Multi-contrast MRI Reconstruction from Undersampled Data

Team course project at ShanghaiTech University

BME1312: Applications of Artificial Intelligence in Medical Imaging

Public web version with student IDs and emails removed

1. Introduction

In clinical Magnetic Resonance Imaging (MRI), reducing scan time is crucial for patient comfort and minimizing motion artifacts [1]. However, accelerating MRI scans by undersampling in k-space inevitably leads to severe aliasing artifacts in the reconstructed images. To recover high-quality, aliasing-free images while maintaining high acceleration factors, deep learning approaches have shown tremendous potential and superiority over traditional compressed sensing methods [3**?**]. This project aims to utilize the BraTS (Brain Tumor Segmentation) dataset to develop a robust, deep learning-based MRI reconstruction algorithm.

In the initial phase (Task 2), we implemented a baseline reconstruction network employing a standard U-Net architecture [4]. This model was trained using an L2 loss (Mean Squared Error) objective to minimize the pixel-wise discrepancy between the reconstructed image and the fully sampled ground truth. However, an in-depth analysis of the baseline results revealed that relying solely on L2 loss constraints inherently causes the network to generate over-smoothed images, resulting in the loss of critical high-frequency edges and fine texture details during reconstruction [5].

More importantly, we observed a significant limitation during evaluation: reconstructed images with a higher Peak Signal-to-Noise Ratio (PSNR) do not necessarily exhibit better image quality from a human visual perspective. To thoroughly investigate and validate this phenomenon, drawing inspiration from studies on image sharpness, we designed and conducted a "shifting pixels" experiment. The experimental results compellingly demonstrated that when an image undergoes a minor spatial shift, the PSNR drops drastically, yet the subjective image quality perceived by the human eye remains completely unchanged. This underscores the severe flaws of traditional pixel-wise metrics in reflecting true visual fidelity [6].

Based on these observations, we made the following core contributions in the advanced reconstruction phase of this project:

- **Integration of Wavelet Loss:** To address the over-smoothing issue in the U-Net baseline, we introduced a wavelet loss constraint in the frequency domain. This structural penalty enables the network to better focus on and recover the high-frequency details of the image [7].
- **Introduction of DISTS Metric and Perceptual Loss:** To tackle the misalignment between PSNR and human visual perception, we incorporated the Deep Image Structure and Texture Similarity (DISTS) metric into our evaluation framework, which explicitly focuses on texture and structural fidelity [**?**]. Concurrently, we integrated a perceptual loss into our training objectives [5]. This guides the network to approximate the ground truth at the feature level rather than merely the pixel level, significantly enhancing subjective visual quality.
- **Multi-modal Fusion and Unrolled Reconstruction:** To achieve clinical-grade quality, we developed a sophisticated multi-modal fusion architecture in Task 3. This design takes two modalities as input: an undersampled T2 image and a fully sampled T1 image, leveraging the anatomical structural prior from T1 to help recover T2 details [9]. Furthermore, we implemented an unrolled reconstruction network by cascading the denoising network with Data Consistency (DC) layers to strictly ensure k-space fidelity [**?**].

Ultimately, we seamlessly integrated the wavelet loss and perceptual loss into the training of the multi-modal unrolled network. Experimental results prove that our complete pipeline not only achieves excellent performance

[*] Equal Contribution.

in quantitative metrics but also makes significant breakthroughs in preserving tumor details and eliminating aliasing artifacts visually.

2. Method

2.1. Data Processing and Undersampling Simulation

To train our reconstruction networks, we first need to simulate the k-space undersampling process that occurs in accelerated clinical MRI scans. As illustrated in Figure 1, we start with fully sampled 2D image slices. We apply a 2D Fast Fourier Transform (FFT) to convert these images into the frequency domain (k-space). Subsequently, we generate a 2D random variable-density undersampling mask with a specific acceleration factor (e.g., $5\times$). This mask is applied to the fully sampled k-space data to simulate the acquisition process. Finally, an Inverse Fast Fourier Transform (IFFT) is applied to the undersampled k-space data to obtain the zero-filled, aliased images, which serve as the inputs to our deep learning models.

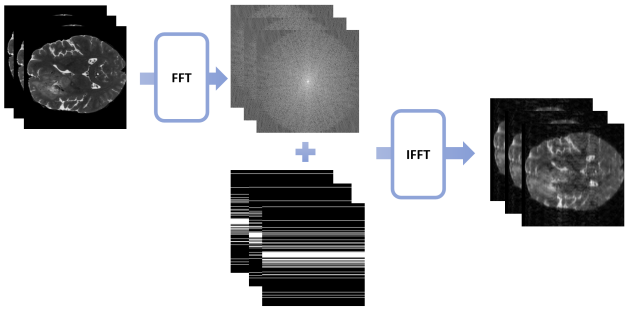


Figure 1. Pipeline of the data preprocessing and undersampling simulation

2.2. Baseline Reconstruction Network (Task 2)

In the initial phase of our project, we formulated the MRI reconstruction as an image-to-image translation problem and developed a baseline model based on the U-Net architecture, as shown in Figure 2. The network takes the blurred, undersampled MRI image as input and processes it through a symmetric encoder-decoder structure.

