

Memory-Guided Random-Direction Feature Disentanglement Model for Multi-Phase MRI Translation

Qianmu Xiao^{1,5}, Xiaoyan Kui^{1*}, Zexin Ji², Yang Gao¹, Wengang Pan³, Wenqi Lu⁴, Jinming Duan⁵, and Beiji Zou¹

¹ School of Computer Science and Engineering, Central South University, China
qianmu@csu.edu.cn, xykui@csu.edu.cn

² Department of Radiation Oncology, Washington University, United States

³ Xiangya School of Nursing, Central South University, China

⁴ Department of Computing and Mathematics, Manchester Metropolitan University, United Kingdom

⁵ School of Health Sciences, University of Manchester, United Kingdom

Abstract. In this paper, we propose a Memory-Guided Random-Direction Feature Disentanglement (MG-RDFD) framework for multi-phase CE-MRI translation. Our method learns an interpretable and parameter-efficient encoder-decoder model without resorting to multiple domain-specific encoders or additional content discriminators. Given posteriors from paired inputs encoded by a shared VAE encoder, MG-RDFD selects a shared content anchor and phase-preferred style anchors along a random 1D latent direction using density-based criteria, which explicitly couples content and style within a unified latent geometry. We further introduce a segmentation-guided style memory bank to retrieve region-wise style prototypes and enhance organ/lesion fidelity. Experiments on multi-phase CE-MRI demonstrate competitive translation quality with favorable inference parameter efficiency, and the memory bank reduces regional texture confusion. The implementation is available at <https://github.com/QianMuXiao/MG-RDFD>.

Keywords: Multi-phase MRI Translation · Content-Style Disentanglement · Segmentation-guided Memory Bank · Medical Image Synthesis

1 Introduction

Multi-phase contrast-enhanced MRI (CE-MRI) provides complementary anatomical and hemodynamic cues that are valuable for diagnosis and treatment planning. In practice, acquiring a complete set of high-quality phases is often constrained by scan time and patient tolerance. It can also be limited by concerns about repeated examinations and contrast-agent exposure. A pragmatic strategy is to acquire only a subset of low-quality phases. A high-quality complete set

* Corresponding author.

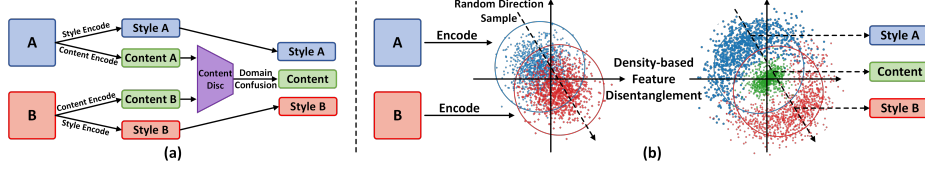


Fig. 1. Content-Style Disentanglement Model Architecture: (a) Independent Encoder and Content Discrimination Framework (b) Our Proposed Density-Driven Random Direction Feature Disentanglement Framework

can then be recovered by combining medical image super-resolution [11,4] and cross-phase translation [6,21,25] models.

Existing translation models largely fall into two paradigms. The first paradigm performs end-to-end image mapping [9,7,28,3] with high-capacity generators. It often achieves strong fidelity, but it remains a black-box transformation and typically comes with a large inference footprint. More recently, diffusion-based generative models [23,24] have shown impressive synthesis quality. However, their training and sampling procedures can be computationally expensive. This makes deployment and extensive experimentation non-trivial in many medical settings.

An alternative paradigm is content-style disentanglement [15,10,16,27,29]. It aims to separate structural content from appearance style to improve interpretability and controllability. It can also be parameter-efficient at inference time. However, many disentanglement-based translators rely on multiple encoders and additional discriminators to enforce content invariance and style separation (Fig. 1(a)). More importantly, content and style are often extracted as two parallel variables. There is usually no explicit mechanism that couples them within a unified latent geometry. As a result, their relationship is implicitly learned and may be under-constrained for medical translation where structure fidelity is critical.

