Hybrid State-Space Models and Denoising Training for Unpaired Medical Image Synthesis

Junming Zhang[1] and Shancheng Jiang[1,2,⋆]

[1] School of Intelligent Systems Engineering, Sun Yat-sen University, Shenzhen 518107, China
[2] Guangdong Provincial Key Laboratory of Fire Science and Intelligent Emergency Technology, Shenzhen 518107, China

Abstract. Unsupervised medical image synthesis faces significant challenges due to the absence of paired data, often resulting in global anatomical distortions and local detail loss. Existing approaches primarily rely on convolutional neural networks (CNNs) for local feature extraction; however, their limited receptive fields hinder effective global anatomical modeling. Recently, Vision Mamba (ViM) has demonstrated efficient global modeling capabilities via state-space models, yet its potential in this task remains unexplored. To address this gap, we propose a hybrid architecture, CRAViM (Convolutional Residual Attention Vision Mamba), which integrates the precise local anatomical feature extraction of CNNs with the long-range dependency modeling of state-space models, thereby enhancing the structural fidelity and detail preservation of synthesized images. Furthermore, we introduce a cycle denoise consistency-based training framework that incorporates transport loss and random denoise loss to jointly optimize global structural constraints and local detail restoration. Experimental results on two public medical imaging datasets demonstrate that CRAViM achieves notable improvements in key metrics such as SSIM and NMI over existing methods, effectively maintaining global anatomical consistency while enhancing local details, thus validating the effectiveness of our approach. The code for this paper can be found at https://github.com/jmzhang-cv/CRAViM.

Keywords: Mamba · State Space Model · Medical Image Synthesis · Unsupervised Image Translation.

1 Introduction

Multimodal medical imaging significantly enhances lesion detection accuracy through integration of heterogeneous imaging data [3], yet faces three major bottlenecks: radiation accumulation, cross-modal registration errors, and exponentially increasing costs [15,13]. Unsupervised medical image synthesis technology addresses these challenges by generating multimodal images from single-scan acquisitions. This approach not only eliminates radiation exposure from repeated

⋆ Corresponding author

scans but also produces anatomically consistent multimodal data, thereby removing registration-induced interference in quantitative analysis. The synthesized images establish a robust data foundation for multiscale feature fusion and pathophysiological mechanism investigation [10]. This technological advancement drives the evolution of medical image analysis from superficial feature observation to deep pathological mechanism exploration, providing novel pathways for developing intelligent diagnostic systems.

Existing unsupervised methods primarily rely on the local feature extraction capability of Convolutional Neural Networks (CNNs) [22,20]. However, their limited receptive field makes it difficult to model long-range spatial dependencies, leading to distortion of global anatomical structures in synthesized images [21,7]. While Vision Transformer (ViT) improves remote modeling through global attention mechanisms [6,4], it still faces a dual challenge in unpaired medical image synthesis: the lack of local inductive bias leads to loss of fine details, and it is difficult to learn fine-grained features under unsupervised conditions [16]. While recent studies show that Mamba-based architectures demonstrate significant advantages in medical image analysis tasks such as classification and segmentation [1], their architectural designs are often not optimized for high-fidelity generation. Consequently, the potential of Vision Mamba (ViM) [23] in the distinct task of unpaired image synthesis has yet to be fully explored.

In this paper, we propose a novel hybrid network, CRAViM (Convolutional Residual Attention Vision Mamba), which innovatively integrates state-space models into unpaired medical image synthesis tasks. This architecture retains the precise local feature extraction capabilities of traditional convolutional neural networks (CNNs) while leveraging state-space models to establish anatomy-aware global dependencies, thereby enhancing the structural consistency of synthesized images. Concurrently, we design a noise-consistency training framework that synergistically optimizes global structural constraints and local detail restoration by combining transport loss with random denoise loss, further improving the model's fidelity in preserving fine-grained details.

Our contributions can be summarized as follows. **1)** We propose CRAViM, a hybrid network that precisely extracts local anatomical features via convolutional operations and models long-range dependencies with state-space models, thereby enhancing the anatomical fidelity and detail restoration of synthesized unpaired medical images. **2)** We design a training framework based on noise-consistency modeling, which effectively improves the structural consistency and global visual quality of synthesized images. **3)** We introduce transport loss as a novel loss function constraint strategy, indirectly optimizing the quality of synthesized images to mitigate potential conflicts among multiple loss functions, thus enhancing the model's stability and generalization capability. **4)** We pioneer the integration of the Mamba structure into unpaired medical image translation tasks, offering a novel research direction for the field and advancing the development of related tasks.