The encoder progressively downsamples the spatial resolution while increasing the channel depth using DoubleConv blocks (a sequence of 3×3 Convolution, Batch Normalization, and ReLU activation) followed by Max Pooling layers. A bottleneck layer captures the deepest latent representations. The decoder then progressively upsamples the feature maps using Transposed Convolutions and fuses them with high-resolution features from the encoder via skip connections. A final 1×1 convolution maps the deep features back to a single-channel reconstructed MRI image.

2.3. Multi-modal Fusion and Unrolled Reconstruction (Task 3)

To further elevate the reconstruction quality to a clinical grade, we advanced our architecture to a Multi-modal Unrolled Reconstruction Network (Figure 4). This advanced pipeline leverages the fully sampled T1 modality to guide the reconstruction of the highly undersampled T2 modality.

The architecture consists of three core components:

1. **Feature Fusion Module:** We utilize shallow CNN encoders to extract features from both the aliased T2 image and the fully sampled T1 image. A Cross-Attention mechanism is introduced where the T2 features act as the *Query*, and the T1 features act as the *Key* and *Value*. This allows the network to borrow structural priors from the T1 image to complement the missing information in T2.
2. **Unrolled Block:** The fused features are fed into a cascade of iterative unrolled blocks. Each block mimics an iteration of traditional optimization algorithms, containing a CNN Denoiser/Regularizer to refine the image features recursively.
3. **Data Consistency (DC) Layer:** To ensure fidelity to the originally acquired data, a DC layer is appended after each unrolled iteration. It transforms the intermediate prediction back to the k-space, replaces the predicted values at sampled positions with the actual measured k-space data, and transforms it back to the image domain.

2.4. Loss Function Design

Traditionally, MRI reconstruction models are optimized using solely the Mean Squared Error (MSE) loss. However, relying purely on pixel-wise loss functions often leads to over-smoothed results. To mitigate this and enhance both the structural fidelity and visual quality, we formulated a comprehensive hybrid loss function combining MSE, Wavelet, and Perceptual losses:

$$\begin{aligned}\mathcal{L}_{total} = &\lambda_{mse}\mathcal{L}_{MSE}(x, \hat{x}) \\ &+ \lambda_{wavelet}\mathcal{L}_{Wavelet}(x, \hat{x}) \\ &+ \lambda_{perceptual}\mathcal{L}_{LPIPS}(x, \hat{x})\end{aligned} \tag{1}$$

where x is the ground truth fully sampled image, $\hat{x}$ is the reconstructed image, and λ represents the balancing weights for each respective component.

Wavelet Loss ($\mathcal{L}_{Wavelet}$): We introduced a wavelet-domain loss to explicitly penalize the loss of high-frequency details. By decomposing the images into four frequency bands using the Discrete Wavelet Transform, we calculate

