FieldMorphUNet: Conditional Multi-Field MRI Translation with Structure-Scale Decoupling

Zengxing Li[1*], Yue Yang[3*], Kaiqiang Xu[1], Zi Yang[2], Weizhi Zhao[1], and Yu Fu[1†]

[1] Lanzhou University, Lanzhou, China
fuyu@lzu.edu.cn
[2] The First Hospital of Lanzhou University, Lanzhou, China
[3] Independent Researcher

Abstract. Multi-field MRI translation must synthesize target-field contrast while preserving anatomy and coping with unstable absolute intensity. We present FieldMorphUNet, a unified conditional model for modality-preserving translation across 15 modality–field domains (five field strengths and three modalities). The final model employs Feature-wise Linear Modulation (FiLM)-modulated multi-contrast 2.5D residual U-Net blocks, Multi-Dconv Head Transposed Attention (MDTA), bottleneck spatial multi-head attention, and a compact residual refiner. Crucially, we propose a structure–scale decoupling strategy to overcome the instability of few-shot intensity scaling: the network first predicts normalized target structures, and a prospective population target-scale prior (pro prior) subsequently recovers raw absolute intensities. Training combines paired prospective translation, unpaired retrospective reconstruction, and staged perceptual refinement with metric-aligned losses. Evaluated on the MRIxFields validation leaderboard, FieldMorphUNet improves all reported metrics across the three challenge tasks. Controlled baseline comparisons evaluate the final model against simpler and alternative generative formulations, while leave-one-subject-out experiments highlight pro calibration as a critical driver of nRMSE improvement.

Keywords: Magnetic resonance imaging translation · Multi-field magnetic resonance imaging · Conditional U-Net · Feature-wise Linear Modulation · Intensity calibration

1 Introduction

MRI appearance depends on field strength, sequence, and subject-specific intensity scale. Cross-field synthesis supports harmonization, but must preserve anatomy while matching field-dependent contrast, noise, and scale. We study three MRIxFields tasks (Fig. 2a): any-field→7T (Task 1), 0.1T→higher (Task 2), and fully controllable field-to-field synthesis (Task 3), covering modalities $m \in$

* Zengxing Li and Yue Yang contributed equally to this work.
† Corresponding author: Yu Fu (fuyu@lzu.edu.cn).

$\{\mathrm{T1W}, \mathrm{T2W}, \mathrm{T2FLAIR}\}$ and fields $f \in \{0.1\mathrm{T}, 1.5\mathrm{T}, 3\mathrm{T}, 5\mathrm{T}, 7\mathrm{T}\}$ (15 modality–field domains).

Training one network per translation direction is impractical when only a few fully paired subjects are available. FieldMorphUNet uses a single generator conditioned on the ordered triplet (m, s, t) and incorporates multi-contrast 2.5D context to exploit complementary anatomical information across modalities. Our central claim is that structure prediction and absolute scale recovery should be decoupled: raw-intensity regression can produce anatomically plausible images yet fail nRMSE when the global scale is incorrect. We therefore learn in a normalized structure space and recover absolute intensity through an explicit prospective population target-scale prior, allowing structural synthesis and intensity calibration to be handled separately.

Contributions: (i) a single FiLM-conditioned multi-contrast 2.5D residual U-Net for all evaluated directions and tasks; (ii) structure–scale decoupling with brain-$p_{99.5}$ normalization and prospective population target-scale recovery; (iii) paired prospective translation and unpaired retrospective reconstruction with metric-aligned losses and staged residual refinement; and (iv) validation across all three challenge tasks, complemented by controlled leave-one-subject-out scale-recovery analysis showing that pro calibration is a primary driver of nRMSE improvement.

