Hallucination-safe deep enhancement of low-field MRI using matched high-field references

Abstract. Low-field magnetic resonance imaging (MRI) has the potential to increase access to medical imaging in facilities where resources are limited due to its affordability and portability in comparison with conventional high-field systems. However, low-field MRI faces challenges such as lower signal-to-noise ratio (SNR), lower resolution, lower contrast, and image artifacts related to acquisition which could affect the diagnostic confidence. Deep learning is a potential solution to improve low-field imaging, but the conventional reconstruction approaches can lead to distorted anatomical details, which are known as hallucinations. Such unreal features do not exist in the standard image and, therefore, may create problems in reconstruction of medical images. In this research, we demonstrate the possibility to avoid hallucinations through the proposed process of deep learning applied to low-field MRI. This work supports the development of trustworthy, anatomically reliable, and safer enhancement methods for low-field MRI in resource-constrained medical environments, while addressing key limitations related to hallucinations, anatomical fidelity, evaluation robustness, and clinical translation.

Keywords: Low-field MRI; Deep learning; MRI enhancement; Hallucination-safe AI; Anatomical fidelity; Image reconstruction; Super-resolution; Artificial intelligence in medical imaging.

1 Introduction

MRI makes it possible to reach neurological diagnoses without the application of ionizing radiation; however, the regular application of either 1.5T or 3T MRI systems is still associated with high costs and depends on infrastructure availability. Low frequency MRI methods represent a more affordable and portable alternative to regular systems that can help boost availability across developing regions, rural areas, and other underprivileged territories [1]. Although LF MRI has made significant progress, it still suffers from challenges like low signal-to-noise ratio, low resolution, low contrast, noise, and artifacts. Unfortunately, the above-mentioned difficulties worsen with decreasing field strength leading to a decrease in the level of detail obtainable from the pictures and its potential in diagnostic examinations [2]. Deep learning can help to enhance the quality of LF MRI by eliminating noise, improving resolution, and bringing back the anatomy. The ultimate aim is to obtain a high-quality picture from low-quality images [3].

However, image enhancement does not assure anatomical correctness. The models may yield visually believable hallucinations that are anatomically inaccurate and not

identical to the prior scan, which may pose a problem in diagnosis as well as clinical decision-making. This indicates the necessity for the creation of hallucination-safe LF MRI enhancement schemes that will utilize paired high-field references [4].
This work proposes a hallucination-safe framework for LF MRI enhancement using paired high-field references. The approach aims to improve image quality, SNR, and resolution while preserving anatomical consistency and reducing hallucination risk.

2 Previous Work

Deep learning has advanced LF MRI enhancement through reconstruction-based, super-resolution, and image-to-image translation methods. Early work addressed reconstruction from noisy LF k-space data. Koonjoo et al. [4] proposed AUTOMAP, an end-to-end model that directly maps k-space data to image space, which improved SNR and reduced artifacts, but its reliance on representative training data may limit robustness to unseen anatomy. Super-resolution methods addressed the limited spatial resolution of LF MRI. De Leeuw den Bouter et al. [3] proposed SRDenseNet to proposed SRDenseNet to enhance LF MRI resolution and brain image quality, but its use of simulated LF data may limit clinical generalization. Image-to-image translation methods also improved LF MRI enhancement. Islam et al. [5] proposed Lo-HiResGAN to generate synthetic high-field images from 64 mT scans, but GANs may introduce realistic yet anatomically inaccurate details [6]. Attention-based methods further improved LF MRI enhancement. Lau et al. [7] used attention-guided 3D super-resolution to preserve anatomical detail, but training on synthetic high-field degradation may limit real-world robustness.

3 Methodology

3.1 Dataset Description

We used the latest available version of the open paired LF/HF brain MRI dataset from Zenodo [8]. The 64 mT LF images were used as inputs, while matched 3 T images served as supervised targets for LF-to-HF enhancement. Figure 1 shows representative paired 64 mT and 3 T brain MRI acquisitions from one subject.