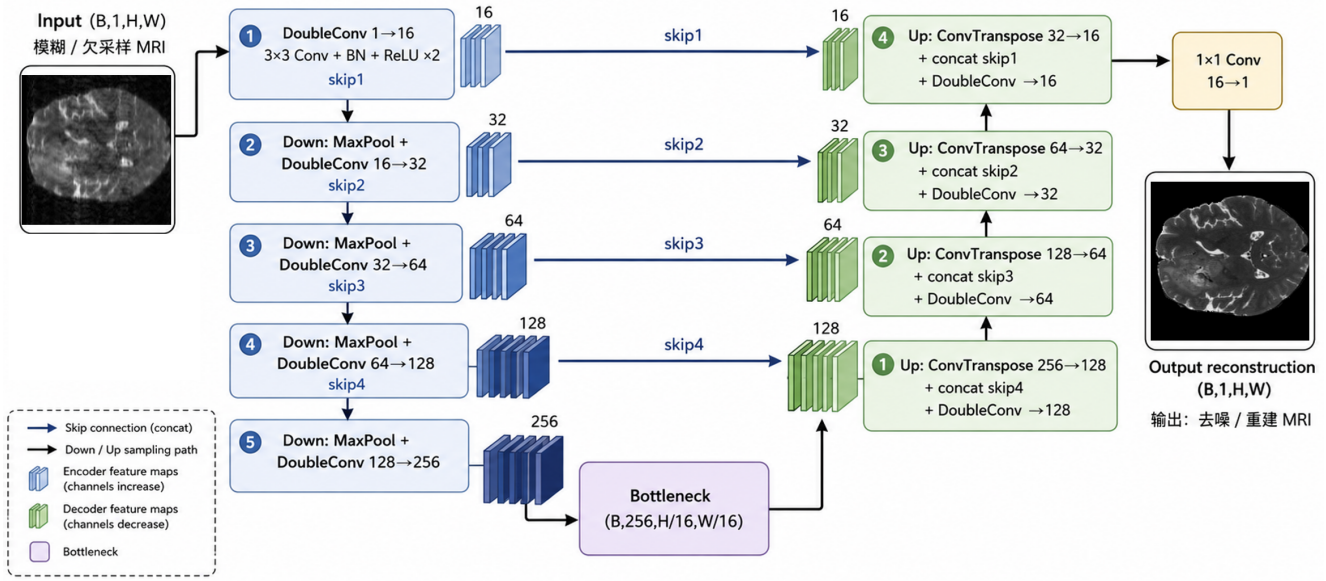


Figure 2. Overall pipeline of the baseline U-Net architecture (Task 2), which employs a symmetric encoder-decoder structure with skip connections to reconstruct high-quality, aliasing-free MRI images from undersampled T2 inputs.

the L_1 distance for each band separately:

$$\mathcal{L}_{wavelet} = w_{LL}\|LL_x - LL_{\hat{x}}\|_1 + w_{LH}\|LH_x - LH_{\hat{x}}\|_1 + w_{HL}\|HL_x - HL_{\hat{x}}\|_1 + w_{HH}\|HH_x - HH_{\hat{x}}\|_1 \quad (2)$$

Here, the LL band preserves the global contrast and basic structure (low-frequency), while the LH, HL, and HH bands extract high-frequency details along the horizontal, vertical, and diagonal directions, respectively. As illustrated in Figure 3, analyzing the error maps of these decomposed bands allows the network to precisely target and recover fine structural edges that are often blurred by standard MSE loss.

Perceptual Loss ($\mathcal{L}_{LPIPS}$): To ensure the reconstructed image aligns with human visual perception and preserves complex texture details, we incorporated the Learned Perceptual Image Patch Similarity (LPIPS) metric as a loss function. This perceptual loss computes the distance between deep feature representations extracted by a pre-trained network, rather than simply comparing raw pixel intensities.

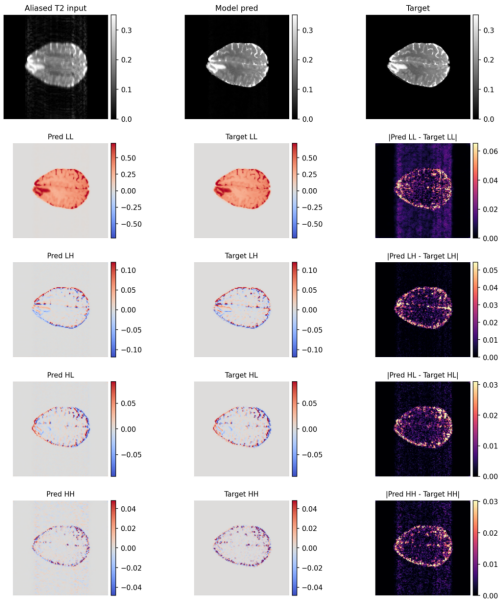


Figure 3. Visualization of the Wavelet Loss components. The figure compares the predicted and target images across different frequency bands (LL, LH, HL, HH), highlighting the structural details captured by the high-frequency components.

3. IMPLEMENTATION DETAILS

3.1. Data Preprocessing

We preprocess the BraTS data at the patient level, where each subject contains multi-modal 3D NIfTI volumes. In our current local copy, the dataset contains 625 subjects. For each subject, we load the T2-weighted volume (`t2w`) as the reconstruction target modality and the native T1-weighted volume (`t1n`) as the auxiliary anatomical modality. Both volumes are converted to `float32` and transposed into a slice-first representation of shape (D, H, W).

Each 3D volume is independently normalized to $[0, 1]$ using min-max normalization. We further discard nearly empty boundary slices using a low-variance threshold.