To address these limitations, we propose a memory-guided random-direction feature disentanglement framework (MG-RDFD) (Fig. 2). Instead of introducing multiple encoders, we design a shared VAE [13] encoder to map paired multi-phase slices into two latent Gaussian distributions. We further regularize these distributions toward a symmetric structure via Flip Distribution Alignment (FDA) [14]. We then construct a random one-dimensional subspace induced by the two distributions. On this subspace, we select a content point and two style points using density-based criteria. The content point maximizes the product of the two densities, while the style points maximize the density discrepancy (Fig. 1(b)). This procedure determines content and style within the same latent geometry. It provides an explicit coupling between shared factors and domain-specific factors for translation. Furthermore, to enhance region-specific appearance modeling, we incorporate a segmentation-guided style memory bank [20]. The memory aggregates and retrieves style cues for organs and lesions. Extensive

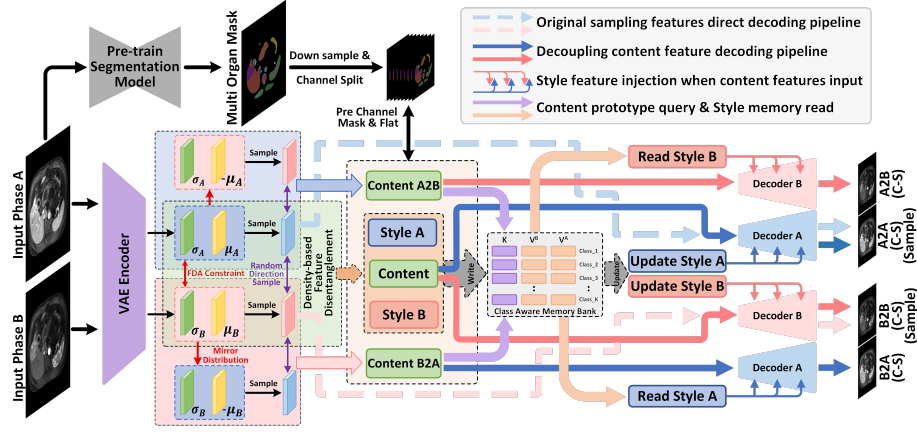


Fig. 2. Main framework of Memory-Guided Random-Direction Feature Disentanglement Model.

experiments on multi-phase abdominal MRI translation demonstrate that our method improves synthesis quality and achieves favorable parameter efficiency compared to representative end-to-end and disentanglement-based baselines.

2 Method

Random Direction Feature Disentanglement (RDFD). Given two normal distributions $q_A = \mathcal{N}(\mu_A, (\sigma_A)^2)$ and $q_B = \mathcal{N}(\mu_B, (\sigma_B)^2)$, we sample $s_A \sim q_A$ and $s_B \sim q_B$ and define a random line $x(t) = s_A + t(s_B - s_A)$. RDFD is applied independently at each spatial location (h, w) by treating channels as feature dimensions; this yields a per-pixel scalar field $t \in \mathbb{R}^{1 \times H \times W}$ and keeps content/style codes spatially aligned for decoding and region-level retrieval. Fig. 3 visualizes RDFD on 2D distributions: the density-overlap maximizer lies in the shared region (content), whereas extrema of density discrepancy move toward domain-preferred regions (styles) across different shifts.

We define the content feature as the point on the line with maximum joint density:

$$z_c = x(t_c), \quad t_c = \arg \max_t p_A(x(t)) p_B(x(t)). \quad (1)$$

For diagonal Gaussians, maximizing $\log p_A(x(t)) + \log p_B(x(t))$ is equivalent to minimizing a per-pixel quadratic $at^2 + bt$, giving the closed-form optimum $t_c = -\frac{b}{2a}$. Let $v = s_B - s_A$ and define $w_{A,c} = \frac{1}{(\sigma_{A,c})^2 + \epsilon}$, $w_{B,c} = \frac{1}{(\sigma_{B,c})^2 + \epsilon}$. The per-pixel coefficients (reduced over channels) are:

$$a = \sum_c (w_{A,c} + w_{B,c}) v_c^2, \quad b = 2 \sum_c [w_{A,c} v_c (s_{A,c} - \mu_{A,c}) + w_{B,c} v_c (s_{A,c} - \mu_{B,c})]. \quad (2)$$

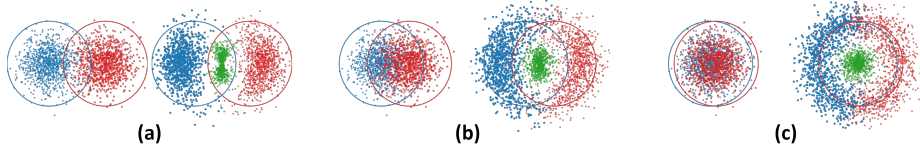


Fig. 3. Visualization of random-direction feature disentanglement under different distribution shifts.