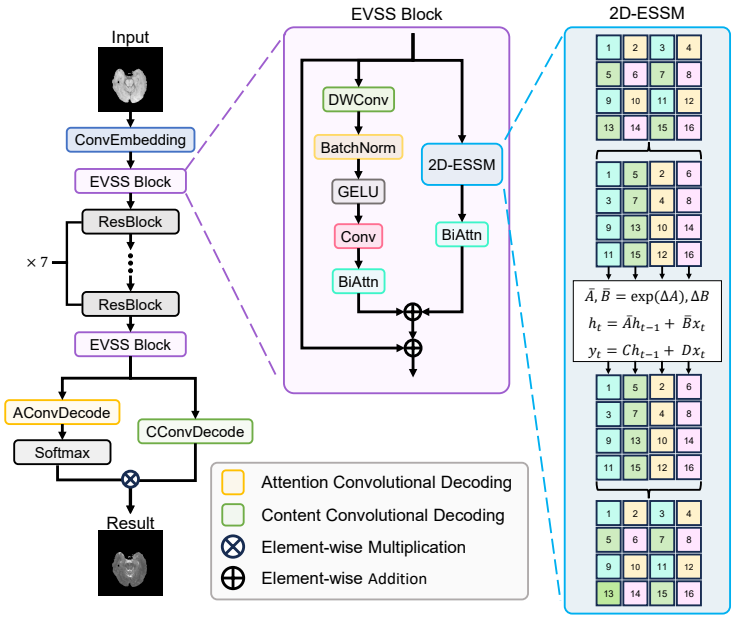

Fig. 1. The architecture of the CRAVIM network. The 2D-ESSM component utilizes a state-space model (SSM) for modeling, and its governing equations are illustrated in the figure.

2 Method

2.1 CRAViM

To address the dual challenges of complex anatomical topology and modality-specific characteristics in medical images, we propose an innovative hybrid architecture, CRAViM (Fig. 1). This architecture integrates the advantages of convolutional operations and state space models, effectively resolving the trade-off dilemma between global structural preservation and local detail reconstruction that commonly exists in traditional methods.

ConvEmbedding The convolutional embedding module uses a three-stage progressive downsampling structure (Eq. 1). To overcome the locality limitation of conventional convolutions, we adapt the EVSS Block [14], whose 2D-Efficient Selective Scan Module (2D-ESSM) establishes long-range dependencies for global context modeling via a selective scanning mechanism.

$$F_{l+1} = \mathrm{ReLU}(\mathrm{IN}(\mathrm{Conv}(F_l))). \tag{1}$$

Information Bottleneck The bottleneck layer constructs a deep feature refinement network by cascading residual blocks [8]. The multi-scale feature fusion

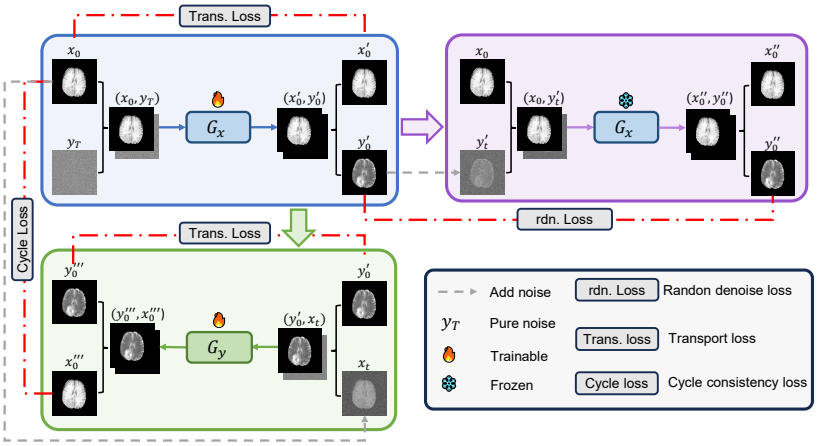

Fig. 2. The proposed cycle denoise consistency framework. The forward (G_x) and backward (G_y) paths establish the cycle. The top right shows the calculation of the random denoise loss (L_{rdn}), which enhances robustness.

mechanism in this design effectively preserves anatomical features across different spatial resolutions, such as gray matter boundaries in MRI images.