2 Related Work

Medical image-to-image translation commonly builds on U-Nets and conditional GANs [10,3,6]. For cross-domain synthesis, recent advances include adversarial multi-objective training (MPGAN [1], AIGAN [2]), cross-dataset decomposition (FADFNet [9]), and frequency-aware mapping for unpaired MRI (FA^2-Net [15]). Flow matching provides an alternative non-adversarial generative formulation by learning vector fields along conditional probability paths [4]. To manage multiple domains within a single model, Feature-wise Linear Modulation (FiLM) [8] provides explicit conditioning, while Restormer-style attention [13] efficiently captures long-range dependencies. Training and evaluation commonly use Structural Similarity (SSIM) and Learned Perceptual Image Patch Similarity (LPIPS) [12,14] as structural and perceptual criteria. Intensity standardization is a classical MRI problem [7,11], but multi-field translation makes field strength a source of both contrast and absolute-scale variation. Accordingly, our method builds on established conditional and attention mechanisms while explicitly decoupling normalized structure prediction from prospective population-based intensity calibration.

3 Method

3.1 Normalized Structure Space and Multi-Contrast Input

Let $I_u^{m,f} \in [0,1]^{H \times W \times Z}$ denote the image of subject u, modality m, and field strength f. We define the robust brain-$p_{99.5}$ reference and the corresponding

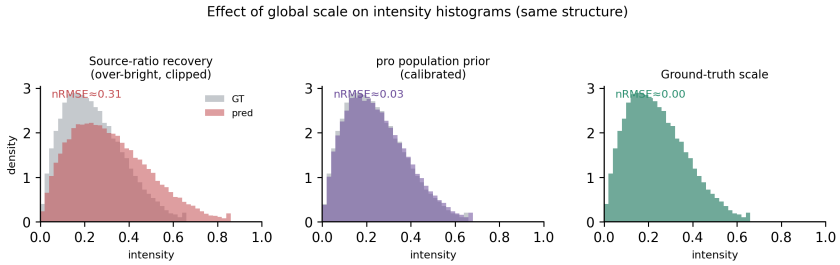

Fig. 1. Global scale error with fixed structure S. Over-bright source-ratio recovery (left) inflates nRMSE and clips high intensities; the pro prior (middle) more closely matches the ground-truth scale distribution (right).

normalized image as

$$r(I) = p_{99.5}(\{v \in I : v > 0.01\}), \qquad \tilde{I} = \mathrm{clip}\big(I/(r(I)+\epsilon), 0, 1\big), \tag{1}$$

where $\epsilon > 0$ prevents numerical instability. Each source contrast is normalized by its own reference, while the paired target is normalized by its target-volume reference.

For a requested primary modality m, source field s, and target field t, the generator receives five neighboring axial slices from each of the three source contrasts. With k=2, the input is

$$X_z^{m,s} = \mathrm{cat}_{m' \in \mathcal{O}(m)} \left[\tilde{I}_{z-2}^{m',s}, \ldots, \tilde{I}_{z}^{m',s}, \ldots, \tilde{I}_{z+2}^{m',s}\right] \in \mathbb{R}^{15 \times H \times W}, \tag{2}$$

where $\mathcal{O}(m)$ places the requested modality first and the other two modalities afterward in a fixed order. Missing superior or inferior neighbors are filled by edge replication. The network predicts the normalized target center slice $\tilde{I}_z^{m,t}$, and a volume is assembled by applying the same operation over all axial indices. This multi-contrast 2.5D representation provides both through-plane context and complementary anatomical information without requiring a full 3D backbone.

At inference, the target reference is unavailable. The classical estimator $\hat{r}_t = r_s R_{m,s\to t}$ couples the output scale to a few-shot source-to-target ratio and is unstable when only a small paired cohort is available; for example, the T1W 5T$\rightarrow$ 7T ratios range from 0.68 to 1.51 (Fig. 3a). We therefore recover the absolute target scale using the population prior described in Section 3.4. If the normalized structure is fixed and only a relative scale error $\delta = (s - s^*)/s^*$ remains, then nRMSE $= |\delta|$ (Fig. 1), showing that scale recovery directly affects quantitative fidelity.