To capture domain preference, we select two style feature points on the same line via density discrepancy. Since Gaussian tails may yield degenerate optima in preference search, we search extrema of $\Delta(t) = p_A(x(t)) - p_B(x(t))$ within a bounded interval:

$$z_{s,A} = x(t_A), \quad t_A = \arg \max_{t \in [-T, T]} \Delta(t), \quad z_{s,B} = x(t_B), \quad t_B = \arg \min_{t \in [-T, T]} \Delta(t). \quad (3)$$

The range is set by directional uncertainty: let $\hat{v} = \frac{v}{\|v\|}$ and set search half-range $T = \kappa \max(\text{std}_A(\hat{v}), \text{std}_B(\hat{v}))$, where $\text{std}_A(\hat{v}) = \sqrt{\sum_c \hat{v}_c^2 \sigma_{A,c}^2}$ (and similarly for B). In practice, we discretize $[-T, T]$ into G grid points $\{t_g\}$, compute $\Delta_g = \Delta(t_g)$, and estimate t_A, t_B using differentiable soft-argmax/soft-argmin:

$$t_A = \sum_g \pi_g^+ t_g, \quad \pi_g^+ = \frac{\exp(\Delta_g/\tau)}{\sum_j \exp(\Delta_j/\tau)}, \quad t_B = \sum_g \pi_g^- t_g, \quad \pi_g^- = \frac{\exp(-\Delta_g/\tau)}{\sum_j \exp(-\Delta_j/\tau)}. \quad (4)$$

Finally, we assign the two candidates by comparing their log-likelihoods under q_A (equivalently q_B): the one with larger $\log p_A(\text{candidate})$ is taken as $z_{s,A}$, and the other as $z_{s,B}$; the selected codes are passed through lightweight projection heads (1×1 conv + GN + SiLU) to match channel dimensions for the decoders and the memory bank. We set $\kappa = 3.0$, $G = 2049$, $\tau = 0.15$, and $\epsilon = 10^{-8}$.

Segmentation-Guided Memory Bank. To enhance local appearance consistency in cross-phase synthesis, we introduce a segmentation-guided style memory bank[27,17]. It organizes RDFS representations into region-level style prototypes and retrieves dense style maps for subsequent decoding. By modeling appearance variation in a region-prototype manner, the memory stabilizes local texture and contrast changes during translation.

We represent the semantic partition by $K=16$ mutually exclusive one-hot region masks $\{\mathbf{m}_k\}_{k=1}^K$ (7 organs, 7 lesions, abdomen, border), where each pixel activates exactly one mask. For each region k , we maintain M_k prototype slots with content keys and domain-specific style values:

$$\mathcal{M} = \left\{ (\mathbf{K}_k, \mathbf{V}_k^A, \mathbf{V}_k^B) \right\}_{k=1}^K, \quad \mathbf{K}_k \in \mathbb{R}^{M_k \times d_c}, \quad \mathbf{V}_k^{A/B} \in \mathbb{R}^{M_k \times d_s}. \quad (5)$$

As shown in Fig. 4, given a content feature map $\mathbf{C} \in \mathbb{R}^{B \times d_c \times H \times W}$, for region- k pixels $\Omega_k = \{i | \mathbf{m}_k(i) = 1\}$, we query region keys with the content

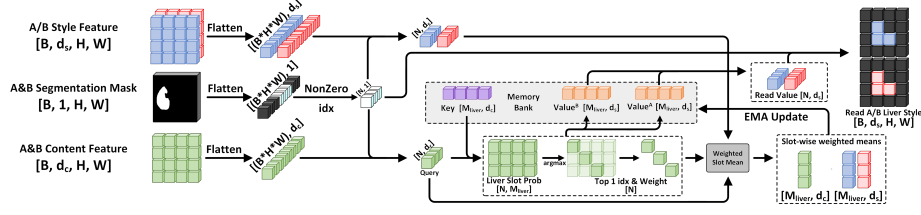


Fig. 4. Memory bank update and read for one semantic class (e.g., liver). We use the content feature as query to produce slot assignment, update slots via top-1 routing with confidence weights, and read style prototypes via soft retrieval.