3.2 Proposed Architecture

We aim to enhance LF MRI while minimizing hallucination risk. To this end, we propose HS-ResRRDBNet-IA, a hallucination-safe reconstruction network based on a Residual U-Net backbone. HS-ResRRDBNet-IA combines Residual-in-Residual Dense Blocks (RRDBs), Squeeze-and-Excitation (SE) attention, hallucination-aware frequency-gated skip fusion (HAFG), input-anchored output fusion, and uncertainty estimation to improve contrast and structural visibility while preserving input-supported anatomy. As shown in Figure 2, the model follows a Residual U-Net encoder–decoder design. The encoder extracts multi-scale features from 64 mT inputs, while residual connections improve training stability and preserve anatomical

information. RRDBs refine bottleneck features, and SE attention emphasizes informative anatomical responses while reducing noise-related features.

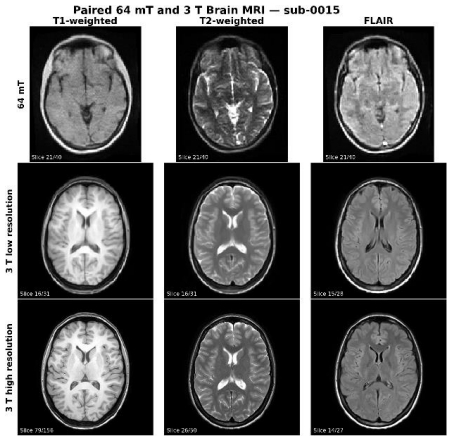

Fig. 1. Representative paired 64 mT and 3 T brain MRI acquisitions for subject

To improve skip reliability, HAFG separates low-frequency structural cues from high-frequency details and adaptively fuses them with decoder features. This preserves stable anatomy while limiting unsupported high-frequency textures that may cause hallucination-like artifacts. The decoder reconstructs the enhanced image through upsampling and convolutional refinement. An input-anchored fusion mechanism then constrains deviations from the 64 mT anatomy using a spatial trust gate, improving contrast and denoising while limiting unsupported anatomical changes. MC-Dropout estimates reconstruction uncertainty at inference. Multiple stochastic predictions are averaged to produce the final image, while voxel-wise variance generates uncertainty maps highlighting low-confidence regions, providing an additional safety mechanism. Overall, HS-ResRRDBNet-IA improves reconstruction quality while promoting hallucination-safe enhancement through frequency-aware skip fusion, input-constrained prediction, mask-guided supervision, and uncertainty-aware inference.

3.3 Hallucination-Safe Reconstruction Strategy

To promote hallucination-safe enhancement, the proposed framework constrains reconstruction using anatomical, structural, frequency, and uncertainty-aware mechanisms. Instead of freely generating an enhanced image, the model is guided to preserve information supported by the 64 mT input while improving contrast, resolution, and noise suppression. Anatomical consistency is encouraged through an image-gradient matching loss computed between the enhanced image and the paired 3 T reference, where gradients refer to image-intensity spatial gradients. This constraint helps preserve anatomical boundaries and fine structural details while reducing excessive smoothing [9]. The model is trained using a composite objective that combines intensity fidelity, structural similarity, edge preservation, frequency consistency, and anatomical gradient matching:

$$L = \lambda_1 L_{L1} + \lambda_2 L_{SSIM} + \lambda_3 L_{edge} + \lambda_4 L_{freq} + \lambda_5 L_{anat}$$

where λ_1, λ_2, λ_3, λ_4 and λ_5 are weighting coefficients controlling the contribution of each loss component.
Here, each term controls a complementary aspect of reconstruction quality and anatomical reliability. During inference, MC-Dropout is used to estimate uncertainty through multiple stochastic forward passes. The mean prediction is used as the final enhanced image, while voxel-wise variance generates uncertainty maps that highlight low-confidence regions. These maps provide an additional safety indicator for identifying less reliable reconstructions.

