OTFusion: Optimal Transport-Driven 3D Multi-modal Medical Image Fusion for Enhanced Cross-Modality Representation

Xinjian Wei, Yafei Xiong, Chunyuan Chen, Haorui Yang, Haotian Lu, Mingyang Yu(✉), and Jing Xu(✉)

The National Key Laboratory of Intelligent Tracking and Forecasting for Infectious Diseases, College of Artificial Intelligence, Nankai University. No. 38, Tongyan Road, Haihe Education Park, Jinnan District. Tianjin, 300350, China
1120240312@mail.nankai.edu.cn, xujing@nankai.edu.cn

Abstract. Multi-modal medical image fusion aims to integrate complementary information across heterogeneous modalities into a unified and informative representation for clinical analysis. However, most existing approaches rely on heuristic or implicit feature aggregation strategies, lacking a principled mechanism to explicitly align cross-modal distributions, which often results in compromised structural consistency and semantic coherence. To address this, we propose a paradigm shift by reformulating multi-modal fusion task as a cross-domain distribution alignment problem, inspired by Optimal Transport (OT) framework. Specifically, we interpret fusion as seeking one minimum-cost matching between the source modalities and the idealized fused distribution, thereby providing a theoretically grounded criterion for cross-modal integration. Building upon this, we develop OTFusion, an end-to-end 3D framework driven by OT principles. OTFusion leverages the collaborative dual-network architecture, the U-Net-based mapping network that generates the fused representation, and the latent potential network estimates Kantorovich potentials to regularize the fusion process toward the optimal transport objective. Furthermore, we introduce a Cross-Dimensional Interaction Module (CroDIM) to explicitly capture complex inter-modal dependencies across spatial and feature dimensions, enabling effective multi-perspective information coupling. An adaptive optimization constraint is additionally incorporated to enhance cross-modal complementarity while preserving feature consistency. Experiments on various datasets demonstrate the OTFusion consistently outperforms state-of-the-art methods in both qualitative visualization and quantitative evaluation, validating its effectiveness and robustness for multi-modal medical image fusion.

Keywords: Multi-modal Medical Image Fusion · 3D Image Fusion · Neural Optimal Transport · Medical Image Analysis.

Xinjian Wei and Yafei Xiong are co-first authors.
Corresponding author: Mingyang Yu and Jing Xu.

1 Introduction

Multi-modal medical imaging provides complementary information for clinical analysis, where different modalities including various Magnetic Resonance Imaging (MRI) and Computed Tomography (CT) capture distinct anatomical structures and functional characteristics [1]. Integrating such heterogeneous information into one unified representation has therefore become an important role in medical image analysis, benefiting downstream tasks such as visualization, diagnosis, and treatment planning [21]. Multi-modal medical image fusion (MMIF) aims to synthesize an informative representation that preserves salient structures while maintaining modality-specific semantics [16]. Conventional methods fuse features via simple extraction and summation [8], but manual weighting often limits both fusion accuracy and computational efficiency. Deep learning-based methods enable end-to-end cross-modal fusion by leveraging convolutional feature learning [15], adversarial domain adaptation [19], and attention-based long-range modeling [17]. However, convolution is limited to local receptive fields, attention may cause uncontrolled cross-modal mixing, and adversarial strategies often rely on empirical objectives and suffer from instability, restricting the effectiveness and stability of MMIF.

Therefore, effective fusion remains challenging due to the inherent cross-modal distribution discrepancy. Different modalities often exhibit substantial variations in intensity statistics, contrast mechanisms, and semantic emphasis, making it difficult to simultaneously preserve structural consistency and semantic coherence. Most existing fusion methods adopt heuristic and implicit feature aggregation strategies [18], which may lead to unstable inter-modal interaction and insufficient distribution alignment, thereby producing fused results that either distort anatomical structures or fail to retain critical modality-specific information.