vector $q_i = \mathbf{C}(i)$ and read style prototypes:

$$p_{i,j} = \text{softmax}_j(q_i^\top \mathbf{K}_{k,j}), \quad \hat{s}_i^{A/B} = \sum_{j=1}^{M_k} p_{i,j} \mathbf{V}_{k,j}^{A/B}. \quad (6)$$

Scattering $\hat{s}^A$ and $\hat{s}^B$ back to spatial locations yields dense style maps $\hat{\mathbf{S}}^A, \hat{\mathbf{S}}^B \in \mathbb{R}^{B \times d_s \times H \times W}$. We assign slots by $j^*(i) = \arg \max_j p_{i,j}$, compute slot-wise weighted means $\tilde{\mathbf{K}}_{k,j}, \tilde{\mathbf{V}}_{k,j}^{A/B}$ (using $p_{i,j}$), and apply momentum updates:

$$\mathbf{K}_{k,j} \leftarrow \text{normalize}\left((1 - \mu)\mathbf{K}_{k,j} + \mu \tilde{\mathbf{K}}_{k,j}\right), \quad \mathbf{V}_{k,j}^{A/B} \leftarrow (1 - \mu)\mathbf{V}_{k,j}^{A/B} + \mu \text{LN}\left(\tilde{\mathbf{V}}_{k,j}^{A/B}\right). \quad (7)$$

To prevent prototype collapse and encourage consistent assignment, we apply cross-entropy on key/value retrieval logits using $j^*(i)$ as the pseudo label, where s_i^D denotes the projected RDFD style features at pixel i for $D \in \{A, B\}$:

$$\mathcal{L}_{\text{reg}} = \lambda_k \text{CE}(\{q_i^\top \mathbf{K}_{k,j}\}_j, j^*(i)) + \lambda_v \sum_{D \in \{A, B\}} \text{CE}(\{s_i^D \mathbf{V}_{k,j}^D\}_j, j^*(i)). \quad (8)$$

Regional supervision alone may not sufficiently separate prototypes across different regions in the global space. We therefore concatenate all region keys into a global dictionary $\mathbf{K}_{\text{all}} = \text{concat}(\mathbf{K}_1, \dots, \mathbf{K}_K) \in \mathbb{R}^{M_{\text{tot}} \times d_c}$, where $M_{\text{tot}} = \sum_{k=1}^K M_k$, map the regional assignment $j^*(i)$ to a global label $g^*(i) \in \{1, \dots, M_{\text{tot}}\}$ via region offsets, and impose an additional global cross-entropy:

$$\mathcal{L}_{\text{glob}} = \lambda_g \text{CE}(\{q_i^\top \mathbf{K}_{\text{all},g}\}_{g=1}^{M_{\text{tot}}}, g^*(i)). \quad (9)$$

The final memory loss is $\mathcal{L}_{\text{mem}} = \mathcal{L}_{\text{reg}} + \mathcal{L}_{\text{glob}}$.

Training & Inference Pipeline. Each iteration takes a registered paired sample (x_A, x_B) . The shared encoder produces two diagonal Gaussian posteriors $q_A(z|x_A)$ and $q_B(z|x_B)$. We jointly optimize (i) direct reconstruction, (ii) disentangled reconstruction, and (iii) paired cross-domain translation.