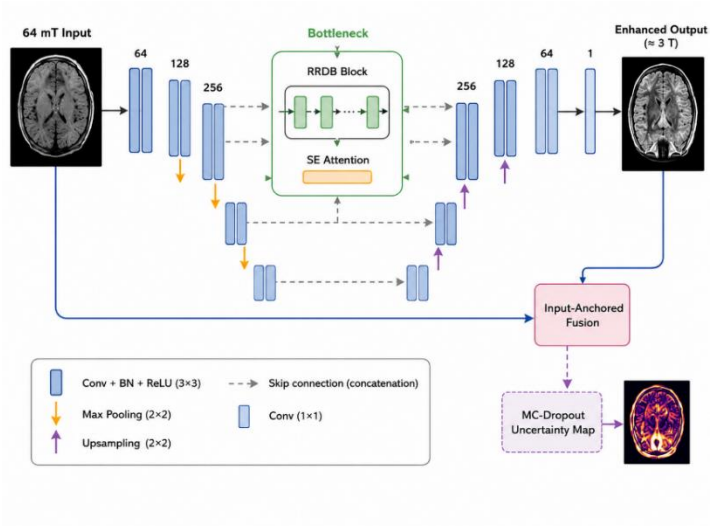

Fig. 2. Architecture of the proposed HS-ResRRDBNet-IA model

4 Experimental Protocol

The model was implemented in PyTorch and trained using AdamW with an initial learning rate and a weight decay of 10^{-4}. Training used a batch size of 8, 256×256 images normalized to [0,1], early stopping, ReduceLROnPlateau scheduling, and exponential moving average weights for validation and checkpoint selection. Model selection was based on validation MAE.
All experiments used paired 64 mT T1-weighted images as inputs and matched low-resolution 3 T T1-weighted images as supervised references. The 3 T images were rigidly registered and resampled onto the native 64 mT grid. No atlas-based normalization or non-linear inter-subject registration was applied to avoid additional interpolation, smoothing, or training-time blurring.
All preprocessing was performed in the 64 mT image space. Non-finite values were removed, invalid regions were set to zero, and robust intensity normalization was applied. Histogram standardization was fitted within each cross-validation fold using only training participants and then frozen for validation and test data to prevent leakage. Mask-guided supervision restricted learning and evaluation to valid anatomical regions,

while background masks enabled leakage monitoring. Reported slice counts represent derived 2D samples from 3D volumes, not independent subjects or scans.

Data augmentation was applied only to the training set. Spatial transformations, including left-right flipping and mild affine perturbations, were applied synchronously to the 64 mT input, 3 T target, and masks. Intensity augmentations were applied only to the 64 mT inputs and included mild gain, bias, gamma changes, Gaussian noise, bias-field modulation, and slight blurring. Aggressive transformations such as vertical flipping, elastic deformation, and atlas normalization were avoided to preserve LF/HF anatomical correspondence [10].

Five-fold cross-validation was performed at the participant level. Each fold included 6 training, 2 validation, and 2 test participants, with all sessions and derived slices from the same participant kept in the same subset to avoid leakage. The 64 mT images were used as network inputs, while the registered 3 T images were used only as supervised targets and evaluation references. No separate all-low-field test set or external quantitative test cohort was used.

Performance was evaluated using MAE, RMSE, NRMSE, PSNR, and SSIM. Anatomical fidelity and hallucination safety were further assessed using HFEN, edge coherence, texture variance error, frequency hallucination score, background leakage, and hallucination risk [11].

An ablation study compared five configurations: ResUNet, ResUNet-RRDB, ResUNet-RRDB-SE, a frequency-gated skip fusion variant, and the final HS-ResRRDBNet-IA. All variants used the same preprocessing, augmentation, optimization, and participant-level cross-validation protocol. The final model was selected because it provided the best balance between reconstruction quality, anatomical consistency, and hallucination control.

5 Results

5.1 Ablation Study

An ablation study evaluated the contribution of each component of the proposed framework. Five models were compared: ResUNet (M0), ResUNet-RRDB (M1), ResUNet-RRDB-SE (M2), ResUNet-RRDB-SE-HAFG-IA (M3), and the final HS-ResRRDBNet-IA (M4) with mask-guided supervision and hallucination-safe losses. All models were evaluated using the same participant-level five-fold cross-validation protocol, with 6 training, 2 validation, and 2 test participants per fold. All 2D slices from the same participant remained in the same subset to prevent data leakage. Table 1 reports the ablation study results after 100 training epochs.

