paired validation inputs for leaderboard-based development scoring. The final N5 paired stage uses the three paired prospective training subjects provided by the organizers over the axial slab $z \in [150, 180)$.

Development scoring follows the challenge submission structure over three paired validation subjects. For each subject, 20 ordered field directions and three contrasts are evaluated, giving $3 \times 20 \times 3 = 180$ prediction files. Validation predictions were submitted to the challenge leaderboard, and the returned metrics were used for development assessment. Because files from the same subject are correlated, the reported standard deviations characterize prediction-file-level variation and should not be interpreted as independent-subject confidence intervals. The final brightness manifest contains 5,784 volume-level entries, including 3,800 synthetic, 1,939 retrospective, and 45 paired prospective training volumes; non-synthetic entries are retained for bookkeeping and statistics rather than implying an applied brightness shift.

4.2 Baselines, Metrics, and Implementation

Baselines. We compare with source copy and the organizer-provided StarGAN v2 checkpoint. Source copy directly submits the source image as the target prediction and provides a co-registered similarity baseline. StarGAN v2 is evaluated through the same leaderboard submission structure.

Metrics. nRMSE is $\|\hat{x} - x\|_2 / \|x\|_2$. SSIM is computed slice-wise in 2D using the target-volume data range. LPIPS uses LPIPS-Alex on 2D grayscale slices replicated to three channels. These match the challenge metric definitions.

Implementation. Experiments used PyTorch 2.4.1, CUDA 12.1, Python 3.10.14, and three NVIDIA RTX 3090 GPUs with 24GB memory each. Paired fine-tuning used Adam, bfloat16 automatic mixed precision, batch size 21, and learning rate 5×10^{-6}. The paired stage was trained to 50 epochs, and checkpoints at epochs 5, 10, 20, 30, 40, and 50 were submitted for leaderboard scoring. The selected checkpoint is epoch 30, which had the highest returned SSIM before selective output calibration. The resulting metrics should therefore be interpreted as adaptively selected leaderboard development evidence rather than independent external-test performance.

4.3 Main Results

N5 improves all three aggregate metrics relative to StarGAN v2 and source copy (Table 1). Before selective calibration, the selected checkpoint reaches SSIM 0.899441, nRMSE 0.253619, and LPIPS 0.112813. Selective calibration is triggered for 15 of 180 validation predictions, increasing SSIM to 0.900799, reducing nRMSE to 0.243617, and slightly increasing LPIPS to 0.113774. Most field directions are unchanged; the largest SSIM gain is for 3T $\rightarrow$ 7T (+0.018415),

Table 1. Development validation results over 180 correlated prediction files. Values are prediction-file-level mean ± standard deviation.

Method	SSIM ↑	nRMSE ↓	LPIPS ↓
Source copy	0.836 ± 0.076	0.522 ± 0.510	0.157 ± 0.069
StarGAN v2 baseline	0.740 ± 0.045	0.344 ± 0.250	0.154 ± 0.020
N5 + selective calibration	$\mathbf{0.901 \pm 0.043}$	$\mathbf{0.244 \pm 0.241}$	$\mathbf{0.114 \pm 0.044}$
Difference vs. source copy	+0.064	−0.278	−0.044
Difference vs. StarGAN v2	+0.161	−0.100	−0.041

while 3T→1.5T (-0.005565) and 0.1T→1.5T (-0.004976) slightly decrease. Per contrast, T2W has the highest SSIM, while T2FLAIR has the lowest nRMSE (Table 2).

Table 2. Per-contrast N5 + selective calibration metrics on development validation.

Contrast	SSIM ↑	nRMSE ↓	LPIPS ↓
T1W	0.897 ± 0.036	0.272 ± 0.192	0.124 ± 0.040
T2W	0.909 ± 0.048	0.263 ± 0.339	0.101 ± 0.040
T2FLAIR	0.896 ± 0.044	0.196 ± 0.144	0.116 ± 0.049

Table 3. Subset-level development validation metrics. Values are means over correlated validation files.