We first sample $z_A \sim q_A$ and $z_B \sim q_B$, and reconstruct $\hat{x}_A = G_A(z_A)$ and $\hat{x}_B = G_B(z_B)$ with a pixel-wise reconstruction loss. We then apply RDFD on the same pair of posteriors to obtain a shared content code z_c and domain-specific style features at each spatial location; the style features are enhanced by the memory bank to yield dense style maps $\tilde{\mathbf{S}}^A, \tilde{\mathbf{S}}^B$. Decoding $(z_c, \tilde{\mathbf{S}}^A)$ and $(z_c, \tilde{\mathbf{S}}^B)$ produces disentangled reconstructions $\hat{x}_A^{cs}$ and $\hat{x}_B^{cs}$, enforcing that RDFD codes can faithfully reconstruct the inputs. We use a single decoder with two modes: in the direct pathway it performs pure decoding (style = $\emptyset$), while in the disentangled pathway we inject the memory-enhanced style map as multi-layer additive residual modulation before each ResBlock (with the style branch upsampled synchronously).

To enable single-input inference, we exploit the symmetric latent structure encouraged by FDA: we construct a mirrored pseudo-target posterior by flipping the mean while keeping the variance unchanged. For $A \rightarrow B$, given x_A we obtain q_A , build $\tilde{q}_B$, and run RDFD on $(q_A, \tilde{q}_B)$ to obtain the content code and a target-domain style code; target styles are enhanced by the memory bank and decoded to generate $\hat{x}_{A \rightarrow B}$. The $B \rightarrow A$ path is defined symmetrically. During training, besides paired reconstruction, we further apply a PatchGAN [9,8] adversarial loss and a perceptual loss [12] on the translated outputs to improve realism and visual quality. We minimize the following weighted objective:

$$\begin{aligned} \mathcal{L} = & \lambda_{\text{rec-samp}} \mathcal{L}_{\text{rec-samp}} + \lambda_{\text{rec-cs}} \mathcal{L}_{\text{rec-cs}} + \lambda_{\text{trans}} \mathcal{L}_{\text{trans}} + \lambda_{\text{gan}} \mathcal{L}_{\text{gan}} \\ & + \lambda_{\text{per}} \mathcal{L}_{\text{per}} + \lambda_{\text{kl}} \mathcal{L}_{\text{kl}} + \lambda_{\text{fda}} \mathcal{L}_{\text{fda}} + \lambda_{\text{mem}} \mathcal{L}_{\text{mem}}. \end{aligned} \quad (10)$$

$\mathcal{L}_{\text{rec-samp}}$ and $\mathcal{L}_{\text{rec-cs}}$ supervise direct and disentangled reconstructions, $\mathcal{L}_{\text{trans}}$ supervises cross-phase translation, $\mathcal{L}_{\text{gan}}$ and $\mathcal{L}_{\text{per}}$ correspond to PatchGAN adversarial and perceptual supervision, $\mathcal{L}_{\text{kl}}$ and $\mathcal{L}_{\text{fda}}$ regularize the posteriors and enforce symmetry, and $\mathcal{L}_{\text{mem}}$ denotes memory-bank related constraints. We set $\lambda_{\text{trans}} = 1$, $\lambda_{\text{mem}} = 0.1$, $\lambda_{\text{kl}} = 10^{-7}$, and all others to 0.01.

3 Experiments

3.1 Experimental Setup

Dataset & preprocessing. We use the LLD-MMRI 2023 dataset [18] and build three cross-phase translation tasks among Pre, CA, and CV phases. To ensure spatial consistency, Pre and CA are registered to CV using ANTsPy [2]. The dataset provides annotations for seven liver-lesion types; in addition, we run TotalSegmentator MRI [1] model on the CV volumes to obtain multi-organ masks and keep liver, lungs, kidneys, spleen, stomach, heart, and spine as major regions, while merging the remaining in-body area into an abdomen category. Before slice extraction, we performed a lesion-stratified patient-level split. Since each patient is associated with one lesion category, patients were stratified by lesion type and divided into training/validation cohorts at a 4:1 ratio to preserve category proportions. After extracting and resizing 2D paired slices to 256×256 , we obtained approximately 32k paired slices for model training and validation.

Table 1. Overview of quantitative results on three translation tasks. Full-image cells report PSNR/SSIM. ROI cells report the average (over three tasks) pixel-weighted PSNR/MAE on liver and lesion masks. Inference params are in millions (M).