The baseline ResUNet achieved the highest conventional reconstruction metrics, with a PSNR of 18.53 dB and an SSIM of 0.6210. The final HS-ResRRDBNet-IA obtained comparable performance, with a PSNR of 18.49 dB and an SSIM of 0.6106, while maintaining a low Hallucination Score of 0.1522. Although the lowest Hallucination Score was observed for the RRDB-based variants, all configurations remained within a narrow range of 0.1521–0.1556.

Therefore, HS-ResRRDBNet-IA was selected not because it maximized PSNR or SSIM, but because it provided a safety-oriented trade-off between reconstruction

quality, anatomical consistency, hallucination control, and computational complexity. With 4.321 M trainable parameters, the final model introduced hallucination-aware skip fusion, input-anchored output fusion, mask-guided supervision, and hallucination-safe losses without substantially increasing model size.

Table 1. Results of the ablation study after 100 training epochs.

Model	Params (M)	PSNR ↑	SSIM ↑	Hallucination Score ↓
M0	4.690	18.53	0.6210	0.1542
M1	3.789	18.49	0.6124	0.1536
M2	3.806	18.41	0.6088	0.1521
M3	4.321	18.47	0.6130	0.1556
M4	4.321	18.49	0.6106	0.1522

5.2 Final Reconstruction Performance

Following the ablation study, HS-ResRRDBNet-IA was selected for the final reconstruction experiments because it provided the best balance between reconstruction quality, anatomical preservation, and hallucination control. The final model combines RRDB refinement, SE attention, hallucination-aware frequency-gated skip fusion, input-anchored output fusion, mask-guided supervision, hallucination-safe losses, and uncertainty estimation.

The model was trained and evaluated using the same participant-level five-fold cross-validation protocol. Model selection was based on validation MAE, while test performance was assessed using both conventional reconstruction metrics and hallucination-safety measures. HS-ResRRDBNet-IA achieved an MAE of 0.0902, PSNR of 18.49 dB, SSIM of 0.6106, and Hallucination Score of 0.1522. Compared with the baseline ResUNet, which achieved an MAE of 0.0893, PSNR of 18.53 dB, SSIM of 0.6210, and Hallucination Score of 0.1542, the proposed model maintained similar reconstruction accuracy while slightly reducing hallucination risk.

These results indicate that HS-ResRRDBNet-IA does not simply optimize pixel-level similarity. Instead, it promotes a safer reconstruction strategy by preserving anatomy supported by the 64 mT input and limiting unsupported high-frequency content and background leakage. The input-anchored fusion constrains the enhanced output toward the original low-field anatomy, while mask-guided supervision and hallucination-safe losses discourage anatomically inconsistent structures.

The enhanced images presented qualitative improvements regarding contrast and noise reduction when compared to the 64 mT inputs but maintaining the structure of the main anatomical elements. Furthermore, uncertainty maps revealed areas with low confidence near the boundary of the tissue and low-signal regions in general; this means there is little confidence in the interpretation of enhanced images. Figure 3 shows a representative example including the 64 mT input, registered 3 T reference, enhanced reconstruction, absolute error map, and uncertainty map.

The HS-ResRRDBNet-IA provided Representative reconstruction. It enhanced contrast and visibility of structures while maintaining anatomical details with the support of 64 mT input. The uncertainty map indicates areas where prediction confidence is low.

Additional hallucination-specific measurements, including FHS, TVE, Edge Coherence, Background Leakage, and Hallucination risk, confirmed the model's ability

to improve image quality while maintaining anatomical integrity. Edge Coherence was almost zero, indicating that anatomical borders were retained without significant sharpening. In conclusion, HS-ResRRDBNet-IA managed to provide a good hallucination-safety trade-off, but more validation on bigger datasets and pathological cases is needed before clinical use.

