Learning Modality-Invariant Structure and Distribution Alignment for Multimodal Medical Image Registration

Siqi Xiao†, Zetian Feng†, Quanling Zou, and Yi Wang ✉

Smart Medical Imaging, Learning and Engineering (SMILE) Lab, Medical UltraSound Image Computing (MUSIC) Lab, Guangdong Key Laboratory for Biomedical Measurements and Ultrasound Imaging, School of Biomedical Engineering, Shenzhen University Medical School, Shenzhen University, Shenzhen, China
onewang@szu.edu.cn

Abstract. Multimodal medical image registration remains challenging due to substantial cross-modal intensity discrepancies. While recent deep learning methods attempt to reduce modality gaps, they often underexploit modality-invariant anatomical structures or decouple representation alignment from deformation estimation. In this work, we propose a multimodal registration framework that jointly enhances shared anatomical representations and mitigates cross-modal discrepancies. Specifically, we introduce a Structure-Aware Frequency Enhancement (SAFE) module to decompose features into frequency components and selectively reinforce structure-sensitive high-frequency responses, promoting modality-invariant anatomical encoding. In addition, we develop a Registration-guided Multimodal Contrastive Alignment (RMCA) strategy that embeds contrastive learning into the registration process, enabling joint optimization of feature alignment and deformation estimation without requiring pre-aligned supervision. Extensive comparative and ablation experiments on three public datasets demonstrate that our method consistently outperforms state-of-the-art approaches, confirming its efficacy for multimodal medical image registration. *The code is available at* https://github.com/47-XIAO/MMreg.

Keywords: Multimodal Image Registration · Contrastive Learning · Frequency Enhancement · Unsupervised Deformable Registration

1 Introduction

Multimodal medical image registration aims to spatially align anatomical structures across different imaging modalities and is fundamental to numerous clinical applications, such as image-guided interventions, radiotherapy planning, and multi-parametric disease assessment [21]. However, due to heterogeneous imaging mechanisms, identical anatomical structures often exhibit markedly different intensity distributions, posing significant challenges for cross-modal registration.

† Siqi Xiao and Zetian Feng contributed equally to this work.

To address this challenge, recent deep learning methods have advanced multimodal registration at three levels. At the feature level, contrastive learning approaches [19,13,6,23] project multimodal images into a shared latent space to alleviate intensity discrepancies; however, many rely on additional supervision, such as pre-aligned image pairs for voxel-to-voxel contrastive learning, and are often decoupled from the registration process. At the image level, GAN-based strategies [26,3,8] mitigate intensity gaps via inter-modal translation, yet their performance is sensitive to data quality and quantity. At the model level, cross-modal interaction modules, such as cross-attention mechanisms, have been introduced to enhance inter-modal feature correlation [20,17,9], but frequently underexploit modality-invariant anatomical structures.

Despite substantial intensity variations, multimodal images preserve modality-invariant structural correspondence. Motivated by this insight, we propose an unsupervised framework for multimodal registration that jointly models modality-invariant structural representations and mitigates cross-modal intensity discrepancies within a unified optimization scheme. Extensive comparative and ablation experiments in three public multimodal datasets demonstrate the efficacy of the proposed method. The main contributions are summarized as follows:

- We propose a unified framework for cross-modal registration that jointly enhances shared anatomical structures and reduces modality-specific intensity variations, enabling robust multimodal alignment.
- We introduce a shared Structure-Aware Frequency Enhancement (SAFE) module to learn modality-invariant structural representations. SAFE decomposes features into low- and high-frequency components and selectively amplifies structure-sensitive high-frequency responses. Cross-modal parameter sharing enforces consistent structure-guided feature learning.
- We design a Registration-guided Multimodal Contrastive Alignment (RMCA) strategy to reduce cross-modal intensity discrepancies at the feature level. Positive pairs are dynamically constructed between warped moving and corresponding fixed features via the predicted deformation field, without pre-aligned supervision, enabling joint optimization of feature alignment and deformation estimation.