Subset / method	SSIM ↑	nRMSE ↓	LPIPS ↓
All / StarGAN v2	0.740	0.344	0.154
All / N5 + selective calibration	**0.901**	**0.244**	**0.114**
Low-to-high / StarGAN v2	0.740	0.372	0.150
Low-to-high / N5 + selective calibration	**0.895**	**0.269**	**0.122**
0.1T-source / StarGAN v2	0.737	0.326	**0.151**
0.1T-source / N5 + selective calibration	**0.874**	**0.222**	0.159

For low-to-high translation, N5 improves SSIM from 0.740 to 0.895 and reduces nRMSE from 0.372 to 0.269 (Table 3). For the 0.1T-source subset, N5 improves SSIM and nRMSE, while LPIPS remains slightly worse.

4.4 Ablation Study

We conduct three targeted experiments on the same 180-prediction development protocol: removing synthetic pretraining, reducing the input from nine slices to one slice, and training the official StarGAN v2 architecture with a two-stage synthetic-to-paired schedule.

Table 4. Targeted ablations on the same 180-prediction development validation protocol. "Cal." denotes selective fixed-brightness output calibration.

Experiment	Epoch	Cal.	SSIM ↑	nRMSE ↓	LPIPS ↓
Full N5	30	yes	**0.900799**	**0.243617**	0.113774
No synthetic pretraining	40	no	0.894826	0.249784	0.130763
No synthetic pretraining	40	yes	0.894849	0.246377	0.131940
One-slice input	20	no	0.896696	0.253117	0.120801
One-slice input	20	yes	0.898104	0.242675	0.121365
StarGAN v2 two-stage	10	no	0.738773	0.344217	0.159438
StarGAN v2 two-stage	10	yes	0.738994	0.342524	0.159544

These targeted, non-compute-matched comparisons suggest associations with synthetic pretraining and slice context; they do not constitute controlled causal

ablations. Synthetic pretraining is associated with approximately 0.006 higher SSIM after calibration, and the nine-slice input is associated with approximately 0.003 higher SSIM than the one-slice variant. Two-stage StarGAN v2 training does not improve the provided baseline; its best calibrated SSIM remains slightly lower than the official checkpoint (0.738994 vs. 0.739694).

4.5 Visualization

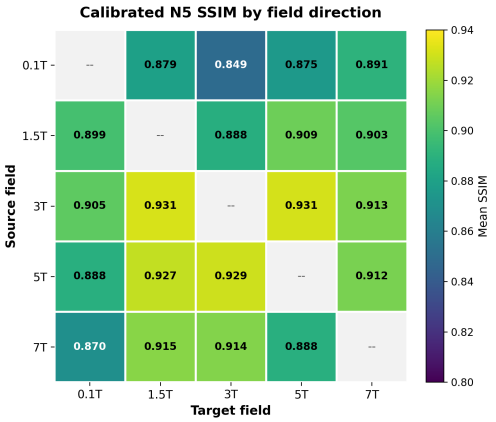

Fig. 2. Per-field-direction SSIM averaged over contrasts and validation subjects.

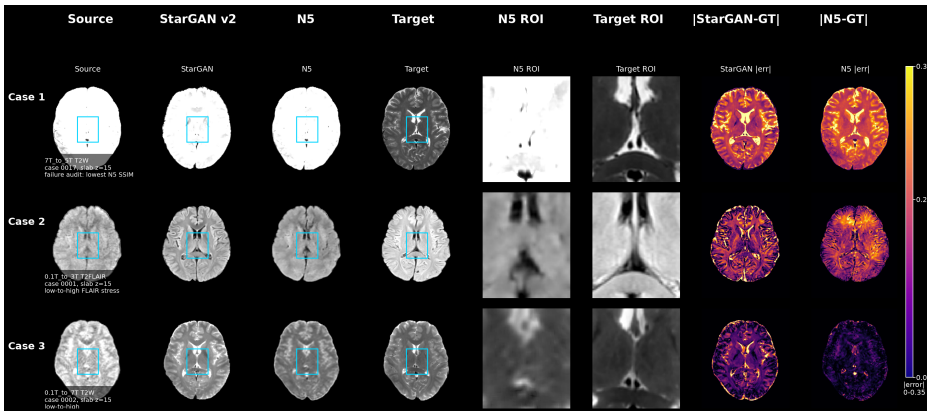

Fig. 3. Qualitative comparison with source, StarGAN v2, N5, target, ROI boxes, and fixed-scale absolute residual maps. Residual maps use a shared absolute-error range of 0–0.35 in normalized intensity.