Convolutional Decoder The decoding stage adopts a collaborative design of transpose convolutions and attention mechanisms. First, the EVSS Block integrates global context information, followed by a mixed spatial-channel attention gating mechanism:

$$F_{out} = \sigma(F_{attention}) \odot F_{content}, \tag{2}$$

where the attention weight map is normalized by the sigmoid activation function $\sigma(\cdot)$, enhancing the focus on lesion regions.

Thus, the design of CRAViM has the following advantages: (1) the cascade architecture of state-space models and convolutional operations effectively integrates local and global features; (2) multi-level feature extraction retains complete spatial structural information; (3) the jump sampling scanning strategy fully utilizes the structural information extracted by the progressive feature fusion strategy while compensating for the computational efficiency of multi-level feature extraction.

2.2 Training Framework

To address the insufficient modeling capacity of existing unsupervised frameworks—often resulting in anatomical structure distortion and feature confusion—we propose a cycle denoise consistency framework (Fig. 2). This architecture synergistically integrates denoising modeling with cycle consistency constraints. The core idea is to leverage noise as a training signal to decouple and

stabilize the optimization process, thereby alleviating the conflict between preserving global structures and recovering fine-grained details.

To achieve this goal, we introduce two dedicated loss functions. First, to preserve global anatomical structures under the unpaired setting, we design a symmetric cross-modal transport loss ($\mathcal{L}_{\text{trans}}$). This loss acts as an indirect structural constraint on both generators, encouraging them to function as identity mappings for their respective input images when conditioned on specific noise inputs. It consists of both a forward and a backward component. Given the forward pass outputs $(x'_0, y'_0) = G_x(x_0, y_T)$ and the subsequent reconstruction pass for the intermediate image $(y'''_0, x'''_0) = G_y(y'_0, x_t)$, the total transport loss is defined as:

$$\mathcal{L}_{\text{trans}} = \mathbb{E}_{x_0 \sim p_{\mathcal{X}}}[\|x'_0 - x_0\|_1 + \|y'''_0 - y'_0\|_1], \tag{3}$$

Second, to enhance the model's robustness and enforce the learning of noise-invariant anatomical representations, we propose a random denoise loss ($\mathcal{L}_{\text{rdn}}$). This mechanism trains the generator G_x to remove randomly added noise from its own synthetic output y'_0. This process, illustrated in the top right of Fig. 2, forces the generator to distinguish authentic anatomical features from noise-induced artifacts. Given the denoising outputs $(x''_0, y''_0) = G_x(x_0, y'_t)$, the loss is defined as:

$$\mathcal{L}_{\text{rdn}} = \mathbb{E}_{x_0 \sim p_{\mathcal{X}}}[\|y''_0 - y'_0\|_1], \tag{4}$$

Here, y'_t is obtained by applying random intensity noise to the generated sample y'_0.

These two loss functions work in a complementary fashion: $\mathcal{L}_{\text{trans}}$ provides a stable anchor for global structural preservation, while $\mathcal{L}_{\text{rdn}}$ focuses on enhancing the fidelity and robustness of local features. This hybrid optimization strategy, combined with the standard adversarial loss ($\mathcal{L}_{\text{adv}}$) and cycle consistency loss ($\mathcal{L}_{\text{cyc}}$), forms our final training objective:

$$\min_{G_x, G_y} \max_{D_{\mathcal{X}}, D_{\mathcal{Y}}} \mathcal{L}_{\text{total}} = \lambda_{\text{adv}}\mathcal{L}_{\text{adv}} + \lambda_{\text{cyc}}\mathcal{L}_{\text{cyc}} + \lambda_{\text{trans}}\mathcal{L}_{\text{trans}} + \lambda_{\text{rdn}}\mathcal{L}_{\text{rdn}}. \tag{5}$$

3 Experiments

3.1 Datasets and Experimental Setup

We conducted experiments on two publicly available neuroimaging datasets: the IXI dataset [2] and the Brats2018 dataset [11]. Cross-sectional slices were extracted from each dataset and resized to 256×256 pixels for input into the deep learning models. To ensure the generalization capability of our models, each dataset was randomly partitioned by patient ID into training, validation, and test sets with a ratio of 25:5:10.