2 Method

2.1 Network Overview

The overall architecture is illustrated in Fig. 1(a). A dual-stream encoder extracts four-scale hierarchical features $\{F_i\}_{i=1}^4$ and $\{M_i\}_{i=1}^4$ for fixed and moving images. The encoders share architecture but not parameters, enabling modality-specific representation learning. Each encoder contains four blocks (Fig. 2(a)): the first preserves local structures without downsampling, while the remaining blocks progressively downsample and double channel dimensions to enlarge the receptive field and capture high-level semantics. Residual connections after

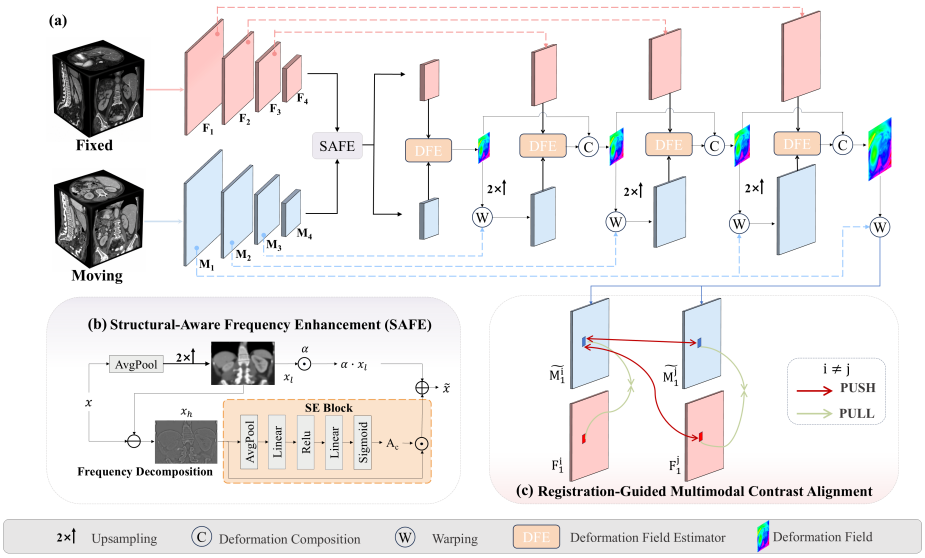

Fig. 1. Overview of the proposed multimodal registration network.

downsampling stabilize training. The deepest features F_4 and M_4 are further refined by SAFE to produce structure-enhanced representations F_4' and M_4'.

The decoder refines deformation fields in a coarse-to-fine manner. At scale i, F_i and the warped M_i are input into the Deformation Field Estimator (DFE) to predict the deformation update, with F_4' and M_4' used at the coarsest level ($i = 4$) and the deformation field $\phi_4 = \varphi_4$. For scale $i = 3, 2, 1$, the deformation is updated as:

$$\begin{cases} \phi_i' = up(\phi_{i+1}) \\ \varphi_i = DFE(F_i, M_i \circ \phi_i') \\ \phi_i = (\phi_i' \circ \varphi_i) + \varphi_i \end{cases} \tag{1}$$

where $up(\cdot)$ denotes tri-linear upsampling, and $\circ$ is the warp operation. As shown in Fig. 2(b), the DFE comprises two convolutional blocks (convolution, instance normalization, ReLU), a convolution layer (DFEConv) for deformation field estimation, and a diffeomorphism layer (Diff) [5] to enforce smoothness.

2.2 Structural-Aware Frequency Enhancement

We develop a SAFE module to strengthen modality-invariant representations and steer deformation estimation toward shared anatomical boundaries. As illustrated in Fig. 1(b), given the input feature map $x \in \mathbb{R}^{C \times D \times H \times W}$, where C is the channel number and D, H, W represent the spatial dimensions, SAFE first

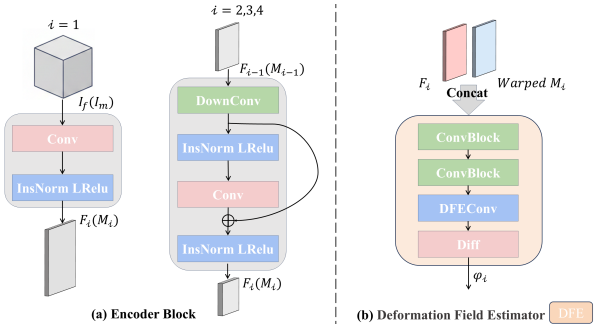

Fig. 2. Details of the encoder block and the deformation field estimator (DFE).

decomposes it into low- and high-frequency components:

$$\begin{aligned} x_l &= up(AvgPool(x)), \\ x_h &= x - x_l. \end{aligned} \tag{2}$$

While high-frequency components encode rich structural details [25,24,17], their contributions to registration are not equally informative. To emphasize consistent structural responses and suppress modality-specific noise, SAFE applies channel attention to the high-frequency branch. Specifically, a squeeze-and-excitation (SE) block [12] learns adaptive channel weights A_c, producing the refined component $A_c \cdot x_h$. The final representation is reconstructed as

$$\tilde{x} = A_c \cdot x_h + \alpha \cdot x_l, \tag{3}$$

where $\alpha = 0.2$ balances structural information and global semantic context.

