MICCAI2025
28th International Conference on
Medical Image Computing and
Computer Assisted Intervention
23-27 September 2025
Daejeon Convention Center
Daejeon REPUBLIC OF KOREA

ClinicalFMamba: Advancing Clinical Assessment using Mamba-based Multimodal Neuroimaging Fusion

MICCAI MLMI 2025

Meng (Simon) Zhou*[1], and Farzad Khalvati [2,3]
[1] TD Bank Group, Toronto, Canada
[2] Department of Computer Science, University of Toronto, Toronto, Canada
[3] Department of Medical Imaging, University of Toronto, Toronto, Canada
*Work done while at University of Toronto

Outline

Problem

Intro

- **Background**
- **Related Works**

Method

- **Preliminaries**
- **Proposed Method**

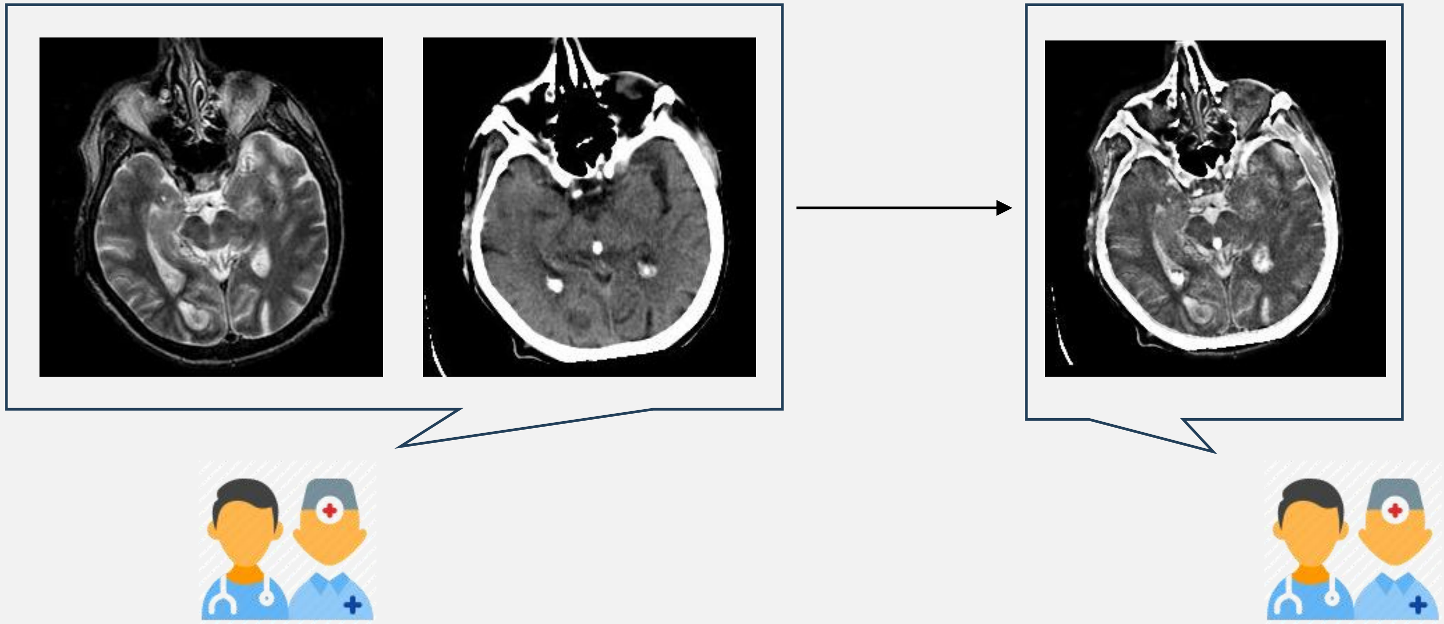
Results

- **Experiments**
- **Conclusions**

Intro

Background

- Multimodal medical image fusion plays an increasingly prominent role in clinical diagnosis.
- It aims to aggregate complementary information from different image modalities to produce higher-quality fused images (e.g., anatomical and functional images)
- Due to hardware constraints and current physical imaging principles, individual modalities can only capture specific aspects of tissue characteristics, leading to incomplete diagnostic information.



Related Works

- CNN-based:
  - MSRPAN [1]: Residual pyramid attention network for multimodal medical image fusion and Feature Energy Ratio Strategy to fuse feature maps
  - MSDRA [2]: Double residual attention network for multimodal medical image fusion and uses weighted L1 Norm to fuse feature maps.
  - EH-DRAN [3]: Dilated residual attention network + edge enhancer to extract multi-scale features and enhance edge details. A family of parameter-free fusion strategies is proposed to fuse feature maps in the latent space.

Limitations…

1. limited by their inherent local receptive fields, which restrict their ability to capture long-range spatial dependencies.
2. Most of the methods are in two-stage, which create another layer of computation

[1] Fu, J., Li, W., Du, J., & Huang, Y. (2021). A multiscale residual pyramid attention network for medical image fusion. *Biomedical Signal Processing and Control*, *66*, 102488.
[2] Li, W., Peng, X., Fu, J., Wang, G., Huang, Y., & Chao, F. (2022). A multiscale double-branch residual attention network for anatomical–functional medical image fusion. *Computers in biology and medicine*, *141*, 105005.
[3] Zhou M, Zhang Y, Xu X, Wang J, Khalvati F. Edge-Enhanced Dilated Residual Attention Network for Multimodal Medical Image Fusion. In2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) 2024 Dec 3 (pp. 4108-4111). IEEE.