For training, we trained our models for 100 epochs using the Adam optimizer with a learning rate of 2×10^{-4}. Cosine Annealing was employed for performance

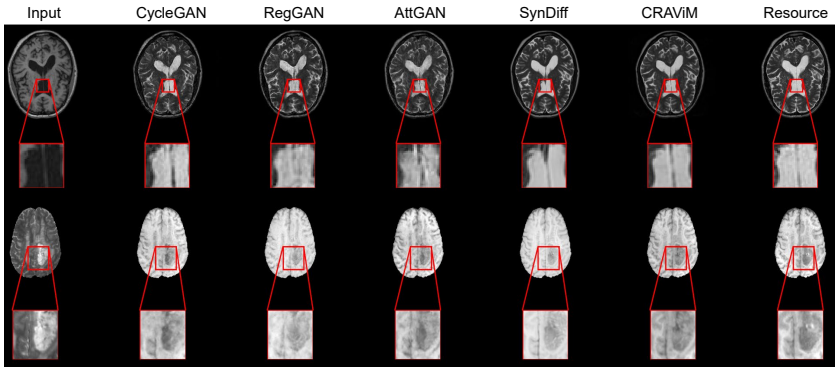

Fig. 3. Visual results of different methods. The first row depicts the T1 → T2 translation on the IXI dataset, while the second row illustrates the T2 → T1 translation on the BraTS2018 dataset.

optimization. The loss weights were set as $\lambda_{\text{adv}} = 1$, $\lambda_{\text{cyc}} = 15$, $\lambda_{\text{trans}} = 1$, and $\lambda_{\text{rdn}} = 5$. All experiments were conducted on NVIDIA RTX 3090 GPUs with a batch size of 4.

For evaluation, we adopted a combination of SSIM [18], NMI, LPIPS [19], and MSE, as recommended by a recent study [5]. Although pixel-wise metrics such as SSIM and MSE provide valuable information on structural similarity and error, they are sensitive to minor spatial misalignments common in unpaired datasets. The deep feature-based LPIPS, however, is more robust to such issues and better correlates with human perceptual judgments. This metric suite therefore enables a comprehensive assessment of image synthesis quality. Specifically, SSIM, LPIPS, and MSE were computed following z-score normalization, and the reported results represent the average over five runs. Additionally, a paired Student's t-test was performed to evaluate the significance of the performance differences between CRAViM and the compared methods ($p = 0.05$).

3.2 Experimental Results

Based on Fig. 3 and Table 1, a direct visual comparison of the conversion results across the IXI and BraTS2018 datasets demonstrates that CRAViM exhibits a significant advantage in cross-modal image translation tasks. Experimental results show that CRAViM significantly outperforms other competing methods in key metrics such as SSIM and MSE ($p < 0.05$), thereby validating its superior performance in both detail restoration and global consistency preservation. This high-quality modality conversion provides clinicians with additional complementary information, enhances the interpretation of pathological regions, and reduces the risk of missed diagnoses.

Furthermore, although the diffusion-based SynDiff exhibits a slight LPIPS advantage and shows no significant difference in NMI compared to CRAViM

Table 1. Quantitative comparison on the IXI and Brats2018 datasets. Results are presented as mean ± standard deviation over five evaluations.

Method	T1 → T2				T2 → T1			
	SSIM ↑	NMI ↑	LPIPS ↓	MSE ↓	SSIM ↑	NMI ↑	LPIPS ↓	MSE ↓
IXI Dataset								
CycleGAN [22]	74.32 ±.42	1.201 ±.003	0.150 ±.002	0.222 ±.008	77.82 ±.28	1.238 ±.001	0.122 ±.002	0.106 ±.002
RegGAN [9]	75.39 ±.21	1.209 ±.003	0.132 ±.001	0.221 ±.001	77.72 ±.34	1.239 ±.002	0.121 ±.001	0.112 ±.003
AttGAN [17]	74.59 ±.47	1.211 ±.004	0.131 ±.004	0.216 ±.001	78.34 ±.24	1.233 ±.003	0.119 ±.005	0.108 ±.003
SynDiff [12]	77.46 ±.36	1.227 ±.004	**0.113** ±.001	0.209 ±.007	80.18 ±.67	1.251 ±.008	0.111 ±.003	0.108 ±.004
CRAViM	**78.59** ±.14	**1.228** ±.003	0.124 ±.001	**0.165** ±.001	**81.94** ±.20	**1.265** ±.002	**0.098** ±.001	**0.083** ±.001
Brats2018 Dataset								
CycleGAN [22]	90.38 ±.26	1.323 ±.001	0.064 ±.001	0.060 ±.001	90.75 ±.18	1.315 ±.001	0.049 ±.001	0.021 ±.001
RegGAN [9]	89.31 ±.27	1.315 ±.002	0.068 ±.004	0.065 ±.006	90.25 ±.08	1.311 ±.001	0.051 ±.002	0.022 ±.001
AttGAN [17]	89.81 ±.06	1.320 ±.001	0.065 ±.001	0.063 ±.002	90.52 ±.04	1.315 ±.001	0.048 ±.001	0.020 ±.001
SynDiff [12]	92.24 ±.22	1.339 ±.004	0.053 ±.001	0.055 ±.004	92.31 ±.89	1.326 ±.007	0.047 ±.001	0.018 ±.002
CRAViM	**92.43** ±.15	**1.344** ±.001	**0.051** ±.001	**0.047** ±.001	**93.32** ±.17	**1.335** ±.002	**0.039** ±.001	**0.016** ±.001