Model	Params (M)	Full Slices PSNR(dB)/SSIM(%)			Liver Avg PSNR(dB) /MAE	Lesion Avg PSNR(dB) /MAE
		Pre→CA	Pre→CV	CA→CV		
Pix2Pix[9]	51.89	23.65/70.81	22.80/71.75	23.94/73.49	19.93/0.077	16.83/0.112
PTNet[28]	26.83	24.07/76.69	23.44/75.08	24.66/78.46	20.54/0.070	17.36/0.103
TransUnet[5]	100.45	24.78/78.25	24.16/76.25	25.51/79.65	21.24/0.064	17.87/0.098
ResVit[7]	117.72	25.04/78.01	24.25/73.27	25.37/76.39	21.04/0.064	17.59/0.101
DRIT++[15]	11.81	25.24/80.11	23.47/76.96	25.18/80.16	21.70/0.060	16.93/0.111
MGUIT[10]	16.27	23.15/77.31	22.92/76.52	24.79/80.25	20.17/0.072	17.17/0.106
FDA_VAE[14]	11.78	25.15/80.40	24.17/77.35	25.65/80.19	21.14/0.064	17.51/0.102
RDFD-only	13.06	24.85/78.59	24.28/77.27	25.79/79.84	21.14/0.064	17.81/0.100
MG-RDFD	13.15	25.36/79.99	24.63/77.36	25.98/80.87	21.40/0.062	18.09/0.096

Evaluation metrics. We evaluate full-slice quality using PSNR and SSIM [26]. We further compute average ROI PSNR and MAE on valid pixels within the liver and lesion regions in three tasks. We also calculate the minimum set of parameters activated for cross-phase inference.

Training details. The maximum number of training epochs for all models is 200. We use Adam optimizer with a learning rate of 1×10^{-4} on 4×RTX 5090 GPUs, the batch size is set to 24.

3.2 Ablation Studies

We build our method upon the FDA-VAE[14] backbone[22,19], where a shared encoder maps paired images into two symmetric latent distributions. This structured latent space is the foundation that makes our subsequent disentanglement feasible. On top of it, we introduce RDFD to decouple reliable content and style anchors from the two distributions, leading to an independently runnable variant (RDFD-only, Tab.1), where the reconstruction style and the target style for single-domain inference are both obtained by RDFD. Finally, we integrate the segmentation-guided style memory bank, and enforce that all style features are retrieved from the memory for style injection, yielding the full model (MG-RDFD, Tab.1). Compared with FDA-VAE, RDFD-only yields gains on lesion ROI fidelity and improves PSNR on two tasks, while its Pre→CA SSIM drop may partly stem from residual inter-phase misregistration that penalizes pixel-wise structural scores. After integrating the memory bank, MG-RDFD becomes consistently stronger on both full-slice and ROI metrics, suggesting that memory-based style retrieval effectively compensates local texture inconsistencies and stabilizes translation.

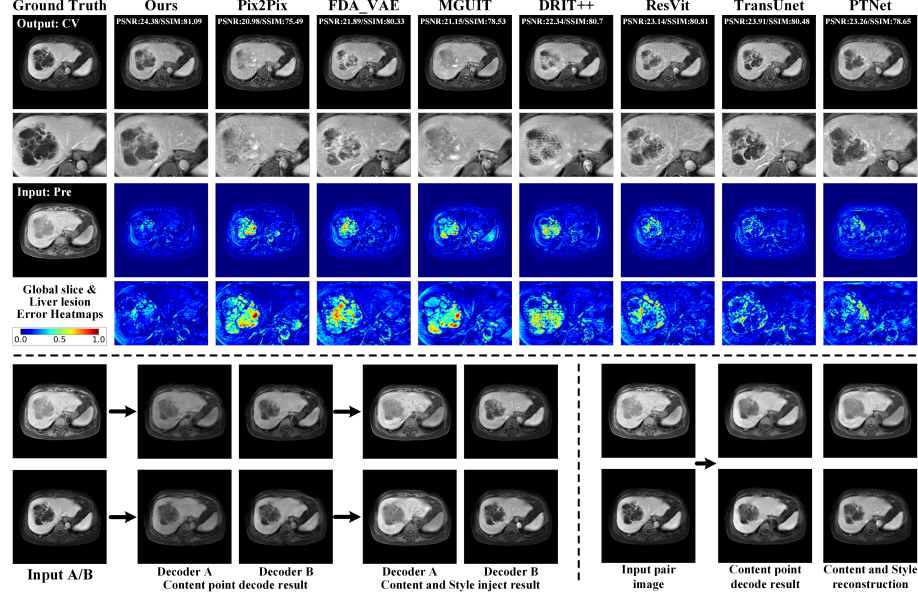


Fig. 5. Visualization of translation results, error maps and RDFD content anchors on a challenging lesion slice.