To emulate accelerated MRI acquisition, we transform each fully sampled T2 slice to k-space and apply a variable-density Cartesian undersampling mask. Unless otherwise stated, we use an acceleration factor of $R = 5$ with center fraction 0.08. The inverse FFT magnitude is used as the

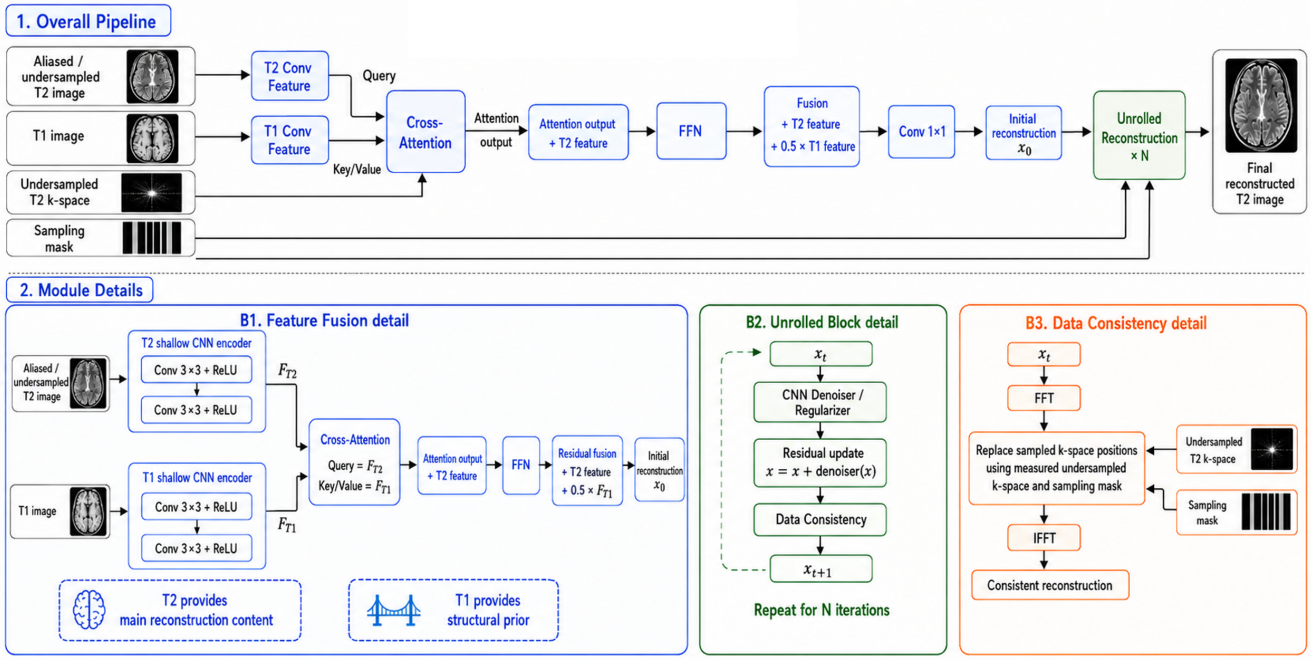


Figure 4. Overall pipeline of the Multi-modal Unrolled Reconstruction Network (Task 3).

aliased T2 input.

For each retained slice, we store the aliased T2 image, the fully sampled T2 image, the fully sampled T1 image, the undersampling mask, and the undersampled complex k-space.

We split the dataset at the patient level to avoid leakage across slices from the same subject. Specifically, subjects are randomly shuffled with seed 42 and divided into training, validation, and test sets with ratios of 70%, 15%, and 15%, respectively. Under the current dataset size of 625 subjects, this produces 437 training subjects, 93 validation subjects, and 95 test subjects.

3.2. U-Net (Task2) Training Settings

For Task 2, we train a baseline single-modal U-Net reconstructor on the preprocessed dataset. The network takes the aliased T2 slice as input and predicts the fully sampled T2 slice in the image domain.

The baseline model is a standard encoder-decoder U-Net with depth 4 and base width 16.

We optimize the model using MSE loss and Adam with learning rate 1×10^{-3}. Training runs for 50 epochs with batch size 128, and validation is performed every epoch.

We use a short warmup stage followed by `ReduceLROnPlateau` learning-rate decay, together with gradient clipping and early stopping.

During evaluation, we report NMSE, PSNR, SSIM, and LPIPS.

We keep the best checkpoint according to validation loss.

3.3. Multi-modal Unrolled Reconstruction Network (Task3) Training Settings

For Task 3, we train a multi-modal unrolled reconstruction network on the same preprocessed dataset. The model takes the aliased T2 image, fully sampled T1 image, undersampling mask, and undersampled T2 k-space as inputs, and predicts the fully sampled T2 image.

The model uses a T1-guided unrolled architecture with 5 iterations. Each iteration consists of a learned denoising block followed by a hard data-consistency step in k-space. In our main setting, the denoiser width is 32 channels.

The training objective combines MSE loss, Haar high-frequency wavelet loss, and LPIPS perceptual loss, with weights 1.0, 0.03, and 0.1, respectively. We train the model with Adam for 80 epochs using batch size 32 and gradient accumulation over 2 steps, giving an effective batch size of 64. The initial learning rate is 1.5×10^{-4}.

We use one-epoch warmup followed by `ReduceLROnPlateau` learning-rate decay, together with gradient clipping and mixed-precision training.