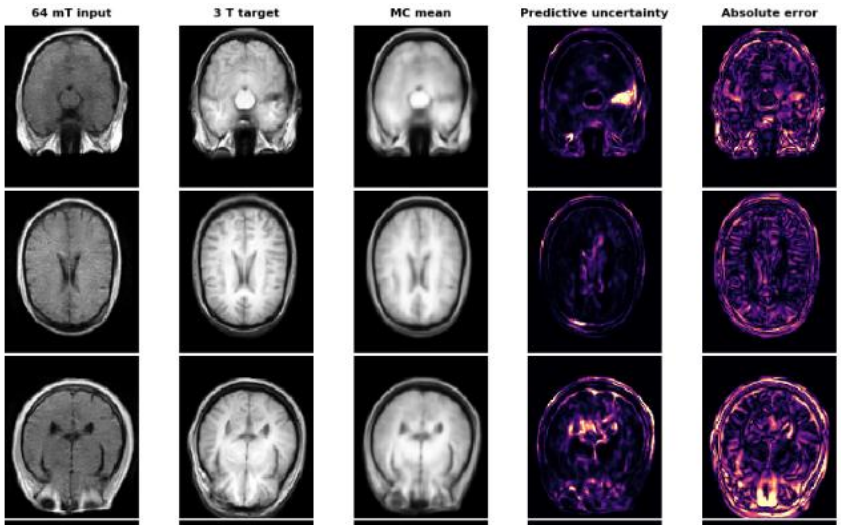

Fig. 3. Enhanced MRI reconstruction and uncertainty map

Table 2 compares the proposed model with representative state-of-the-art LF MRI enhancement methods. Because these methods were evaluated on different datasets and experimental protocols, the comparison should be interpreted as contextual rather than strictly direct.

Table 2. Comparison with representative state-of-the-art LF MRI enhancement methods

Method	PSNR (dB)	SSIM	RMSE	Hallucination score
AUTOMAP [4]	21.72	0.716	0.232	-
SRDenseNet [3]	35.39	0.9506	-	-
LoHiResGAN [5]	33.842	0.959	0.104	-
HS-ResRRDBNet-IA	18.49	0.6106	0.1229	0.1522

6 Discussion

This study introduced HS-ResRRDBNet-IA, a hallucination-safe framework for LF MRI enhancement using paired 64 mT and 3 T brain MRI data. The model combines RRDB refinement, SE attention, frequency-aware skip fusion, input-anchored prediction, mask-guided supervision, hallucination-safe losses, and uncertainty estimation to improve image quality while preserving anatomical fidelity.

HS-ResRRDBNet-IA achieved reconstruction performance close to the baseline ResUNet, with slightly lower PSNR and SSIM but a better hallucination-safety trade-off. This suggests that the model is not optimized only for pixel-level similarity, but for safer reconstruction. The ablation study showed that RRDB blocks improved feature refinement, SE attention supported anatomical feature selection, and input-anchored fusion constrained the output toward structures supported by the 64 mT input.

Qualitative results showed improved contrast and reduced noise, while uncertainty maps identified low-confidence regions near tissue boundaries and weak-signal areas. These maps provide useful guidance for cautious interpretation of enhanced images.

The study remains limited by the small healthy cohort, repeated sessions, 2D slice-based training, and possible geometric mismatch between 64 mT and 3 T images. Comparisons with previous methods are contextual because datasets and protocols differ. Future work should include larger pathological datasets, 3D models, external validation, and radiologist assessment before clinical deployment.

7 Conclusion

This study introduced HS-ResRRDBNet-IA, a hallucination-safe deep learning framework for low-field MRI enhancement. The model combines a Residual U-Net backbone, RRDB blocks, SE attention, anatomical and frequency constraints, input-anchored prediction, and uncertainty estimation to improve image quality while preserving anatomical fidelity.

Experimental results on paired 64 mT and 3 T brain MRI images showed that the proposed model achieved reliable reconstruction performance while limiting hallucination risk. The ablation study indicated that RRDB blocks improved high-frequency feature refinement, SE attention supported anatomical feature selection, and MC-Dropout provided useful uncertainty information for identifying less reliable regions.