SAFE is applied at the deepest scale because high-level features contain stronger modality-invariant semantic, making them more suitable for cross-modal enhancement. To encourage consistent structural enhancement across modalities, SAFE shares parameters between the two encoder branches, thus promoting modality-agnostic feature learning and improving cross-modal alignment.

2.3 Registration-Guided Multimodal Contrastive Alignment

The primary challenge in multimodal registration stems from substantial discrepancies in intensity distribution. To address this, we propose a RMCA strategy at the feature level, which reduces cross-modal representation gaps by enforcing alignment within a shared embedding space during deformation estimation.

Unlike most multimodal contrastive learning that rely on pre-aligned samples, RMCA dynamically constructs spatially aligned feature maps as positive pairs during training. RMCA is applied at the finest-resolution feature level because these features preserve richer spatial and anatomical details, while the learned alignment naturally propagates through the encoder-decoder architecture. Specifically, at the first encoder level, the predicted deformation ϕ_1 is applied to the moving feature M_1, producing the warped feature $\tilde{M}_1 = M_1 \circ \phi_1$. As

the deformation ϕ_1 establishes spatial correspondence, $\tilde{M}_1$ and the corresponding fixed feature F_1 are spatially aligned and treated as positive pairs, while other features in the batch serve as negatives, as shown in Fig. 1(c).

By embedding positive pair construction into the registration process, RMCA requires no additional supervision. As the deformation field progressively improves, spatial correspondence between modalities is continuously refined, leading to a mutually reinforced optimization between contrastive learning and deformation estimation. The contrastive loss is defined as:

$$\mathcal{L}_{cont} = -\frac{1}{2N}\sum_{i=1}^{2N}\log\left[\frac{\exp\left(sim(z_i, z_{p(i)})/\tau\right)}{\sum_{j=1,j\neq i}^{2N}\exp\left(sim(z_i, z_j)/\tau\right)}\right], \tag{4}$$

where $\tau = 0.5$ is the temperature parameter, z_i is the i-th sample feature, $z_{p(i)}$ is the cross-modal positive counterpart, and $sim(\cdot)$ is voxel-wise cosine similarity followed by global averaging to measure pairwise feature similarity.

The overall training objective is:

$$\mathcal{L} = \lambda_1\mathcal{L}_{MIND}(I_f, I_m \circ \phi) + \lambda_2\mathcal{L}_{cont}(F_1, M_1 \circ \phi) + \lambda_3\mathcal{L}_{reg}(\phi), \tag{5}$$

where λ_1, λ_2, and λ_3 balance the similarity, contrastive, and regularization loss terms, respectively. To account for the instability of the deformation field in early training, a ramp-up strategy is applied to λ_2:

$$\lambda_2(m) = e^{-5\left(1-\frac{m}{\tau}\right)^2}, \quad m \in [10, \tau], \tag{6}$$

where τ is the ramp-up length, m is the epoch number. For $m \in [1, 9]$, $\lambda_2(m) = 0$. The similarity term is computed using the modality independent neighborhood descriptor (MIND) [10], which is robust across different modalities. The regularization term $\mathcal{L}_{reg}$ is defined as L_2-$norm$ of the gradient of deformation field [4].