Figure 2 summarizes SSIM by ordered field direction; 0.1T-source cases remain more difficult. In Figure 3, the ROI regions cover ventricular margins and gray–white matter boundaries. In these examples, N5 shows closer tissue-boundary agreement than StarGAN v2, with residual errors mainly near interfaces and difficult low-to-high translations.

5 Discussion and Conclusion

N5 is a compact conditional model for research-use cross-field MRI harmonization. A single checkpoint is relevant to low-field and portable MRI, where reviews emphasize lower infrastructure requirements and improved access [18]; small local models may also reduce deployment and data-transfer burden in edge medical imaging [19]. The residual formulation frames synthesis as appearance adjustment around registered anatomy, but the study remains limited by adaptive development-set selection, untested calibration-threshold robustness, and the absence of a simpler conditional U-Net or residual baseline. Independent external-cohort, pathology-aware, reader-based, and uncertainty-aware validation are needed before clinical or diagnostic use.

References

1. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: ICCV (2017)
2. Park, T., Efros, A.A., Zhang, R., Zhu, J.Y.: Contrastive learning for unpaired image-to-image translation. In: ECCV (2020)
3. Choi, Y., Uh, Y., Yoo, J., Ha, J.W.: StarGAN v2: Diverse image synthesis for multiple domains. In: CVPR (2020)
4. Skandarani, Y., Jodoin, P.M., Lalande, A.: GANs for medical image synthesis: An empirical study. Journal of Imaging 9(3), 69 (2023)
5. Jung, H.K., Kim, K., et al.: Image-based generative artificial intelligence in radiology: comprehensive updates. Korean Journal of Radiology 25(11), 959–981 (2024)
6. Perez, E., Strub, F., de Vries, F., Dumoulin, V., Courville, A.: FiLM: Visual reasoning with a general conditioning layer. In: AAAI (2018)
7. Dumoulin, V., Shlens, J., Kudlur, M.: A learned representation for artistic style. In: ICLR (2017)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
9. Bahrami, K., Rekik, I., Shi, F., Shen, D.: Joint reconstruction and segmentation of 7T-like MR images from 3T MRI based on cascaded convolutional neural networks. In: MICCAI, LNCS 10433, pp. 764–772 (2017)
10. Qu, L., Zhang, Y., Wang, S., Yap, P.T., Shen, D.: Synthesized 7T MRI from 3T MRI via deep learning in spatial and wavelet domains. Medical Image Analysis 62, 101663 (2020)
11. Nyúl, L.G., Udupa, J.K.: On standardizing the MR image intensity scale. Magnetic Resonance in Medicine 42(6), 1072–1081 (1999)
12. Shah, M., Xiao, Y., Subbanna, N., Francis, S., Arnold, D.L., Collins, D.L., Arbel, T.: Evaluating intensity normalization on MRIs of human brain with multiple sclerosis. Medical Image Analysis 15(2), 267–282 (2011)
13. Shinohara, R.T., Sweeney, E.M., Goldsmith, J., et al.: Statistical normalization techniques for magnetic resonance imaging. NeuroImage: Clinical 6, 9–19 (2014)
14. Fortin, J.P., Sweeney, E.M., Muschelli, J., Crainiceanu, C.M., Shinohara, R.T., et al.: Removing inter-subject technical variability in magnetic resonance imaging studies. NeuroImage 132, 198–212 (2016)
15. Fortin, J.P., Parker, D., Tunç, B., et al.: Harmonization of multi-site diffusion tensor imaging data. NeuroImage 161, 149–170 (2017)
16. Reinhold, J.C., Dewey, B.E., Carass, A., Prince, J.L.: Evaluating the impact of intensity normalization on MR image synthesis. In: Medical Imaging 2019: Image Processing, vol. 10949, 109493H (2019)
17. Abbasi, S., Lan, H., Choupan, J., et al.: Deep learning for the harmonization of structural MRI scans: a survey. BioMedical Engineering OnLine 23, 90 (2024)
18. Kravchenko, D., Hagar, M.T., Vecsey-Nagy, M., et al.: Low-field and portable MRI technology: advancements and innovations. European Radiology Experimental 9, 103 (2025)
19. Xu, Y., Khan, T.M., Song, Y., Meijering, E.: Edge deep learning in computer vision and medical diagnostics: a comprehensive survey. Artificial Intelligence Review 58, 93 (2025)
20. 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)

21. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018)
22. MRIxFields2026 Organizers: MRIxFields2026 challenge documentation (2026)