($p > 0.05$) on the IXI T1→T2 task, CRAViM achieves superior overall performance with only 11.77M parameters (35% of NCSNpp) and 265.16G FLOPs (4.8% of NCSNpp). The proposed framework reduces storage requirements by 64.9% (44.89MB vs. 128.16MB) and accelerates training by 110% (0.10s/iter vs. 0.21s/iter), achieving an optimal balance between computational efficiency and accuracy, thereby providing an efficient and practical solution for real-time clinical applications.

3.3 Ablation Study

To validate the effectiveness of our core innovations, we conducted ablation studies on the IXI dataset. As shown in Table 2, the introduction of the cycle denoise consistency training strategy significantly improves the ResNet baseline, achieving a 4.9% increase in SSIM and a 21.2% reduction in MSE for the T1→T2 task, substantially outperforming the traditional cycle consistency strategy.

Notably, even under our high-performance cycle denoise consistency training framework, CRAViM outperforms the structurally similar convolutional generative network AttGAN in the T2→T1 task, achieving a 0.96% improvement in SSIM and a 7.55% reduction in LPIPS. This enhancement is attributed to the EVSS module, which establishes global dependencies through its selective scanning mechanism, overcoming the limited receptive field of convolutional operations and better aligning with the continuity of anatomical structures in medical images.

Table 2. Performance of different networks under various architectures on the IXI dataset. Results are presented as the mean over five evaluations.

Network	**T1 → T2**				**T2 → T1**			
	SSIM ↑	**NMI ↑**	**LPIPS ↓**	**MSE ↓**	**SSIM ↑**	**NMI ↑**	**LPIPS ↓**	**MSE ↓**
Cycle Consistency								
ResNet [22]	74.32	1.201	0.150	0.222	77.82	1.238	0.122	0.106
AttGAN [17]	74.59	1.211	0.131	0.216	78.34	1.233	0.119	0.108
CRAViM	75.28	1.204	0.144	0.214	78.77	1.246	0.116	0.099
Cycle Denoise Consistency								
ResNet [22]	77.96	1.226	0.128	0.175	81.56	1.263	0.103	0.083
AttGAN [17]	77.73	1.225	0.127	0.173	81.16	1.260	0.106	0.086
CRAViM (w/o rdn.)	77.41	1.225	0.128	0.178	78.45	1.244	0.116	0.103
CRAViM (w/o trans.)	78.21	1.226	0.125	0.171	81.53	1.262	0.101	0.084
CRAViM	78.59	1.228	0.124	0.165	81.94	1.265	0.098	0.083

Table 2 further reveals the synergistic effects of the innovative components: (1) The cross-modal transport loss (Equation 3), as a novel indirect constraint strategy, optimizes the feature space rather than directly supervising the image, thereby avoiding conflicts with adversarial loss optimization; (2) When combined with the random denoise loss (Equation 4), the MSE for the T1→T2 task is reduced by an additional 7.30%, effectively enhancing the robustness of detail reconstruction; (3) The complete model produces super-additive improvements in both tasks, validating the orthogonal advantage of indirect control and direct constraints. This hybrid optimization strategy successfully addresses the trade-off between global consistency and local detail in medical image synthesis: the transport loss guarantees the preservation of macroscopic anatomical structures, while the denoise loss focuses on local feature fidelity. The synergy between these two losses significantly enhances the model's generalization ability.