3.2 Conditional Multi-Contrast Generator

The stage-one generator F_θ is a residual U-Net with GroupNorm, SiLU activations, skip connections, and a sigmoid output head (Fig. 2b). The condition vector is constructed as

$$c = \mathrm{MLP}([E_m(m), E_s(s), E_t(t)]), \tag{3}$$

where E_m, E_s, and E_t are learnable embeddings for the requested modality, source field, and target field, respectively. Each residual block applies Feature-wise Linear Modulation after GroupNorm:

$$\mathrm{FiLM}(h, c) = (1 + \gamma(c)) \odot \mathrm{GN}(h) + \beta(c), \tag{4}$$

where $\gamma(c)$ and $\beta(c)$ are channel-wise linear projections of c.

The encoder uses a base width of C=96 and increases the channel dimensions as $C, 2C, 4C, 8C$; the decoder mirrors this hierarchy with skip concatenations. Multi-Dconv Head Transposed Attention (MDTA) is applied at intermediate encoder and decoder resolutions to model channel-wise dependencies efficiently, while spatial Multi-Head Attention (MHA) operates on the lowest-resolution features around the bottleneck to provide global spatial context. The sigmoid head produces the coarse normalized target slice $\hat{y}_c$. Separate embeddings for s and t preserve the ordering of transformations such as 3T$\rightarrow$ 7T and 7T$\rightarrow$ 3T, while the condition triplet (m, s, t) specifies the requested translation.

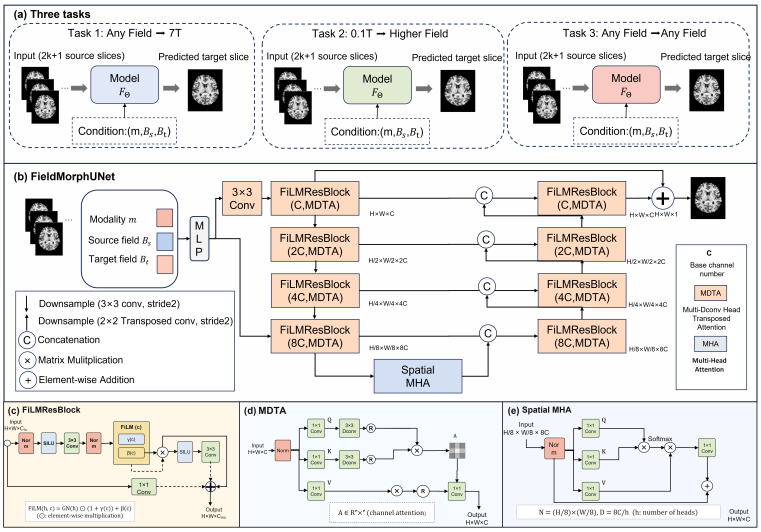

Fig. 2. FieldMorphUNet architecture. (a) Three challenge tasks share one model F_θ conditioned on modality and source/target fields. (b) 2.5D FiLM residual U-Net with MDTA at lower resolutions and spatial MHA at the bottleneck. (c) FiLM residual block. (d) MDTA channel attention. (e) Spatial multi-head attention.

3.3 Paired and Unpaired Training with Residual Refinement

Training alternates between paired prospective translation and unpaired retrospective reconstruction. The paired stream samples (m, s, t, u, z) and supervises the normalized target slice for an ordered field transformation. The retrospective stream corrupts a real 2.5D context and reconstructs its clean center slice under the self-domain condition (m, f, f). This stream regularizes anatomical reconstruction and robustness to degraded inputs without requiring cross-domain

retrospective pairs. The stage-one objective is

$$\mathcal{L}_{\text{stage1}} = \lambda_1 \mathcal{L}_1 + \lambda_s \mathcal{L}_{\text{SSIM}} + \lambda_p \mathcal{L}_{\text{LPIPS}} + \lambda_g \mathcal{L}_{\nabla}, \tag{5}$$

where $\mathcal{L}_{\nabla}$ is the image-gradient loss. All terms are computed in normalized space with $(\lambda_1, \lambda_s, \lambda_p, \lambda_g) = (1.0, 1.0, 0.5, 0.15)$. The final model does not use adversarial training.