Validation is performed every epoch, and we report NMSE, PSNR, SSIM, and LPIPS.

We keep the best checkpoint according to validation loss.

4. Experimental Results

4.1. Quantitative Comparison: Task 2 vs. Task 3

To comprehensively evaluate the reconstruction performance of the baseline U-Net (Task 2) and the proposed Multi-modal Unrolled Network (Task 3), we adopted four

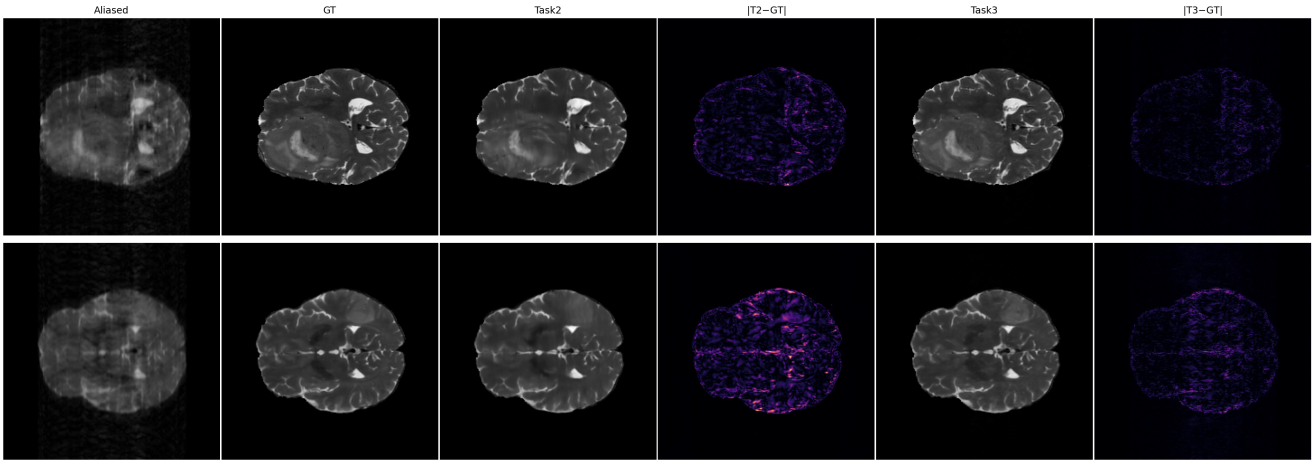


Figure 5. Qualitative comparison of reconstruction results between Task 2 and Task 3.

complementary metrics, including Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM), Learned Perceptual Image Patch Similarity (LPIPS), and Deep Image Structure and Texture Similarity (DISTS). While PSNR and SSIM mainly measure pixel-level fidelity and structural consistency, LPIPS and DISTS further evaluate perceptual similarity and texture preservation from a human visual perspective.

As summarized in Table 1, the proposed Task 3 framework consistently achieves superior performance across all evaluation metrics. Specifically, Task 3 improves PSNR from 37.21 dB to 41.11 dB and increases SSIM from 0.885 to 0.954, demonstrating substantially better reconstruction accuracy and structural consistency. Meanwhile, LPIPS decreases from 0.0205 to 0.0090 and DISTS decreases from 0.1558 to 0.0936, indicating that the proposed method produces reconstructions with significantly improved perceptual quality and more realistic texture recovery.

The qualitative comparisons in Figure 5 further support these observations. Compared with the baseline Task 2 results, the Task 3 reconstructions exhibit sharper anatomical boundaries, clearer tumor structures, and fewer residual artifacts. Moreover, the corresponding error maps ($|T2-GT|$ and $|T3-GT|$) reveal that the proposed method significantly suppresses reconstruction errors, particularly around fine structural details and high-frequency regions. These improvements demonstrate the effectiveness of incorporating multi-modal priors together with iterative data consistency refinement in the unrolled reconstruction framework.

4.2. Qualitative Results and Training Logs

Figure 6 shows the five worst Task 3 cases ranked by PSNR. Despite severe aliasing in the inputs, the reconstructions preserve the main anatomical structures and lesion regions, demonstrating the robustness of the proposed model on challenging samples. Remaining errors mainly appear as mild over-smoothing of fine textures and residual boundary artifacts, especially near cortical folds, tumor margins, and image borders.

Table 1. Quantitative comparison between Task 2 (Baseline) and Task 3 (Multi-modal Unrolled Network).

Model	**DISTS** ↓	**LPIPS** ↓	**PSNR** ↑	**SSIM** ↑
Task 2 (U-Net)	0.1558	0.0205	37.21	0.885
Task 3 (Proposed)	0.0936	0.0090	41.11	0.954

Figure 7 shows the validation curves of the Task 2 baseline. The validation loss and NMSE decrease rapidly in the early training stage and then gradually converge, while PSNR and SSIM increase and stabilize. This indicates that the baseline U-Net learns a reasonable reconstruction mapping from undersampled inputs. However, the PSNR and SSIM curves exhibit noticeable early-stage fluctuations, suggesting that the baseline remains less stable on challenging validation samples.