In this work, we introduce an optimal transport (OT)-inspired perspective by reformulating multi-modal fusion as a cross-domain distribution alignment problem. Specifically, we interpret fusion as seeking a minimum-cost coupling between source modality distributions and an idealized fused distribution, which provides a theoretically grounded criterion for cross-modal integration. Rather than explicitly computing transport plans, we leverage the Kantorovich dual formulation and estimate transport-induced potentials to regularize the fusion process toward OT-consistent alignment. Based on this formulation, we propose **OTFusion**, an end-to-end 3D fusion framework driven by OT principles. OTFusion adopts a collaborative dual-network architecture, consisting of a U-Net-based mapping network for generating fused representations and a latent potential network for estimating Kantorovich potentials to guide OT-based regularization. Furthermore, we introduce a Cross-Dimensional Interaction Module (CroDIM) to explicitly capture complex inter-modal dependencies across spatial and feature dimensions, enabling effective multi-perspective information coupling. An adaptive optimization constraint is additionally incorporated to enhance cross-modal complementarity while preserving feature consistency.

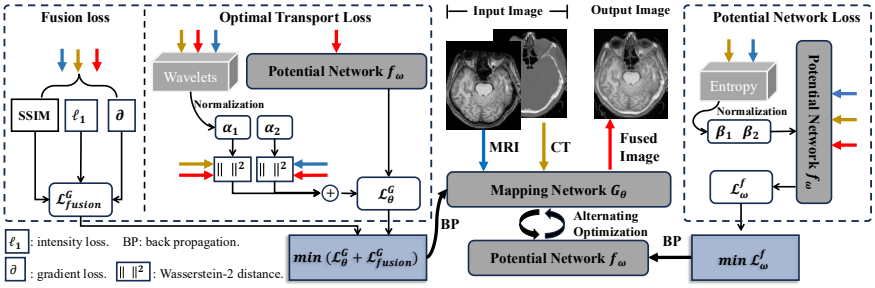

Fig. 1. Overview of the proposed OT-based optimization pipeline for multi-modal fusion.

Our contributions are summarized as follows: 1) We propose an OT-inspired formulation that reformulates multi-modal medical image fusion as a cross-domain distribution alignment problem. 2) We develop an end-to-end 3D OT-driven fusion framework (OTFusion) with a dual-network design based on Kantorovich potential regularization. 3) We introduce CroDIM and an adaptive optimization constraint to better model inter-modal dependencies and enhance complementarity while maintaining feature consistency. 4) Experiments on multiple datasets validate the effectiveness and robustness of the proposed.

2 Method

Given two 3D medical modalities $x, y \in \mathbb{R}^{C\times H\times W\times D}$, where C, H, W, and D denote the channel, height, width, and depth, respectively, we formulate multi-modal fusion as a distribution-guided mapping problem. Unlike supervised fusion, no explicit ground-truth fused target is available. Therefore, the fused representation is encouraged to align with an implicit distribution jointly induced by the source modalities. To achieve principled distribution alignment, we adopt an Optimal Transport (OT) perspective and optimize a coupled mapping–potential objective in an unsupervised manner. Specifically, OTFusion (see Fig. 1) consists of two collaborative components: (i) a mapping network G_θ that generates the fused output F_{OT},

$$F_{\mathrm{OT}} = G_\theta(x, y), \tag{1}$$

and (ii) a potential network f_ω that estimates the Kantorovich dual potential to quantify distributional discrepancy and provide gradient-based regularization for optimizing G_θ,

$$\nabla\mathcal{L}_\theta^G \leftarrow f_\omega(\cdot). \tag{2}$$

Optimal Transport: Different from classical OT formulations that explicitly optimize transport plans, the proposed OTFusion follows a neural OT paradigm

that performs implicit cross-modal distribution and alignment via learned transport mappings and Kantorovich potentials[4,3]. In OT, the Wasserstein-1 distance between two distributions admits the Kantorovich dual form [4]:

$$W_1(P,Q) = \sup_{\|f\|_L \leq 1} \left(\mathbb{E}_{u\sim P}[f(u)] - \mathbb{E}_{v\sim Q}[f(v)]\right), \tag{3}$$

where the optimal f is the Kantorovich potential. P and Q are the source and target distribution.