A compact RefineUNet further corrects the coarse prediction $\hat{y}_c$. It receives $\hat{y}_c$ together with the five-slice primary-modality context $S_z^{m,s} = [\tilde{I}_{z-2}^{m,s}, \ldots, \tilde{I}_{z+2}^{m,s}]$ and is conditioned on the same triplet (m, s, t). The refiner predicts a residual Δ and a spatial gate $g \in [0, 1]$:

$$\hat{y}_f = \text{clip}(\hat{y}_c + \alpha \tanh(\Delta), 0, 1), \qquad \hat{y}_r = \text{clip}\big(\hat{y}_c + g \odot (\hat{y}_f - \hat{y}_c), 0, 1\big), \tag{6}$$

where α=0.1 bounds the correction amplitude. Refinement is optimized in stages: an initial residual model is trained with the generator frozen, after which the full five-slice context and adaptive gate are introduced while the generator decoder is jointly fine-tuned. A final perceptual fine-tuning stage increases LPIPS supervision while retaining structural, gradient, and coarse-preservation objectives.

At inference, the identity and anterior–posterior flip paths are mapped to the same orientation and averaged:

$$\hat{y} = \frac{1}{2}\left[\hat{y}_r(X) + \mathcal{F}_{\text{AP}}^{-1}\big(\hat{y}_r(\mathcal{F}_{\text{AP}}(X))\big)\right]. \tag{7}$$

The refined output is used directly (equivalently, β=1) without an additional classical post-filter, followed by pro calibration.

3.4 Prospective Population Target-Scale Calibration

Normalized inference produces a structural prediction in $[0, 1]$, which must be rescaled to recover the absolute target intensity. The conventional source-ratio estimate $s^{\text{ratio}} = r_s R_{m,s\to t}$ depends on a median ratio computed from a small paired cohort. As illustrated in Fig. 3a, the subject-level ratios $r(I_u^{m,t})/r(I_u^{m,s})$ can vary substantially, making the estimated target scale sensitive to the source intensity.

We instead use a prospective population target-scale prior (pro prior). Let $\mathcal{U}_{\text{pro}}^{\text{train}}$ denote the prospective training cohort and define

$$\mathcal{R}_{\text{pro}}^{m,t} = \left\{r(I_u^{m,t}) : u \in \mathcal{U}_{\text{pro}}^{\text{train}}\right\}, \qquad \pi_{\text{pro}}^{m,t} = \text{median}\big(\mathcal{R}_{\text{pro}}^{m,t}\big). \tag{8}$$

The prior is computed once for each modality–target-field pair and provides a fixed target-domain scale. For the calibration study, we also define $\pi_{\text{retro}}^{m,t}$ as the median brain-$p_{99.5}$ of samples from the corresponding retrospective training domain. The final prediction uses the pro prior:

$$s_{m,t} = \pi_{\text{pro}}^{m,t}, \qquad \hat{I}^{m,t} = \text{clip}(\hat{y}\, s_{m,t}, 0, 1). \tag{9}$$

This formulation separates normalized structural synthesis from absolute target-intensity recovery.

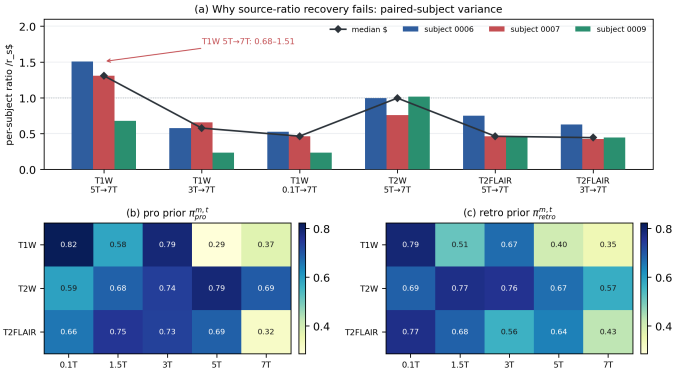

Fig. 3. Calibration evidence. (a) Per-subject ratios $r(I_u^{m,t})/r(I_u^{m,s})$; diamonds mark the classical median $R_{m,s\to t}$ (T1W 5T→ 7T spans 0.68–1.51). (b)–(c) Population target priors $\pi_{\text{pro}}^{m,t}$ and $\pi_{\text{retro}}^{m,t}$ (brain-$p_{99.5}$ in max-normalized scoring space).