Figure 8 further reports the Task 3 training dynamics. The training and validation losses decrease rapidly in the early epochs and then gradually stabilize, suggesting effective optimization without obvious overfitting. Meanwhile, PSNR increases steadily, while LPIPS and DISTS decrease, indicating simultaneous improvements in pixel fidelity, perceptual similarity, and structural texture preservation.

Overall, the qualitative results and training logs are consistent with Table 1. Task 3 produces sharper and more faithful reconstructions than the Task 2 baseline, while the remaining failures are mainly concentrated in fine anatomical details, lesion boundaries, and image-edge artifacts.

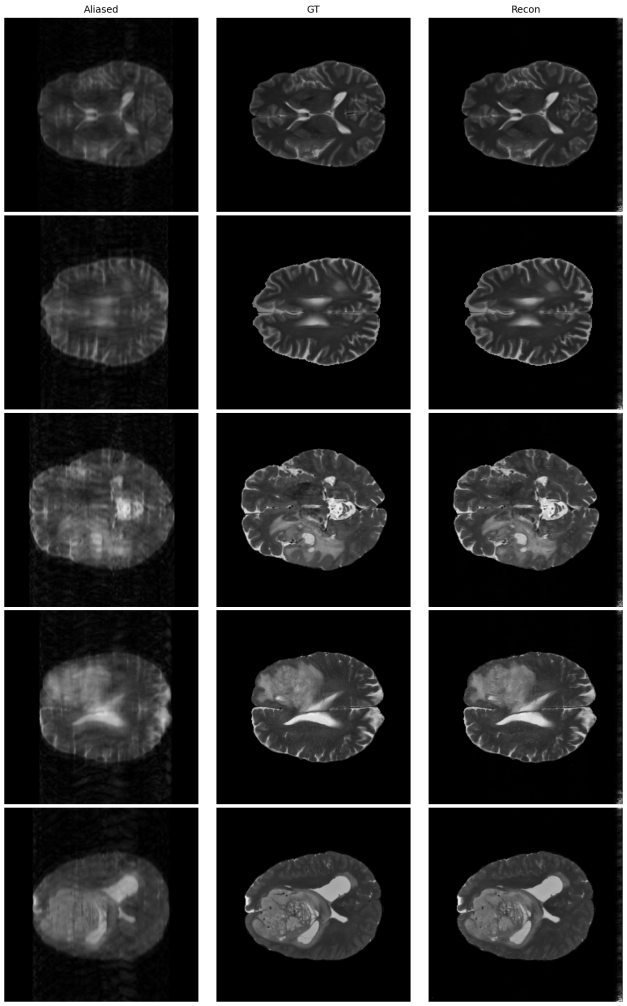


Figure 6. The five worst-performing reconstruction cases from the task3 baseline (calculated using PSNR as the metric).

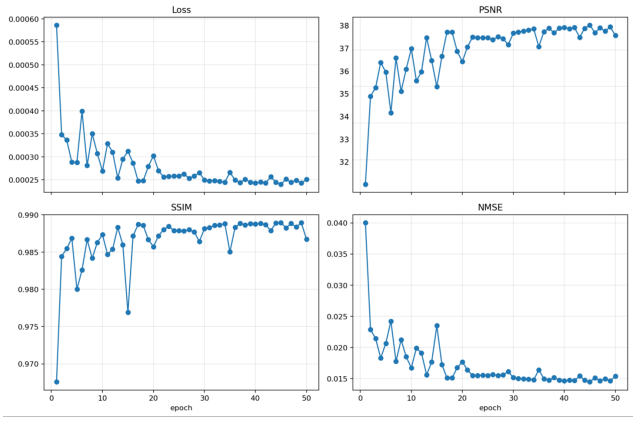


Figure 7. Validation loss curves for Task 2.

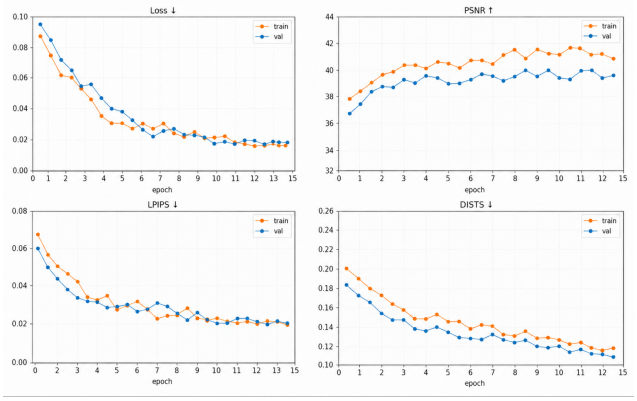


Figure 8. Training and Validation loss curves for Task 3.