OT-guided Optimization Objective: Motivated by above formulation, we parameterize f using the potential network f_ω, and employ it to guide the fused distribution Q_f toward the implicit target distribution. Concretely, the mapping network G_θ is optimized using a joint objective that combines a W2-consistent quadratic cost to preserve modality-consistent structural information with the W1 dual potential f_ω to minimize the distributional OT discrepancy with respect to both input modalities:

$$\mathcal{L}_\theta^G = \frac{1}{|B|} \sum_{(x,y)\in B} \Big[\alpha_1 \|x - G_\theta(x,y)\|_2^2 + \alpha_2 \|y - G_\theta(x,y)\|_2^2 - f_\omega\big(G_\theta(x,y)\big)\Big], \tag{4}$$

where $|B|$ denotes the batch size, $\|\cdot\|^2$ defines the quadratic cost consistent with the Wasserstein-2 distance, and α_1, α_2 are modality-adaptive coefficients to balance structural contributions from each modality. Meanwhile, the potential network f_ω is optimized to approximate the OT dual potential by minimizing the discrepancy between the fused output and each modality distribution:

$$\mathcal{L}_\omega^f = \frac{1}{|B|} \sum_{(x,y)\in B} \Big[\beta_1\big(f_\omega(G_\theta(x,y)) - f_\omega(x)\big) + \beta_2\big(f_\omega(G_\theta(x,y)) - f_\omega(y)\big)\Big], \tag{5}$$

where β_1 and β_2 control the relative contributions of the two modalities in the dual optimization. The two networks are optimized in an asymmetric alternating manner: G_θ progressively generates fused outputs with improved distributional alignment, while f_ω refines the Kantorovich potential estimation. Notably, Q_f is not a fixed explicit target but an implicit distribution jointly induced by P_x and P_y. Guaranteed by the uniqueness of the Kantorovich potential for absolutely continuous distributions, our alternating optimization empirically converges to a stable joint distribution rather than acting merely as a distance regularizer.

Adaptive Weighting: To highlight structurally informative regions, the coefficients α_1 and α_2 are adaptively computed from the high-frequency energies E_x and E_y of the two modalities via discrete wavelet transform (DWT) [7]:

$$\alpha_1 = \frac{E_x}{E_x + E_y + \varepsilon}, \quad \alpha_2 = \frac{E_y}{E_x + E_y + \varepsilon}, \tag{6}$$

where $\varepsilon = 10^{-6}$ ensures numerical stability. Similarly, modality-wise information richness is quantified by information entropy [11], based on which β_1 and β_2 are computed as:

$$\beta_1 = \frac{H_x}{H_x + H_y + \varepsilon}, \quad \beta_2 = \frac{H_y}{H_x + H_y + \varepsilon}, \tag{7}$$

where H_x and H_y denote the average entropies of the two modalities.
Auxiliary Fusion Loss: The auxiliary constraints including structural similarity, intensity consistency, and gradient preservation are incorporated to further enhance structural consistency and semantic coherence [6]:

$$\mathcal{L}_{fusion}^{G} = (1 - \mathcal{L}_{ssim}) + \mathcal{L}_{int} + \mathcal{L}_{text}. \tag{8}$$

Total Loss: The final optimization objectives of OTFusion are given by:

$$\min_{\theta} \left(\mathcal{L}_{\theta}^{G} + \mathcal{L}_{fusion}^{G}\right), \min_{\omega} \mathcal{L}_{\omega}^{f}. \tag{9}$$

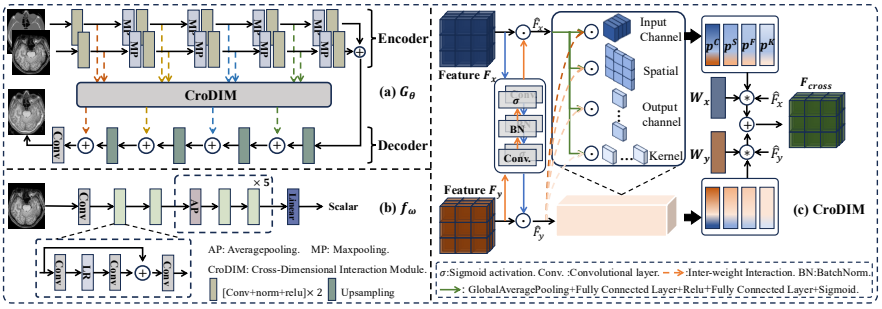

Fig. 2. Illustration of the detailed architecture. (a) The mapping network G_θ. (b) The potential network f_ω. (c) The structure of the proposed CroDIM, where W denotes the convolutional kernel.