4 Experiments

4.1 Protocol and Implementation

We used the challenge-provided prospective paired multi-field volumes for supervised translation and retrospective unpaired volumes for self-domain reconstruction. Image fidelity was evaluated using three-dimensional normalized Root Mean Squared Error (nRMSE, $\|\hat{I} - I\|_2/\|I\|_2$), mean axial Structural Similarity (SSIM), and slice-wise Learned Perceptual Image Patch Similarity (LPIPS). Tasks 1–2 additionally reported adjusted Dice and volume-consistency scores. Checkpoints were monitored using a modality-balanced set of translation directions.

The stage-one generator used five neighboring slices from each of the three source contrasts (k=2, 15 input channels), a base width of 96, channel multipliers $(1, 2, 4, 8)$, and a condition dimension of 256. It was trained for 120k iterations using AdamW [5], a batch size of 6, an initial learning rate of 2×10^{-4}, a retrospective batch ratio of 0.3, and an auxiliary-contrast dropout probability of 0.1. We used exponential moving average weights, mixed-precision training, and gradient clipping.

The RefineUNet used a base width of 48 and a residual bound of α=0.1. Its initial, decoder-joint, and perceptual refinement stages were optimized for 4k, 8k, and 12k iterations, with learning rates of 2×10^{-5}, 1×10^{-5}, and 3×10^{-6}, respectively. The first stage used a batch size of 4, and the two joint stages used a batch size of 1. Reported results used anterior–posterior flip test-time augmentation and pro calibration.

4.2 Main Results

Table 1. Challenge validation results on Tasks 1–3, reported as adjusted means over subtasks. Dice and Volume apply only to Tasks 1–2.

Task	Setting	Method	SSIM↑	LPIPS↓	nRMSE↓	Dice↑	Volume↑
1	Any→7T	Baseline	0.905	0.075	0.304	0.854	0.860
		FieldMorphUNet	**0.929**	**0.060**	**0.259**	**0.887**	**0.885**
2	0.1T→Higher	Baseline	0.866	0.102	0.218	0.766	0.779
		FieldMorphUNet	**0.903**	**0.081**	**0.192**	**0.879**	**0.888**
3	Any→Any	Baseline	0.684	0.217	0.502	—	—
		FieldMorphUNet	**0.916**	**0.069**	**0.216**	—	—

FieldMorphUNet improved all reported metrics over the challenge baseline (Table 1). On Task 1, SSIM increased from 0.905 to 0.929 and LPIPS decreased from 0.075 to 0.060, while nRMSE decreased from 0.304 to 0.259. Task 2 showed improvements in both image fidelity and segmentation consistency, including Dice from 0.766 to 0.879 and Volume from 0.779 to 0.888. The largest margin was observed on Task 3: SSIM increased from 0.684 to 0.916, while LPIPS and nRMSE decreased from 0.217 to 0.069 and from 0.502 to 0.216, respectively.

Figure 4 presents representative translations from 0.1T to higher target fields across T1W, T2W, and T2FLAIR. The predictions preserved the major anatomical structures while reproducing the target-field contrast across the evaluated modalities.