3 Experiments

Datasets. We evaluate the proposed method using three public datasets:

- **Learn2Reg (Abdomen)** [11]: It contains 8 paired and 89 unpaired MR-CT samples. Each case provides segmentation labels for four structures: liver, spleen, left kidney, and right kidney. We use 89 unpaired samples (40 MR, 49 CT) for training and 8 paired cases for testing. The data is resampled to a spacing of $2.0 \times 2.0 \times 2.0$ mm^3. All volumes are normalized and resized to $160 \times 160 \times 128$ via cropping and padding.
- **AMOS(Abdomen)** [14]: We select 30 annotated unpaired MR and CT images, split into training/test sets (24/6). For consistency, we use segmentation labels of the same four anatomical structures for the evaluation. The data is resampled to a spacing of $1.5 \times 1.5 \times 2.0$ mm^3. All volumes are normalized and resized to $208 \times 144 \times 128$ via cropping and padding.
- **SR-Reg (Brain)** [3]: We select 50 annotated MR-CT sample pairs and split them into training/validation/test sets (50/10/20). The data is resampled to a spacing of $1.0 \times 1.0 \times 1.0$ mm^3. All volumes are normalized and resized to a resolution of $144 \times 176 \times 160$ via cropping and padding.

Table 1. Comparative results of different methods on three public datasets.

Methods	Learn2Reg		AMOS		SR-Reg	
	DSC (%) ↑	$\%\lvert J_\phi\rvert \leq 0$ ↓	DSC (%) ↑	$\%\lvert J_\phi\rvert \leq 0$ ↓	DSC (%) ↑	$\%\lvert J_\phi\rvert \leq 0$ ↓
Initial	46.1±13.2	–	48.7±9.50	–	50.9±9.40	–
SyN [1]	53.9±15.4	$\mathbf{0.00 \times 10^{0}}$	59.2±13.0	$\mathbf{1.00 \times 10^{-7}}$	57.6±8.50	$\mathbf{1.00 \times 10^{-7}}$
VM [2]	53.3±14.2	2.99×10^{1}	56.4±8.10	2.35×10^{-3}	59.8±7.70	1.31×10^{-3}
TM [4]	52.4±16.9	1.30×10^{0}	58.0±8.10	6.57×10^{-2}	62.1±7.60	3.40×10^{-3}
FN [15]	53.8±15.4	3.03×10^{-1}	57.8±7.00	2.20×10^{-2}	63.0±7.40	5.55×10^{-2}
DGMIR [17]	52.0±16.4	2.71×10^{-1}	58.9±6.70	1.18×10^{-2}	62.5±8.30	3.94×10^{-3}
ARDMR [13]	51.4±18.3	5.52×10^{-2}	61.9±7.20	4.40×10^{-5}	64.5±7.50	9.00×10^{-6}
MM [22]	54.6±17.2	1.10×10^{-1}	58.0±7.60	8.70×10^{-1}	61.9±7.90	3.05×10^{-0}
Ours	**62.5±19.2**	2.00×10^{-3}	**69.4±8.10**	2.00×10^{-5}	**68.7±6.30**	6.10×10^{-5}

Evaluation Metrics. The Dice score (DSC) [7] is used to evaluate the degree of overlap between the corresponding regions. In addition, the quality of the predicted non-rigid deformation field is assessed by the percentage of voxels with non-positive Jacobian determinant. These evaluation metrics are calculated in 3D. A better registration shall have larger DSC, and smaller Jacobian.

Implementation Details. The proposed method is implemented using PyTorch and all experiments are conducted on an NVIDIA GeForce RTX 4090 GPU with 48GB memory. We use the Adam [16] optimizer with a learning rate decay strategy, which is defined as:

$$lr_m = lr_{init} \cdot (1 - (m-1)/M)^{0.9}, \quad m = 1, 2, \ldots, M \tag{7}$$

where lr_m denotes the learning rate at the m-th epoch, and $lr_{init} = 0.0001$ represents the initial learning rate. The batch size is 2. The model is trained for 30 epochs without RMCA and 40 epochs with RMCA. The weights in Eq 5 are set to $\lambda_1 = 1$ and $\lambda_3 = 1$. The ramp-up length τ is set to 40.

4 Results and Discussion

Comparative Results. We compare our method with representative registration approaches: SyN [1], a classical iterative algorithm; VoxelMorph (VM) [2], a U-Net-based model; TransMorph (TM) [4], which integrates a Swin Transformer [18] encoder; Fourier-Net (FN) [15], a band-limited model that learns displacement representations in the Fourier domain; and three recent multimodal frameworks, DGMIR [17], ARDMR [13] and MambaMorph (MM) [22]. For fair comparison, all models are retrained under a unified training pipeline and carefully tuned. All unimodal methods are trained using MIND loss.