Mapping network G_θ: The mapping network G_θ adopts a 3D encoder–decoder architecture to model the deterministic fusion mapping between two input modalities (see Fig. 2). Given inputs x and y, modality-specific features are first extracted through encoders:

$$F_x = \mathrm{Enc}(x), \quad F_y = \mathrm{Enc}(y), \tag{10}$$

where $\mathrm{Enc}(\cdot)$ consists of stacked 3D convolutions with progressive downsampling to capture multi-scale contextual representations. The extracted features preserve modality-specific structural cues and semantic saliency.
Cross-Dimensional Interaction Module (CroDIM): While parallel encoders capture complementary representations, effective fusion requires explicit cross-modal interaction to align and integrate salient information. To this end, we introduce CroDIM as the core fusion operator, which performs (i) bidirectional cross-modal conditioning and (ii) quad-dimensional adaptive convolution to model multi-perspective dependencies across spatial and feature dimensions.

CroDIM first performs bidirectional conditioning between modalities, allowing each modality to adaptively refine its feature representations according to

the complementary information from the other:

$$\hat{F}_x = F_x \odot \sigma(\mathrm{BN}(\mathrm{Conv}(F_y))), \quad \hat{F}_y = F_y \odot \sigma(\mathrm{BN}(\mathrm{Conv}(F_x))), \tag{11}$$

where $\sigma(\cdot)$ denotes the sigmoid function and $\odot$ indicates element-wise modulation with residual connection. Building upon the refined features, CroDIM further performs quad-dimensional adaptive convolution by generating dimension-specific attention weights along the spatial, input-channel, output-channel, and kernel dimensions [5]. These weights are then cross-modulated between modalities to obtain interacted weights, which are used to reconstruct adaptive convolutional kernels. Specifically, for each dimension $z \in \{S, C, F, K\}$ and kernel index $n \in 1, \ldots, N$, the attention weights $\delta^z_{m,n}$ from two modalities are interacted through element-wise multiplication followed by residual aggregation, enabling bidirectional modulation across modalities:

$$p^z_{x,n} = \delta^z_{x,n} + \delta^z_{x,n} \cdot \delta^z_{y,n}, \quad p^z_{y,n} = \delta^z_{y,n} + \delta^z_{x,n} \cdot \delta^z_{y,n}. \tag{12}$$

The interacted weights $p^z_{m,n}$ are used to construct adaptive convolutional kernels, which are applied to the modulated features $\tilde{F}_x$ and $\tilde{F}_y$ as follows:

$$\tilde{F}_m = \left(\sum_{n=1}^{N} p^z_{m,n} W_{m,n} \right) * \hat{F}_m, \quad m \in \{x, y\}. \tag{13}$$

The two cross-modal features are aggregated and fed into the decoder, where they are reconstructed in conjunction with the original features F_x and F_y via successive upsampling and skip connections to produce the final fusion output:

$$F_{OT} = \mathrm{Dec}\Big((\tilde{F}_x + \tilde{F}_y),\, F_x \oplus F_y \Big), \tag{14}$$

where $\mathrm{Dec}(\cdot)$ denotes the decoder in 3D U-Net backbone [3], and $\oplus$ represents element-wise addition.

Potential Network f_ω: The potential network f_ω, implemented with a ResNet backbone (see Fig. 2), approximates the Kantorovich dual in Wasserstein-1 OT, acting as a distribution-level critic that guides G_θ by measuring discrepancies between fused and source distributions.