This work remains a proof of concept due to the limited healthy cohort and the use of 2D slices. Future studies should include larger and more diverse populations, pathological MRI data, 3D reconstruction, external validation, and radiologist assessment.

Overall, HS-ResRRDBNet-IA provides a promising direction for safer low-field MRI enhancement. Its potential impact is particularly relevant in resource-constrained settings, where reliable enhancement of portable low-field MRI could improve imaging accessibility while supporting more trustworthy clinical interpretation.

References

[1] T. C. Arnold, C. W. Freeman, B. Litt, and J. M. Stein, 'Low-field MRI: Clinical promise and challenges', *J. Magn. Reson. Imaging JMRI*, vol. 57, no. 1, pp. 25–44, Jan. 2023, doi: 10.1002/jmri.28408.

[2] M. Xia *et al.*, 'On Hallucinations in Artificial Intelligence–Generated Content for Nuclear Medicine Imaging (the DREAM Report)', *J. Nucl. Med.*, vol. 67, no. 2, pp. 166–174, Feb. 2026, doi: 10.2967/jnumed.125.270653.

[3] M. L. de Leeuw den Bouter, G. Ippolito, T. P. A. O'Reilly, R. F. Remis, M. B. van Gijzen, and A. G. Webb, 'Deep learning-based single image super-resolution for low-field MR brain images', *Sci. Rep.*, vol. 12, no. 1, p. 6362, Apr. 2022, doi: 10.1038/s41598-022-10298-6.

[4] N. Koonjoo, B. Zhu, G. C. Bagnall, D. Bhutto, and M. S. Rosen, 'Boosting the signal-to-noise of low-field MRI with deep learning image reconstruction', *Sci. Rep.*, vol. 11, no. 1, p. 8248, Apr. 2021, doi: 10.1038/s41598-021-87482-7.

[5] K. T. Islam *et al.*, 'Improving portable low-field MRI image quality through image-to-image translation using paired low- and high-field images', *Sci. Rep.*, vol. 13, no. 1, p. 21183, Dec. 2023, doi: 10.1038/s41598-023-48438-1.

[6] S. Kim, H. F. J. Tregidgo, M. Figini, C. Jin, S. Joshi, and D. C. Alexander, 'Tackling Hallucination from Conditional Models for Medical Image Reconstruction with DynamicDPS', arXiv.org. Accessed: Jul. 07, 2026. [Online]. Available: https://arxiv.org/abs/2503.01075v2

[7] V. Lau *et al.*, 'Pushing the limits of low-cost ultra-low-field MRI by dual-acquisition deep learning 3D superresolution', *Magn. Reson. Med.*, vol. 90, no. 2, pp. 400–416, Aug. 2023, doi: 10.1002/mrm.29642.

[8] van den Broek, R., Lena, B., & webb, . andrew . (2025). Paired 64mT and 3T Brain MRI Scans of Healthy Subjects for Neuroimaging Research [Data set]. Zenodo. https://doi.org/10.5281/zenodo.15374450

[9] S. Kim, H. F. J. Tregidgo, C. Jin, M. Figini, and D. C. Alexander, 'HalluGen: Synthesizing Realistic and Controllable Hallucinations for Evaluating Image Restoration', Dec. 03, 2025, *arXiv*: arXiv:2512.03345. doi: 10.48550/arXiv.2512.03345.

[10] J. Lee, B. Kim, and H. Park, 'MC2 -Net: motion correction network for multi-contrast brain MRI', *Magn. Reson. Med.*, vol. 86, no. 2, pp. 1077–1092, Aug. 2021, doi: 10.1002/mrm.28719.

[11] J. Yu, S. Mdletshe, H. Abbasi, E. Kwon, S. Holdsworth, and A. Wang, 'Deep Learning for Brain MRI Artifact Correction: Current Challenges and Future Directions', *Bioengineering*, vol. 13, no. 6, p. 699, Jun. 2026, doi: 10.3390/bioengineering13060699.