As shown in Table 1, our method achieves the highest DSC across all three datasets while maintaining an extremely low folding ratios, indicating superior

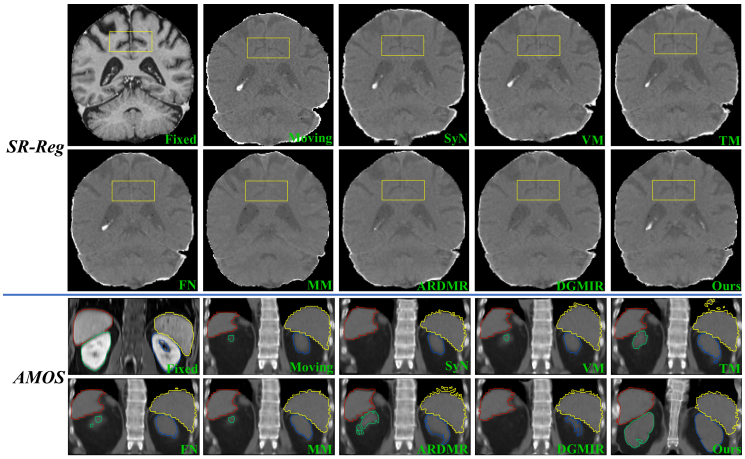

Fig. 3. Visual comparison of different methods on SR-Reg and AMOS. Color-coded contours in AMOS results indicate different organs (red-spleen; yellow-liver; blue-right kidney; green-left kidney).

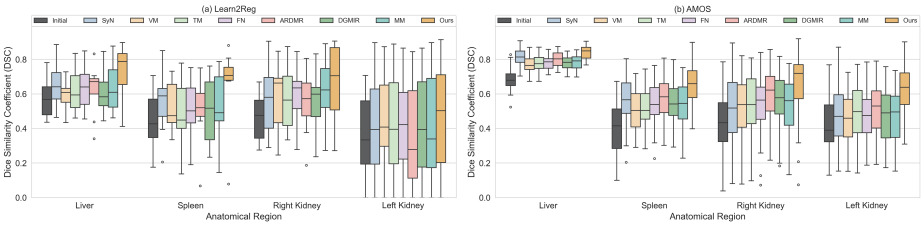

Fig. 4. The DSC distribution of different methods across various anatomical regions on (a) Learn2Reg and (b) AMOS.

accuracy and deformation regularity. On Learn2Reg, it improves DSC to 62.5%, surpassing the strongest baseline (MM, 54.6%) by 7.9%, with a markedly reduced folding ratio (2.00×10^{-3}). On AMOS, it achieves 69.4%, exceeding the best competing methods (ARDMR, 61.9%) by 7.5%, while preserving deformation smoothness (2.00×10^{-5}). On SR-Reg, it attains 68.7%, outperforming ARDMR (64.5%) by 4.2% and FN (63.0%) by 5.7%, again with a very low folding ratio (6.10×10^{-5}). Note that all these DSC improvements are statistically significant (t-test).

Fig. 3 visualizes two example registration results of different methods. Fig. 4 further shows the DSC distribution of different methods across various anatomical regions on Learn2Reg and AMOS. It can be observed that our method not only generates overall favorable performance but also consistently attains accurate and robust registration for each of the four abdominal organs.

Overall, unimodal architectures (VM, TM) show limited generalization in multimodal settings, often yielding inferior accuracy or unstable deformations.

Table 2. Ablation results of the proposed method.

		Learn2Reg		AMOS		SR-Reg	
SAFE	RMCA	DSC (%) ↑	$\%\lvert J_\phi\rvert \leq 0$ ↓	DSC (%) ↑	$\%\lvert J_\phi\rvert \leq 0$ ↓	DSC (%) ↑	$\%\lvert J_\phi\rvert \leq 0$ ↓
		52.3±16.7	5.13×10^{-4}	57.8±7.10	$\mathbf{0.00 \times 10^{0}}$	61.1±7.40	$\mathbf{0.00 \times 10^{0}}$
✓		55.4±16.8	$\mathbf{2.02 \times 10^{-4}}$	63.5±7.10	$\mathbf{0.00 \times 10^{0}}$	66.4±7.20	$\mathbf{0.00 \times 10^{0}}$
	✓	59.6±19.0	3.49×10^{-3}	68.6±8.20	2.20×10^{-5}	68.3±6.50	3.50×10^{-5}
✓	✓	**62.5±19.2**	2.00×10^{-3}	**69.4±8.10**	2.00×10^{-5}	**68.7±6.30**	6.10×10^{-5}