3 Experiments

Datasets and Implementation Details: Extend experiments are evaluated on the publicly available 3D medical datasets: SynthRAD [12] and BraTS[1]. All volumes are center-cropped to fixed sizes of $(160, 160, 192)$ for SynthRAD and $(160, 192, 144)$ for BraTS. For SynthRAD, 160 paired CT–MRI volumes are used for training and 20 for testing. For BraTS, 150 paired MRI-T1 and MRI-T2 volumes are used for training and 20 for testing. All experiments are implemented

[1] https://www.med.upenn.edu/cbica/brats/

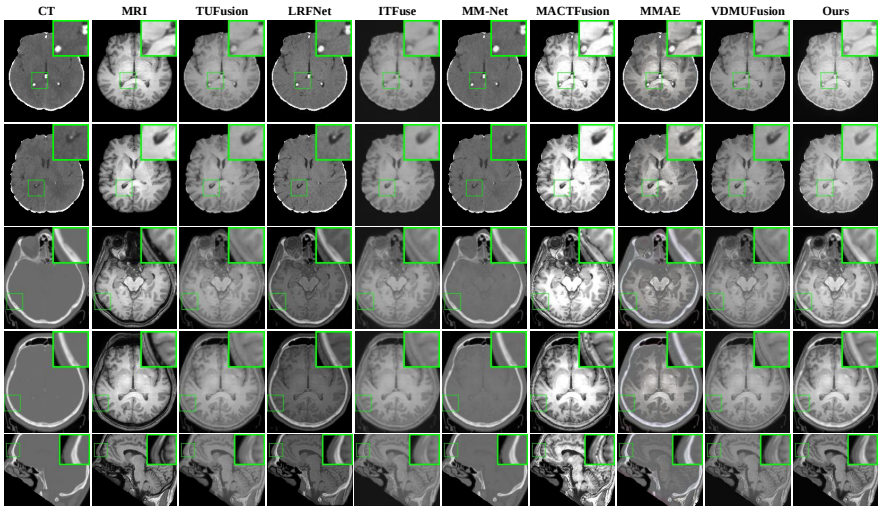

Fig. 3. Qualitative comparison of the proposed with the state-of-the-art methods on *SynthRAD* dataset.

in PyTorch and conducted on an H20-NVLink GPU. The network is optimized using Adam with an initial learning rate of 2×10^{-4} and a batch size of 32.

Compared Methods and Metrics: The proposed is compared with several state-of-the-art multi-modal fusion approaches, including TUFusion [20], LRFNet [2], ITFuse [10], MM-Net [13], MACTFusion [18], MMAE [14], and VDMUFusion [9]. All methods are evaluated using the official implementations. Fusion performance is assessed using widely adopted quantitative metrics: $SSIM$, Q_G, Q_{TE}, NCC, LMI, Q_{NCIE}, and Q_P.

Results: Qualitative results on SynthRAD are illustrated in Fig. 3. Most existing approaches exhibit evident modality preference. TUFusion, ITFuse and VDMUFusion effectively preserve MRI structural information but compromise CT intensity consistency, leading to attenuated contrast in high-density regions. In contrast, LRFNet and MM-Net emphasizes CT intensity characteristics at the expense of structural sharpness, resulting in blurred anatomical boundaries. Although MACTFusion and MMAE present relatively balanced visual appearances, they still suffer from contrast bias or subtle texture degradation. Compared with these methods, the proposed OTFusion achieves the faithful integration of complementary information: it preserves salient CT intensity characteristics while simultaneously maintaining fine-grained MRI textures and structural continuity, producing visually consistent and natural fusion results. Quantitative comparisons are reported in Table 1. OTFusion attains the best or the suboptimal performance on both datasets, demonstrating strong overall competitiveness. In particular, higher Q_{NCIE} and LMI values reflect effective cross-modal information interaction, while consistently strong NCC and $SSIM$ scores indicate improved structural similarity and spatial consistency with the source

Table 1. Quantitative Comparison between OTFusion and other Methods. The Best and Suboptimal are highlighted in Red and Blue.