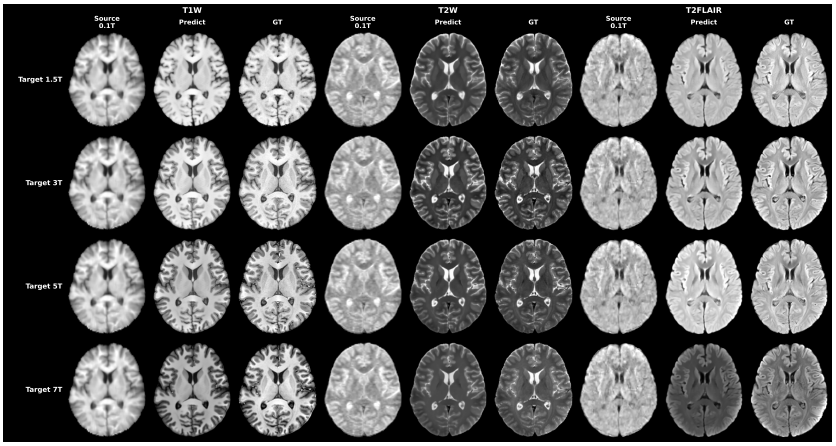

Fig. 4. Qualitative multi-modality translations from source 0.1T to target fields 1.5T, 3T, 5T, and 7T. Each group shows the source, prediction, and target images for T1W, T2W, and T2FLAIR.

4.3 Controlled Baseline Comparisons

We further evaluated two controlled baselines on Task 3. The plain conditional U-Net uses a standard encoder–decoder with spatially broadcast one-hot condi-

tioning and excludes FiLM, attention, residual refinement, and population calibration. A conditional flow-matching baseline [4] provides an alternative generative formulation. Table 2 reports their adjusted validation metrics together with the final FieldMorphUNet.

Table 2. Controlled comparisons on the Task 3 challenge validation set.

Method	SSIM↑	LPIPS↓	nRMSE↓
Plain conditional U-Net	0.8974	0.0860	0.2712
Conditional flow matching	0.8688	0.1450	0.2907
FieldMorphUNet	**0.9160**	**0.0693**	**0.2160**

FieldMorphUNet achieved higher SSIM and lower LPIPS and nRMSE than both controlled baselines. The plain conditional U-Net remained closer to the final model than conditional flow matching, but the complete FieldMorphUNet formulation improved all three metrics.

4.4 Scale-Recovery Analysis

To isolate the effect of absolute scale recovery, we fixed the generator and refiner and changed only the scalar $s_{m,t}$. In each leave-one-subject-out fold, the scale statistics were recomputed from the remaining prospective subjects. Table 3 compares source-ratio recovery, retrospective and prospective population priors, and a per-volume oracle scale.

Table 3. Leave-one-subject-out scale-recovery analysis with a fixed generator and refiner. Bold denotes the best deployable result; the oracle is a closed-form per-volume reference bound computed using the target image.

Scale recovery	nRMSE↓	SSIM↑	LPIPS↓
Source ratio $r_s R_{m,s\to t}$	0.192	0.934	0.052
Retro prior $\pi^{m,t}_{\text{retro}}$	0.183	0.935	0.050
Pro prior $\pi^{m,t}_{\text{pro}}$	**0.154**	**0.939**	**0.049**
Per-volume oracle scale	0.076	0.946	0.044

Among the deployable strategies, the pro prior achieved the best overall metric vector. Compared with source-ratio recovery, it reduced nRMSE from 0.192 to 0.154, increased SSIM from 0.934 to 0.939, and reduced LPIPS from 0.052 to 0.049. The retrospective prior provided smaller improvements, whereas the per-volume oracle reached an nRMSE of 0.076. The remaining gap to the oracle reflects subject-specific intensity variation that cannot be fully represented by a single population-level target scale.

5 Conclusion

In this paper, we presented FieldMorphUNet, a conditional framework for multi-field MRI translation. It separates normalized structural synthesis from absolute

intensity recovery using a prospective population target-scale prior, while integrating multi-contrast 2.5D context, conditional feature modulation, attention, and gated residual refinement. Results across the three MRIxFields tasks support its applicability to multi-field translation, and the leave-one-subject-out analysis underscores the importance of explicit scale recovery.