3.3 Comparisons with State-of-the-Art Methods

Quantitative Analysis. We compare our method with representative supervised end-to-end Unet-like baselines (Pix2Pix[9], PTNet [28], TransUNet[5] and ResViT[7]) and disentanglement models (DRIT++[15] and MGUIT[10]). Note that DRIT++ and MGUIT were originally designed for unpaired translation; to adapt them to our paired CE-MRI setting, we train them with paired inputs data and adding a cross-phase pixels ℓ_1 loss and removing random synthesis branches. As shown in Tab.1, our full model (MG-RDFD) achieves a favorable quality-efficiency trade-off under a lightweight inference pipeline, delivering strong full-slice fidelity across tasks while improving liver/lesion ROI metrics. Overall, the results support our goal of training an interpretable, single-pass encoder-decoder model with competitive translation performance.

Qualitative Analysis. Fig. 5 presents a challenging liver lesion slice for qualitative evaluation. Compared with competing methods, our synthesis better preserves global intensity patterns and reduces local errors within clinically relevant regions, as reflected by the error maps. In addition, the lower panel visualizes the intermediate content anchor decoded from the RDFD-only pipeline: even without explicit supervision on this intermediate decoding, it captures the main anatomy and contrast layout shared across phases. After injecting the corresponding style representation, the decoder recovers phase-specific details, illustrating the interpretability of our anchor-based disentanglement.

4 Conclusion

In this paper, we propose a content-style disentanglement framework for multi-phase MRI translation. A shared encoder maps images from different phases into a symmetric latent space, where we introduce RDFD, a density-driven random-direction anchoring mechanism to identify reliable content and style points for translation. Unlike prior pipelines that rely on separate encoders and content discriminators, our method performs interpretable anchoring and style injection within a unified latent space. We further incorporate a segmentation-guided style memory bank to enhance region-level texture consistency and improve synthesis quality. Experiments demonstrate a favorable trade-off between multi-metric translation fidelity and inference parameter efficiency.

Acknowledgments. This work was supported by the National Natural Science Foundation of China (Nos. U22A2034 and 62301616), the Scientific Research Fund of Hunan Provincial Education Department (No. 24A0018), the Major Program of Xiangjiang Laboratory (No. 23XJ02005), the Central South University Research Programme of Advanced Interdisciplinary Studies (No. 2023QYJC020), and the Natural Science Foundation of Hunan Province (No. 2024JJ6530).

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

References

1. Akinci D’Antonoli, T., Berger, L.K., Indrakanti, A.K., Vishwanathan, N., Weiss, J., Jung, M., Berkarda, Z., Rau, A., Reisert, M., Küstner, T., et al.: Totalsegmentator mri: robust sequence-independent segmentation of multiple anatomic structures in mri. *Radiology* **314**(2), e241613 (2025)
2. ANTsX Team: ANTsPy: Advanced Normalization Tools in Python (2023), <https://github.com/ANTsX/ANTsPy>, accessed: 2023-10-25
3. Atli, O.F., Kabas, B., Arslan, F., Yurt, M., Dalmaz, O., Çukur, T.: I2i-mamba: Multi-modal medical image synthesis via selective state space modeling. *arXiv preprint arXiv:2405.14022* (2024)
4. Borges, P., Fernandez, V., Tudosi, P.D., Nachev, P., Ourselin, S., Cardoso, M.J.: Using mr physics for domain generalisation and super-resolution. In: *International Workshop on Simulation and Synthesis in Medical Imaging*. pp. 177–186. Springer (2024)
5. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y.: Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306* (2021)
6. Chu, Y., Yang, C., Luo, G., Qiu, Z., Gao, X.: Anatomic-constrained medical image synthesis via physiological density sampling. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. pp. 69–79. Springer (2024)
7. Dalmaz, O., Yurt, M., Çukur, T.: Resvit: residual vision transformers for multi-modal medical image synthesis. *IEEE Transactions on Medical Imaging* **41**(10), 2598–2614 (2022)

8. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. *Advances in neural information processing systems* **27** (2014)
9. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 1125–1134 (2017)
10. Jeong, S., Kim, Y., Lee, E., Sohn, K.: Memory-guided unsupervised image-to-image translation. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 6558–6567 (2021)
11. Ji, Z., Zou, B., Kui, X., Vera, P., Ruan, S.: Deform-mamba network for mri super-resolution. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. pp. 242–252. Springer (2024)
12. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part II* 14. pp. 694–711. Springer (2016)
13. Kingma, D.P.: Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013)
14. Kui, X., Xiao, Q., Li, Q., Ji, Z., Zhang, J., Zou, B.: Flip distribution alignment vae for multi-phase mri synthesis. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. pp. 208–218. Springer (2025)
15. Lee, H.Y., Tseng, H.Y., Huang, J.B., Singh, M., Yang, M.H.: Diverse image-to-image translation via disentangled representations. In: *Proceedings of the European conference on computer vision (ECCV)*. pp. 35–51 (2018)
16. Li, C., Chen, Z., Zhang, Y., Zhong, L., Yang, W.: Boosting medical image synthesis via registration-guided consistency and disentanglement learning. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. pp. 78–88. Springer (2025)
17. Li, J., Liu, J., Kuang, H., Yue, H., He, M., Wang, J.: Mico: Multiple instance learning with context-aware clustering for whole slide image analysis. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. pp. 376–385. Springer (2025)
18. Lou, M., Ying, H., Liu, X., Zhou, H.Y., Zhang, Y., Yu, Y.: Sdr-former: A siamese dual-resolution transformer for liver lesion classification using 3d multi-phase imaging. *Neural Networks* p. 107228 (2025)
19. MONAI, P.: Brats mri axial slices generative diffusion model. https://github.com/Project-MONAI/model-zoo/tree/dev/models/brats_mri_axial_slices_generative_diffusion (2024), accessed: 2024-05-22
20. Park, H., Noh, J., Ham, B.: Learning memory-guided normality for anomaly detection. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 14372–14381 (2020)
21. Phan, V.M.H., Xie, Y., Zhang, B., Qi, Y., Liao, Z., Perperidis, A., Phung, S.L., Verjans, J.W., To, M.S.: Structural attention: Rethinking transformer for unpaired medical image synthesis. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. pp. 690–700. Springer (2024)
22. Pinaya, W.H., Graham, M.S., Kerfoot, E., Tudosiu, P.D., Dafflon, J., Fernandez, V., Sanchez, P., Wolleb, J., Da Costa, P.F., Patel, A., et al.: Generative ai for medical imaging: extending the monai framework. *arXiv preprint arXiv:2307.15208* (2023)

23. Pinaya, W.H., Tudosi, P.D., Dafflon, J., Da Costa, P.F., Fernandez, V., Nachev, P., Ourselin, S., Cardoso, M.J.: Brain imaging generation with latent diffusion models. In: MICCAI Workshop on Deep Generative Models. pp. 117–126. Springer (2022)
24. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
25. Tian, H., Lv, T., Fan, J., Pan, D., Li, L., Pan, X.: Coarse-to-fine medical image translation by incorporating deterministic guidance and probabilistic refinement. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 128–137. Springer (2025)
26. 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)
27. Yu, Z., Zhao, B., Zhang, S., Chen, X., Yan, F., Feng, J., Peng, T., Zhang, X.Y.: Hifi-syn: Hierarchical granularity discrimination for high-fidelity synthesis of mr images with structure preservation. *Medical Image Analysis* **100**, 103390 (2025)
28. Zhang, X., He, X., Guo, J., Ettehadi, N., Aw, N., Semanek, D., Posner, J., Laine, A., Wang, Y.: Ptnet: A high-resolution infant mri synthesizer based on transformer. *arXiv preprint arXiv:2105.13993* (2021)
29. Zhang, Y., Yu, Z., Zhang, X., Zhang, S., Chen, X., Yang, H., Zhang, X.Y.: Disdiff: Disentanglement diffusion network for mr imaging translation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 153–163. Springer (2025)