Dataset	Method	$Q_{NCIE}\uparrow$	$Q_{TE}\uparrow$	$Q_G\uparrow$	$Q_P\uparrow$	$LMI\uparrow$	$NCC\uparrow$	$SSIM\uparrow$
SynthRAD	TUFusion	0.8167	1.0294	0.3815	0.4772	0.6908	0.3148	0.4200
	LRFNet	0.8136	1.0563	0.4069	0.4394	0.6668	0.2784	0.3642
	ITFuse	0.8131	1.0889	0.2375	0.3660	0.6676	0.2713	0.2894
	MM-Net	0.8154	0.9794	0.4323	0.4256	0.6950	0.3072	0.3780
	MACTFusion	0.8129	1.0923	0.5914	0.4392	0.6590	0.2658	0.3974
	MMAE	0.8164	1.0081	0.5210	0.4496	0.7037	0.3122	0.4049
	VDMUFusion	0.8161	1.1053	0.3943	0.4953	0.7020	0.3074	0.3608
	Ours	0.8167	1.1001	0.6642	0.5107	0.7087	0.3154	0.4233
BraTS	TUFusion	0.8108	0.9169	0.4112	0.5669	0.6802	0.2368	0.6421
	LRFNet	0.8084	0.8772	0.4646	0.2796	0.7116	0.2005	0.6166
	ITFuse	0.8103	1.2726	0.2453	0.5133	0.6416	0.2288	0.2223
	MM-Net	0.8090	0.8723	0.4704	0.2941	0.7253	0.2138	0.6578
	MACTFusion	0.8112	0.9112	0.4793	0.5679	0.7400	0.2448	0.6526
	MMAE	0.8105	0.9249	0.4607	0.5394	0.7073	0.2409	0.5791
	VDMUFusion	0.8110	1.2620	0.4450	0.5784	0.6936	0.2391	0.2937
	Ours	0.8112	1.2847	0.4794	0.5795	0.7284	0.2412	0.6594

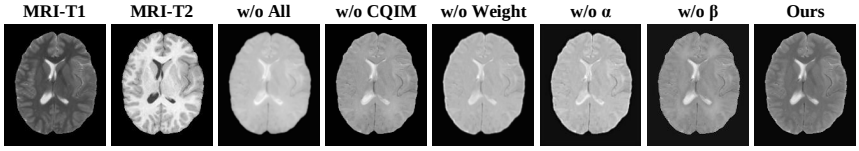

Fig. 4. Qualitative visualization of ablation study results on the *BraTS*.

Table 2. Evaluation results of the ablation study. "w/o" denotes without.

Name	CroDIM	wavelet	entropy	$Q_{NCIE}\uparrow$	$Q_{TE}\uparrow$	$Q_G\uparrow$	$Q_P\uparrow$	$LMI\uparrow$	$NCC\uparrow$	$SSIM\uparrow$
w/o All	–	–	–	0.8093	1.2605	0.4612	0.5016	0.6243	0.2214	0.6400
w/o CroDIM	–	✓	✓	0.8096	1.2675	0.4653	0.5227	0.6630	0.2242	0.6416
w/o Weight	✓	–	–	0.8097	1.2774	0.4751	0.5512	0.6830	0.2354	0.6519
w/o β	✓	✓	–	0.8105	1.2793	0.4754	0.5631	0.6942	0.2397	0.6561
w/o α	✓	–	✓	0.8102	1.2769	0.4735	0.5559	0.6857	0.2395	0.6526
Ours	✓	✓	✓	0.8112	1.2847	0.4794	0.5795	0.7284	0.2412	0.6594

modalities. Moreover, superior Q_G and Q_P results further confirm enhanced edge preservation and perceptual quality. Overall, OTFusion maintains the balanced and stable performance across diverse criteria, highlighting its robustness and comprehensive fusion capability. In future work, extended experiments will be conducted to further validate its practical utility through downstream clinical tasks (e.g., brain tumor segmentation on the BraTS dataset).