4.3. Rethinking PSNR: Alignment with Human Perception

During our initial analysis of the Task 2 results, we encountered a significant contradiction: images with higher PSNR values did not consistently yield better subjective visual quality. As illustrated in Figure 9, Slice 43 achieves a higher PSNR (39.10 dB) than Slice 28 (35.34 dB). However, inspecting the Error Maps and the perceptual LPIPS metric reveals the opposite story: Slice 43 has a significantly worse LPIPS score (0.0364 vs. 0.0194) and exhibits denser, more structured errors, aligning better with human visual dissatisfaction.

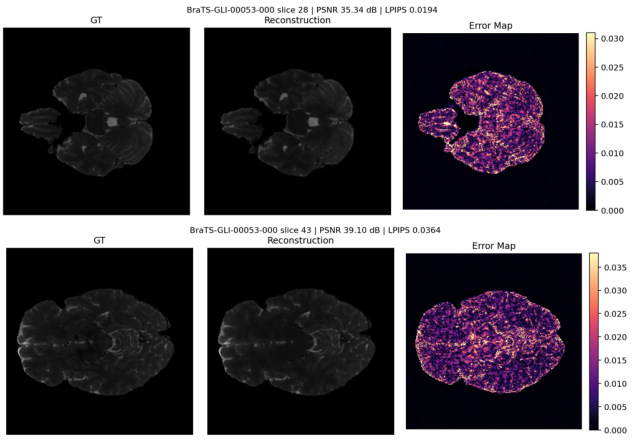
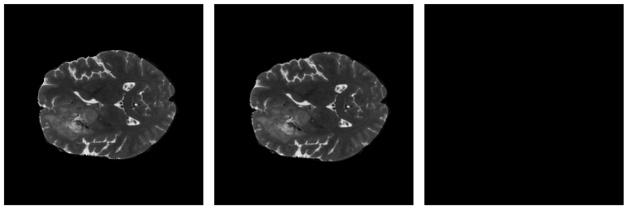


Figure 9. Contradiction between PSNR and Perceptual Quality. Slice 43 (bottom) has a higher PSNR but a noticeably worse Error Map and LPIPS score compared to Slice 28 (top).

To systematically investigate this metric failure, we drew inspiration from the supplemental materials of Apple's research on image sharpness [10]. We conducted a "shifting pixels" experiment by translating the reference MRI image by minor spatial percentages.

As shown in Figure 10 and detailed in Table 2, shift-



(a) Reference image (MRI) (b)Translated (1.0%) (b)Mean Image

Figure 10. The translation experiment. A visually identical translated image suffers massive PSNR degradation compared to the reference.

ing the image by merely 1.0% causes the PSNR to plummet drastically to 20.9 dB, a score typically associated with severe degradation. Yet, to the human eye, the 1.0% translated image is nearly indistinguishable from the reference. In contrast, perceptual metrics like DISTS and LPIPS remain relatively stable and robust to these imperceptible shifts. This experiment definitively proves that purely pixel-aligned metrics like PSNR and MSE heavily penalize slight spatial misalignments while ignoring structural integrity, justifying our critical decision to incorporate Wavelet and Perceptual (LPIPS/DISTS) losses into our final Task 3 pipeline.

Table 2. Results of the spatial translation experiment on evaluation metrics. Perceptual metrics (DISTS, LPIPS) degrade gracefully, whereas pixel-wise metrics (PSNR, SSIM) collapse under minor shifts.

Comparison	DISTS ↓	LPIPS ↓	PSNR ↑	SSIM ↑
Translated (0.1%)	0.028	0.007	35.9	0.988
Translated (1.0%)	0.073	0.059	20.9	0.760
Translated (5.0%)	0.070	0.220	17.4	0.651
Translated (15.0%)	0.120	0.422	15.2	0.538
Mean Image	0.655	0.612	16.8	0.085

4.4. Ablation Study

To validate the individual contributions of our proposed hybrid loss components, we conducted a rigorous ablation study on our network. We incrementally added the Wavelet Loss and the Perceptual (LPIPS) Loss to the baseline MSE objective.

Table 3. Quantitative ablation results comparing the full proposed multi-modal model against simplified configurations.