4 Conclusion

In this paper, we propose a hybrid model, CRAVIM, along with a cycle denoise consistency training framework to address the challenges of global structural distortion and local detail loss in unpaired medical image synthesis. To the best

of our knowledge, this is the first application of the Mamba architecture in unsupervised image synthesis tasks. As such, our approach offers an efficient and robust solution for unpaired medical image synthesis. A direct comparison with other Mamba-based medical models remains challenging, as their architectures—typically based on U-Net structures with patch merging—are generally optimized for segmentation rather than synthesis. Future work will focus on extending our framework to 3D applications and validating its performance on more diverse modalities and tasks involving long-range dependencies. We hope this study will promote further advancements in the field of unpaired medical image analysis.

Acknowledgments. This work was supported in part by the Natural Science Foundation of Guangdong Province (2025A1515012230).

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

References

1. Bansal, S., Madisetty, S., Rehman, M.Z.U., Raghaw, C.S., Duggal, G., Kumar, N., et al.: A comprehensive survey of mamba architectures for medical image analysis: Classification, segmentation, restoration and beyond. arXiv preprint arXiv:2410.02362 (2024)
2. Biomedical Image Analysis Group, I.C.L., Centre for the Developing Brain, K.C.L.: Information extraction from images (2018), https://brain-development.org/ixi-dataset/
3. Cui, C., Yang, H., Wang, Y., Zhao, S., Asad, Z., Coburn, L.A., Wilson, K.T., Landman, B.A., Huo, Y.: Deep multimodal fusion of image and non-image data in disease diagnosis and prognosis: a review. Progress in Biomedical Engineering **5**(2), 022001 (2023)
4. Dalmaz, O., Yurt, M., Çukur, T.: Resvit: residual vision transformers for multimodal medical image synthesis. IEEE Transactions on Medical Imaging **41**(10), 2598–2614 (2022)
5. Dohmen, M., Klemens, M., Baltruschat, I., Truong, T., Lenga, M.: Similarity metrics for mr image-to-image translation. arXiv preprint arXiv:2405.08431 (2024)
6. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. ICLR (2021)
7. Guo, H., Li, J., Dai, T., Ouyang, Z., Ren, X., Xia, S.T.: Mambair: A simple baseline for image restoration with state-space model. In: European Conference on Computer Vision. pp. 222–241. Springer (2024)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016)
9. Kong, L., Lian, C., Huang, D., Hu, Y., Zhou, Q., et al.: Breaking the dilemma of medical image-to-image translation. Advances in Neural Information Processing Systems **34**, 1964–1978 (2021)

10. Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B., Sánchez, C.I.: A survey on deep learning in medical image analysis. Medical image analysis **42**, 60–88 (2017)
11. Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., et al.: The multimodal brain tumor image segmentation benchmark (brats). IEEE transactions on medical imaging **34**(10), 1993–2024 (2014)
12. Özbey, M., Dalmaz, O., Dar, S.U., Bedel, H.A., Özturk, Ş., Güngör, A., Çukur, T.: Unsupervised medical image translation with adversarial diffusion models. IEEE Transactions on Medical Imaging (2023)
13. Pei, X., Zuo, K., Li, Y., Pang, Z.: A review of the application of multi-modal deep learning in medicine: bibliometrics and future directions. International Journal of Computational Intelligence Systems **16**(1), 44 (2023)
14. Pei, X., Huang, T., Xu, C.: Efficientvmamba: Atrous selective scan for light weight visual mamba. arXiv preprint arXiv:2403.09977 (2024)
15. Pfeiffer, M., Funke, I., Robu, M.R., Bodenstedt, S., Strenger, L., Engelhardt, S., Roß, T., Clarkson, M.J., Gurusamy, K., Davidson, B.R., et al.: Generating large labeled data sets for laparoscopic image processing tasks using unpaired image-to-image translation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part V 22. pp. 119–127. Springer (2019)
16. 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: proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. vol. LNCS 15007. Springer Nature Switzerland (October 2024)
17. Tang, H., Liu, H., Xu, D., Torr, P.H., Sebe, N.: Attentiongan: Unpaired image-to-image translation using attention-guided generative adversarial networks. IEEE Transactions on Neural Networks and Learning Systems (TNNLS) (2021)
18. 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)
19. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)
20. Zhang, Z., Yang, L., Zheng, Y.: Translating and segmenting multimodal medical volumes with cycle- and shape-consistency generative adversarial network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)
21. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2881–2890 (2017)
22. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2223–2232 (2017)
23. Zhu, L., Liao, B., Zhang, Q., Wang, X., Liu, W., Wang, X.: Vision mamba: Efficient visual representation learning with bidirectional state space model. arXiv preprint arXiv:2401.09417 (2024)