Ablation Study: Ablation studies on the BraTS dataset (Table 2, Fig. 4) evaluate the contributions of CroDIM in the mapping network and the adaptive weighting strategy. Removing CroDIM degrades most metrics, especially Q_{TE}, Q_P, and LMI, and produces blurred boundaries with weaker textures, highlighting its role in cross-modal feature interaction and structural preservation. Excluding the wavelet-based α or entropy-based β similarly reduces performance,

with the largest drop when both are removed, confirming that adaptive weighting effectively balances modality contributions and preserves information.

4 Conclusion

In this paper, we propose OTFusion, an end-to-end Optimal Transport–driven framework for multi-modal 3D medical image fusion. By reformulating fusion as a cross-modal distribution alignment problem, OTFusion provides a principled mechanism to guide the fused representation toward structurally consistent and semantically coherent integration. The proposed CroDIM enables explicit cross-dimensional interaction, while the adaptive optimization strategy further improves modality complementarity and feature consistency. Extensive experiments across multiple datasets demonstrate that OTFusion achieves consistent and robust improvement.

Acknowledgments. This research is supported by Tianjin Science and Technology Major Project under the grant No. 24ZXZSSS00140.

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

References

1. Azam, M.A., Khan, K.B., Salahuddin, S., Rehman, E., Khan, S.A., Khan, M.A., Kadry, S., Gandomi, A.H.: A review on multimodal medical image fusion: Compendious analysis of medical modalities, multimodal databases, fusion techniques and quality metrics. Computers in Biology and Medicine **144**, 105253 (2022). https://doi.org/https://doi.org/10.1016/j.compbiomed.2022.105253, https://www.sciencedirect.com/science/article/pii/S0010482522000452
2. He, D., Li, W., Wang, G., Huang, Y., Liu, S.: Lrfnet: A real-time medical image fusion method guided by detail information. Computers in Biology and Medicine **173**, 108381 (2024). https://doi.org/https://doi.org/10.1016/j.compbiomed.2024.108381, https://www.sciencedirect.com/science/article/pii/S0010482524004657
3. Kim, B., Zhuang, Y., Mathai, T.S., Summers, R.M.: Otmorph: Unsupervised multi-domain abdominal medical image registration using neural optimal transport. IEEE Transactions on Medical Imaging **44**(1), 165–179 (2025). https://doi.org/10.1109/TMI.2024.3437295
4. Korotin, A., Selikhanovych, D., Burnaev, E.: Neural optimal transport. arXiv preprint arXiv:2201.12220 (2022)
5. Li, C., Zhou, A., Yao, A.: Omni-dimensional dynamic convolution. arXiv preprint arXiv:2209.07947 (2022)
6. Ma, J., Tang, L., Fan, F., Huang, J., Mei, X., Ma, Y.: Swinfusion: Cross-domain long-range learning for general image fusion via swin transformer. IEEE/CAA Journal of Automatica Sinica **9**(7), 1200–1217 (2022). https://doi.org/10.1109/JAS.2022.105686