Configuration	DISTS ↓	LPIPS ↓	PSNR ↑	SSIM ↑
w/o Wavelet	0.1289	0.0185	38.77	0.888
w/o Perceptual	0.1395	0.0198	37.56	0.885
w/o Cross-Attention	0.1334	0.0191	37.79	0.887
Ours (Task 3)	**0.0936**	**0.0090**	**41.11**	**0.954**

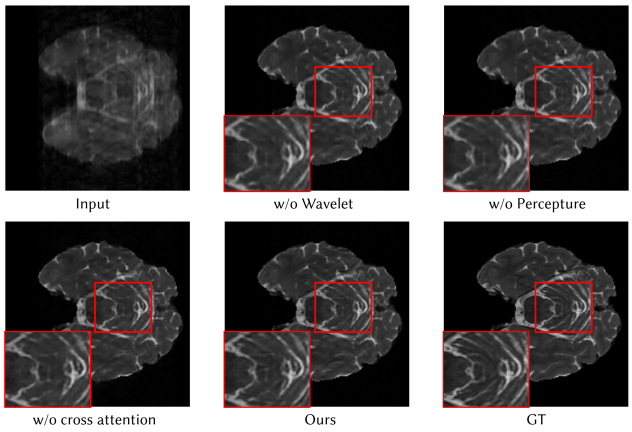


Figure 11. Qualitative ablation study on key training components.

As observed in Table 3 and Figure 11, relying solely on MSE leads to over-smoothed textures. The addition of the Wavelet loss successfully recovers high-frequency edges, while the further integration of the Perceptual loss maximally restores the organic texture of the brain tissues, yielding the best overall DISTS and LPIPS scores.

5. Conclusion

In this work, we studied multi-contrast MRI reconstruction from undersampled k-space data. We first established a single-modal U-Net baseline for Task 2, which effectively suppresses aliasing artifacts but tends to produce over-smoothed reconstructions. To overcome this limitation, we proposed a Multi-modal Unrolled Reconstruction Network for Task 3, which leverages the fully sampled T1 modality as anatomical guidance for undersampled T2 reconstruction. By combining cross-attention-based feature fusion, iterative unrolled refinement, and explicit k-space data consistency, the proposed method achieves more faithful and perceptually realistic reconstructions.

Extensive experiments show that our Task 3 model consistently outperforms the Task 2 baseline across PSNR, SSIM, LPIPS, and DISTS. Qualitative results further demonstrate that the proposed framework better preserves anatomical boundaries, tumor regions, and fine structural details while reducing residual undersampling artifacts. In addition, our analysis of PSNR reveals its limitation in reflecting perceptual image quality, motivating the use of perceptual and structure-aware metrics. The ablation study confirms the effectiveness of the proposed components, including wavelet loss, perceptual loss, and cross-attention-based multi-modal fusion.

Despite these improvements, challenging pathological cases still exhibit mild over-smoothing and boundary artifacts. Future work may explore stronger edge-aware constraints, more robust multi-modal registration, and ad-

vanced frequency-domain modeling. Overall, our results demonstrate that integrating multi-modal anatomical priors with unrolled data consistency and perceptually motivated objectives is an effective direction for high-quality accelerated MRI reconstruction.

References

[1] Lustig, M., Donoho, D., and Pauly, J. M. Sparse MRI: The application of compressed sensing for rapid MR imaging. *Magnetic Resonance in Medicine*, 58(6):1182–1195, 2007. 1

[2] Schlemper, J., Caballero, J., Hajnal, J. V., Price, A. N., and Rueckert, D. A deep cascade of convolutional neural networks for dynamic MR image reconstruction. *IEEE Transactions on Medical Imaging*, 37(2):491–503, 2018.

[3] Wang, S., Su, Z., Ying, L., Peng, X., Zhu, S., Liang, F., Feng, D., Liang, D. Accelerating magnetic resonance imaging via deep learning. In *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*, pages 514–517, 2016. 1

[4] Ronneberger, O., Fischer, P., and Brox, T. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, 2015. 1

[5] Johnson, J., Alahi, A., and Fei-Fei, L. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision (ECCV)*, pages 694–711, 2016. 1

[6] Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., and Shi, W. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4681–4690, 2017. 1

[7] Kang, E., Min, J., and Ye, J. C. Wavelet domain residual network (WavResNet) for low-dose X-ray CT reconstruction. *IEEE Transactions on Medical Imaging*, 37(12):2986–2997, 2018. 1

[8] Ding, K., Ma, K., Wang, S., and Simoncelli, E. P. Image quality assessment: Unifying structure and texture similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5):2567–2581, 2022.

[9] Zhou, B. and Zhou, S. K. DuDoRNet: Learning a dual-domain recurrent network for fast MRI reconstruction with deep T1 prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4273–4282, 2020. 1

[10] Mescheder, L., Dong, W., Li, S., Bai, X., Santos, M., Hu, P., Lecouat, B., Zhen, M., Delaunoy, A., Fang, T., Tsin, Y., Richter, S. R., and Koltun, V. Sharp monocular view synthesis in less than a second. *arXiv preprint arXiv:2512.10685*, 2025. 6