Frequency-based FN improves alignment compared with conventional CNN- or Transformer-based unimodal models, highlighting the advantage of frequency-domain deformation modeling for capturing global spatial correlations; however, it remains insufficient for complex multimodal discrepancies. Multimodal-specific frameworks (DGMIR, ARDMR, and MM) further improve performance, underscoring the importance of explicit cross-modal modeling. Nevertheless, their gains are moderate or accompanied by compromised deformation regularity, particularly for MM on SR-Reg and AMOS. In contrast, our method combines structure-aware frequency enhancement with registration-guided contrastive alignment, jointly strengthening multimodal feature consistency and deformation stability. This integration enables consistently superior accuracy while maintaining low folding ratios, achieving a better balance between alignment performance and topological preservation.

Ablation Study. We perform ablation experiments on three datasets to evaluate the contributions of SAFE and RMCA. The results are reported in Table 2.

The baseline achieves moderate performance (e.g., 52.3% DSC on Learn2Reg). Adding SAFE alone yields consistent gains (+3.1%, +5.7%, +5.3% DSC on Learn2Reg, AMOS, and SR-Reg) while maintaining extremely low folding ratios, indicating improved structural modeling without compromising deformation smoothness. RMCA alone produces larger improvements on DSC, particularly on Learn2Reg (+7.3%) and AMOS (+10.8%), demonstrating its effectiveness in reducing cross-modal discrepancies. Although the proportion of non-positive Jacobians slightly increases, it remains at a low magnitude. The full model achieves the best performance on all datasets, improving DSC by +10.2%, +11.6%, and +7.6% over the baseline, while preserving low folding ratios. These results indicate that SAFE and RMCA are complementary: SAFE enhances modality-invariant structural representations, and RMCA strengthens feature alignment, jointly enabling accurate and topology-preserving registration.

5 Conclusion

We propose a multimodal medical image registration framework that jointly enhances modality-invariant anatomical representations and alleviates cross-modal intensity discrepancies. The SAFE module reinforces structure-sensitive frequency

components to promote consistent anatomical encoding, while the RMCA strategy embeds contrastive learning into deformation estimation to align features within a shared space. Extensive comparative and ablation studies on three public datasets prove the effectiveness and robustness of the proposed approach.

Acknowledgments. This work was supported in part by the Shenzhen Medical Research Fund under Grant D2402010, in part by the National Natural Science Foundation of China under Grant 62471306, and in part by the Shenzhen Major Scientific and Technological Project under Grant KJZD20230923114615031.

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

References

1. Avants, B.B., Grossman, M., Gee, J.C.: Symmetric diffeomorphic image registration: Evaluating automated labeling of elderly and neurodegenerative cortex and frontal lobe. In: Biomedical Image Registration. pp. 50–57. Springer (2006)
2. Balakrishnan, G., Zhao, A., Sabuncu, M.R., Guttag, J., Dalca, A.V.: VoxelMorph: A learning framework for deformable medical image registration. IEEE Transactions on Medical Imaging **38**(8), 1788–1800 (2019)
3. Billot, B., Greve, D.N., Puonti, O., Thielscher, A., Van Leemput, K., Fischl, B., Dalca, A.V., Iglesias, J.E.: SynthSeg: Segmentation of brain MRI scans of any contrast and resolution without retraining. Medical Image Analysis **86**, 102789 (2023)
4. Chen, J., Frey, E.C., He, Y., Segars, W.P., Li, Y., Du, Y.: TransMorph: Transformer for unsupervised medical image registration. Medical Image Analysis **82**, 102615 (2022)
5. Dalca, A.V., Balakrishnan, G., Guttag, J., Sabuncu, M.R.: Unsupervised learning for fast probabilistic diffeomorphic registration. In: Medical Image Computing and Computer Assisted Intervention. pp. 729–738. Springer (2018)
6. Dey, N., Schlemper, J., Salehi, S.S.M., Zhou, B., Gerig, G., Sofka, M.: ContraReg: Contrastive learning of multi-modality unsupervised deformable image registration. In: Medical Image Computing and Computer Assisted Intervention. pp. 66–77. Springer (2022)
7. Dice, L.R.: Measures of the amount of ecologic association between species. Ecology **26**(3), 297–302 (1945)
8. Fan, J., Cao, X., Wang, Q., Yap, P.T., Shen, D.: Adversarial learning for mono-or multi-modal registration. Medical Image Analysis **58**, 101545 (2019)
9. Feng, Z., Ni, D., Wang, Y.: Salient region matching for fully automated MR-TRUS registration. In: 2025 IEEE 22nd International Symposium on Biomedical Imaging. pp. 1–5 (2025)
10. Heinrich, M.P., Jenkinson, M., Bhushan, M., Matin, T., Gleeson, F.V., Brady, S.M., Schnabel, J.A.: MIND: Modality independent neighbourhood descriptor for multi-modal deformable registration. Medical Image Analysis **16**(7), 1423–1435 (2012)
11. Hering, A., Hansen, L., Mok, T., Chung, A., Siebert, H., Hager, S., Lange, A., Kuckertz, S., Heldmann, S., Shao, W.: Learn2Reg: comprehensive multi-task medical image registration challenge, dataset and evaluation in the era of deep learning. IEEE Transactions on Medical Imaging **42**, 697–712 (2023)

12. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: IEEE Conference on Computer Vision and Pattern Recognition (2018)
13. Hu, Y., Zhang, Q., Zhao, Z., Ding, X., Li, W., Yang, J., Sun, J., Xu, L.X.: ARDMR: Adaptive recursive inference and representation disentanglement for multimodal large deformation registration. Medical Image Analysis **107**, 103844 (2026)
14. Ji, Y., Bai, H., Ge, C., Yang, J., Zhu, Y., Zhang, R., Li, Z., Zhanng, L., Ma, W., Wan, X., et al.: Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. Advances in Neural Information Processing Systems **35**, 36722–36732 (2022)
15. Jia, X., Bartlett, J., Chen, W., Song, S., Zhang, T., Cheng, X., Lu, W., Qiu, Z., Duan, J.: Fourier-net: Fast image registration with band-limited deformation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 1015–1023 (2023)
16. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. CoRR **abs/1412.6980** (2014)
17. Le, G., Shu, Y., Qiao, L., Yang, L., Xiao, B., Li, W., Gao, X.: DGMIR: dual-guided multimodal medical image registration based on multi-view augmentation and on-site modality removal. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 152–162. Springer (2025)
18. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: 2021 IEEE/CVF International Conference on Computer Vision. pp. 9992–10002 (2021)
19. Pielawski, N., Wetzer, E., Öfverstedt, J., Lu, J., Wählby, C., Lindblad, J., Sladoje, N.: CoMIR: Contrastive multimodal image representation for registration. In: Advances in Neural Information Processing Systems. vol. 33, pp. 18433–18444 (2020)
20. Song, X., Chao, H., Xu, X., Guo, H., Xu, S., Turkbey, B., Wood, B.J., Sanford, T., Wang, G., Yan, P.: Cross-modal attention for multi-modal image registration. Medical Image Analysis **82**, 102612 (2022)
21. Sotiras, A., Davatzikos, C., Paragios, N.: Deformable medical image registration: A survey. IEEE Transactions on Medical Imaging **32**, 1153–1190 (2013)
22. Wang, Y., Guo, T., Yuan, W., Shu, S., Meng, C., Bai, X.: Mamba-based deformable medical image registration with an annotated brain MR-CT dataset. Computerized Medical Imaging and Graphics **123**, 102566 (2025)
23. Xu, H., Yuan, J., Ma, J.: Murf: Mutually reinforcing multi-modal image registration and fusion. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(10), 12148–12166 (2023)
24. Yan, F., Jiang, X., Lu, Y., Cao, J., Chen, D., Xu, M.: Wavelet and prototype augmented query-based transformer for pixel-level surface defect detection. In: 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 23860–23869 (2025)
25. Zhou, Y., Huang, J., Wang, C., Song, L., Yang, G.: Xnet: Wavelet-based low and high frequency fusion networks for fully- and semi-supervised semantic segmentation of biomedical images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21085–21096 (2023)
26. Zhu, J., Zheng, B., Xiong, B., Zhang, Y., Cui, M., Sun, D., Cai, J., Xie, Y., Qin, W.: SynMSE: A multimodal similarity evaluator for complex distribution discrepancy in unsupervised deformable multimodal medical image registration. Medical Image Analysis **103**, 103620 (2025)