7. Nawaz, S.A., Li, J., Shoukat, M.U., Bhatti, U.A., Raza, M.A.: Hybrid medical image zero watermarking via discrete wavelet transform-resnet101 and discrete cosine transform. Computers and Electrical Engineering **112**, 108985 (2023). https://doi.org/https://doi.org/10.1016/j.compeleceng.2023.108985, https://www.sciencedirect.com/science/article/pii/S0045790623004093
8. Shehanaz, S., Daniel, E., Guntur, S.R., Satrasupalli, S.: Optimum weighted multimodal medical image fusion using particle swarm optimization. Optik **231**, 166413 (2021). https://doi.org/https://doi.org/10.1016/j.ijleo.2021.166413, https://www.sciencedirect.com/science/article/pii/S0030402621001467
9. Shi, Y., Liu, Y., Cheng, J., Wang, Z.J., Chen, X.: Vdmufusion: A versatile diffusion model-based unsupervised framework for image fusion. IEEE Transactions on Image Processing **34**, 441–454 (2025). https://doi.org/10.1109/TIP.2024.3512365
10. Tang, W., He, F., Liu, Y.: Itfuse: An interactive transformer for infrared and visible image fusion. Pattern Recognition **156**, 110822 (2024). https://doi.org/https://doi.org/10.1016/j.patcog.2024.110822, https://www.sciencedirect.com/science/article/pii/S0031320324005739
11. Tang, W., He, F., Liu, Y., Duan, Y.: Matr: Multimodal medical image fusion via multiscale adaptive transformer. IEEE Transactions on Image Processing **31**, 5134–5149 (2022). https://doi.org/10.1109/TIP.2022.3193288
12. Thummerer, A., van der Bijl, E., Galapon Jr, A., Verhoeff, J.J., Langendijk, J.A., Both, S., van den Berg, C.N.A., Maspero, M.: Synthrad2023 grand challenge dataset: Generating synthetic ct for radiotherapy. Medical physics **50**(7), 4664–4674 (2023)
13. Triaridis, K., Doumanidis, C., Chatzidiamantis, N.D., Karagiannidis, G.K.: Mm-net: A multi-modal approach toward automatic modulation classification. IEEE Communications Letters **28**(2), 328–331 (2024). https://doi.org/10.1109/LCOMM.2023.3342604
14. Wang, X., Fang, L., Zhao, J., Pan, Z., Li, H., Li, Y.: Mmae: A universal image fusion method via mask attention mechanism. Pattern Recognition **158**, 111041 (2025). https://doi.org/https://doi.org/10.1016/j.patcog.2024.111041, https://www.sciencedirect.com/science/article/pii/S0031320324007921
15. Wei, X., Qiu, Y., Xu, J., Zhang, J.: Hcfmanet: A novel holistic cross-modal fusion mamba network for multi-modal medical image fusion. IEEE Transactions on Multimedia pp. 1–15 (2026). https://doi.org/10.1109/TMM.2026.3660130
16. Wei, X., Qiu, Y., Xu, X., Xu, J., Mei, J., Zhang, J.: Ecinfusion: A novel explicit channel-wise interaction network for unified multi-modal medical image fusion. IEEE Transactions on Circuits and Systems for Video Technology **35**(5), 4011–4025 (2025). https://doi.org/10.1109/TCSVT.2024.3516705
17. Wei, X., Xiong, Y., Lu, H., Xu, X., Xu, J.: Xkanfuse: A novel cross-modal fusion method based on kolmogorov-arnold network for multi-modal medical image fusion. Knowledge-Based Systems **326**, 114053 (2025). https://doi.org/https://doi.org/10.1016/j.knosys.2025.114053, https://www.sciencedirect.com/science/article/pii/S0950705125010986
18. Xie, X., Zhang, X., Tang, X., Zhao, J., Xiong, D., Ouyang, L., Yang, B., Zhou, H., Ling, B.W.K., Teo, K.L.: Mactfusion: Lightweight cross transformer for adaptive multimodal medical image fusion. IEEE Journal of Biomedical and Health Informatics **29**(5), 3317–3328 (2025). https://doi.org/10.1109/JBHI.2024.3391620
19. Zhang, Z., Wang, X., Wang, Z., Xu, J., Yang, F.: Pfi-fuse: Parallel frequency-invertible adversarial network for infrared and visible image fusion. Knowledge-Based Systems **330**, 114538 (2025). https://doi.org/https://doi.org/

10.1016/j.knosys.2025.114538, https://www.sciencedirect.com/science/article/pii/S0950705125015771

20. Zhao, Y., Zheng, Q., Zhu, P., Zhang, X., Ma, W.: Tufusion: A transformer-based universal fusion algorithm for multimodal images. IEEE Transactions on Circuits and Systems for Video Technology **34**(3), 1712–1725 (2024). https://doi.org/10.1109/TCSVT.2023.3296745
21. Zhou, T., Cheng, Q., Lu, H., Li, Q., Zhang, X., Qiu, S.: Deep learning methods for medical image fusion: A review. Computers in Biology and Medicine **160**, 106959 (2023). https://doi.org/https://doi.org/10.1016/j.compbiomed.2023.106959, https://www.sciencedirect.com/science/article/pii/S0010482523004249