The method remains limited by its global population prior, which cannot fully capture subject-specific or spatially varying intensity, and by its 2.5D design, which does not enforce full three-dimensional consistency. Evaluation is also restricted to the challenge data and field strengths. Future work will explore spatially adaptive calibration, volumetric modeling, and generalization to unseen field strengths and clinical abnormalities.

Acknowledgments. The authors thank the MRIxFields2026 organizers and data contributors. This work was supported in part by the National Natural Science Foundation of China (Grant No. 82502508), the Explorer Program of Shanghai (Grant No. 26TS1403100), the Central Government Guides Local Science and Technology Development Fund Projects (Grant No. 25ZYJA017), and the Youth Science and Technology Talent Innovation Project of Lanzhou City (Grant No. 2025-QN-037).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Fu, Y., Dong, S., Huang, Y., Niu, M., Ni, C., Yu, L., Shi, K., Yao, Z., Zhuo, C.: MPGAN: Multi Pareto generative adversarial network for the denoising and quantitative analysis of low-dose PET images of human brain. Medical Image Analysis **98**, 103306 (2024). https://doi.org/10.1016/j.media.2024.103306
2. Fu, Y., Dong, S., Niu, M., Xue, L., Guo, H., Huang, Y., Xu, Y., Yu, T., Shi, K., Yang, Q., Shi, Y., Zhang, H., Tian, M., Zhuo, C.: AIGAN: Attention-encoding integrated generative adversarial network for the reconstruction of low-dose CT and low-dose PET images. Medical Image Analysis **86**, 102787 (2023). https://doi.org/10.1016/j.media.2023.102787
3. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5967–5976 (2017). https://doi.org/10.1109/CVPR.2017.632
4. Lipman, Y., Chen, R.T.Q., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. In: International Conference on Learning Representations (2023), `https://openreview.net/forum?id=PqvMRDCJT9t`
5. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (ICLR) (2019)
6. Mirza, M., Osindero, S.: Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784 (2014)
7. Nyúl, L.G., Udupa, J.K.: On standardizing the MR image intensity scale. Magnetic Resonance in Medicine **42**(6), 1072–1081 (1999). https://doi.org/10.1002/(SICI)1522-2594(199912)42:6<1072::AID-MRM11>3.0.CO;2-M
8. Perez, E., Strub, F., de Vries, H., Dumoulin, V., Courville, A.: FiLM: Visual reasoning with a general conditioning layer. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 32 (2018). https://doi.org/10.1609/aaai.v32i1.11671
9. Qian, F., Wang, W., Huang, Y., Niu, M., Gao, Y., Zhao, Z., Shi, K., Yu, L., Fu, Y., Zhuo, C.: FADFNet: A fine-tunable and adaptive decomposition-fusion network for cross-dataset low-dose CT and low-dose PET image reconstruction. Medical Image Analysis **111**, 104016 (2026). https://doi.org/10.1016/j.media.2026.104016
10. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
11. Shinohara, R.T., Sweeney, E.M., Goldsmith, J., Shiee, N., Mateen, F.J., Calabresi, P.A., Jarso, S., Pham, D.L., Reich, D.S., Crainiceanu, C.M.: Statistical normalization techniques for magnetic resonance imaging. NeuroImage: Clinical **6**, 9–19 (2014). https://doi.org/10.1016/j.nicl.2014.08.008
12. 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). https://doi.org/10.1109/TIP.2003.819861
13. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H.: Restormer: Efficient transformer for high-resolution image restoration. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5718–5729 (2022). https://doi.org/10.1109/CVPR52688.2022.00564
14. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: IEEE/CVF Confer-

ence on Computer Vision and Pattern Recognition (CVPR). pp. 586–595 (2018). https://doi.org/10.1109/CVPR.2018.00068
15. Zhao, Z., Fu, Y.: FA^2-Net: A frequency-aware asymmetric dual-stage network for unpaired cross-sequence MRI synthesis. In: Pattern Recognition and Computer Vision (PRCV). pp. 206–219. LNCS, Springer, Singapore (2026). https://doi.org/10.1007/978-981-95-5631-1_15