Related Works

- Transformer-based:
  - SwinFusion [4]: Combining a CNN feature extractor with a cross-domain transformer model to fuse local and global information.
  - MRSCFusion [5]: Combining a multiscale CNN model and applied residual Swin Transformer layers to fuse cross-domain information.
  - MACTFusion [6]: Light-weight cross modality transformer with window and grid attention.

Limitations…

1. Self-Attention mechanism requires quadratic computational complexity, limiting the practical applications on large image
2. Applying the fusion method to real clinical diagnosis tasks is not fully explored.

[4] Ma, J., Tang, L., Fan, F., Huang, J., Mei, X., & Ma, Y. (2022). SwinFusion: Cross-domain long-range learning for general image fusion via swin transformer. *IEEE/CAA Journal of Automatica Sinica*, *9*(7), 1200-1217.

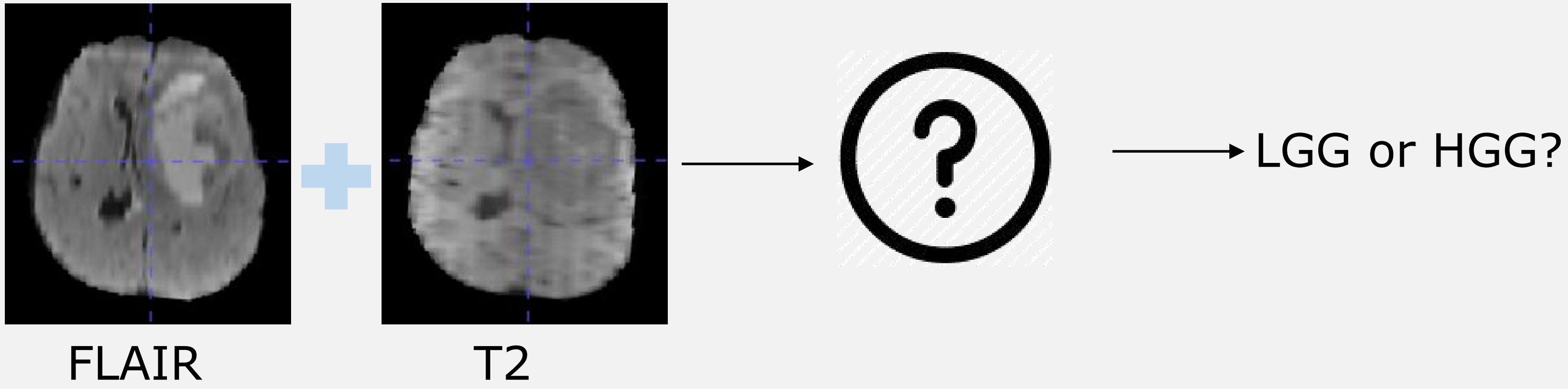
[5] Xie, X., Zhang, X., Ye, S., Xiong, D., Ouyang, L., Yang, B., ... & Wan, Y. (2023). MRSCFusion: Joint residual Swin transformer and multiscale CNN for unsupervised multimodal medical image fusion. *IEEE Transactions on Instrumentation and Measurement*.

[6] Xie, X., Zhang, X., Tang, X., Zhao, J., Xiong, D., Ouyang, L., ... & Teo, K. L. (2024). MACTFusion: Lightweight Cross Transformer for Adaptive Multimodal Medical Image Fusion. *IEEE Journal of Biomedical and Health Informatics*.

Problem

- Two most common fusion tasks in medical imaging: MRI-CT and MRI-SPECT fusion tasks (MRI-CT and MRI-SPECT from Harvard Whole Brain Atlas datasets)

- To further validate the effectiveness of the fusion method, we apply it to a downstream clinical brain tumor pathology classification task between Low-Grade and High-Grade Gliomas (BraTS 2019 dataset).



Method

- Preliminaries
- Proposed Method
- Scaning Strategy

Preliminaries

Mamba is a framework that model sequential data through a hidden state that evolves over time. The transformation from a 1-D sequence x(t) to an output y(t) is achived through a hidden state h(t). The implementation is typically achieved through linear ODEs:

$$h'(t) = Ah(t) + Bx(t), y(t) = Ch(t)$$

To discritize, we use zero-order hold:

$$\overline{A} = exp(\Delta A), \overline{B} = (\Delta A)^{-1}(exp(\Delta A) - I) \cdot \Delta B$$

We can rewrite the first equation as:

$$h_t = \overline{A}h_{t-1} + \overline{B}x_t, y_t = Ch_t$$

Selective-scan:

1. can focus on or ignore particular information, like self-attention in Transformer
2. B, C, Δ is dynamic to the input, allowing the model to filter relevent information and focus on important context within long sequences
3. B: filter out irrelevant data or to focus on important data to allow only relevant data to enter into the new state.
4. C: selective to decide which information from the state is required to materialize the output
5. Δ: weighting between the importance of new samples compared to the previous state.

Proposed Method

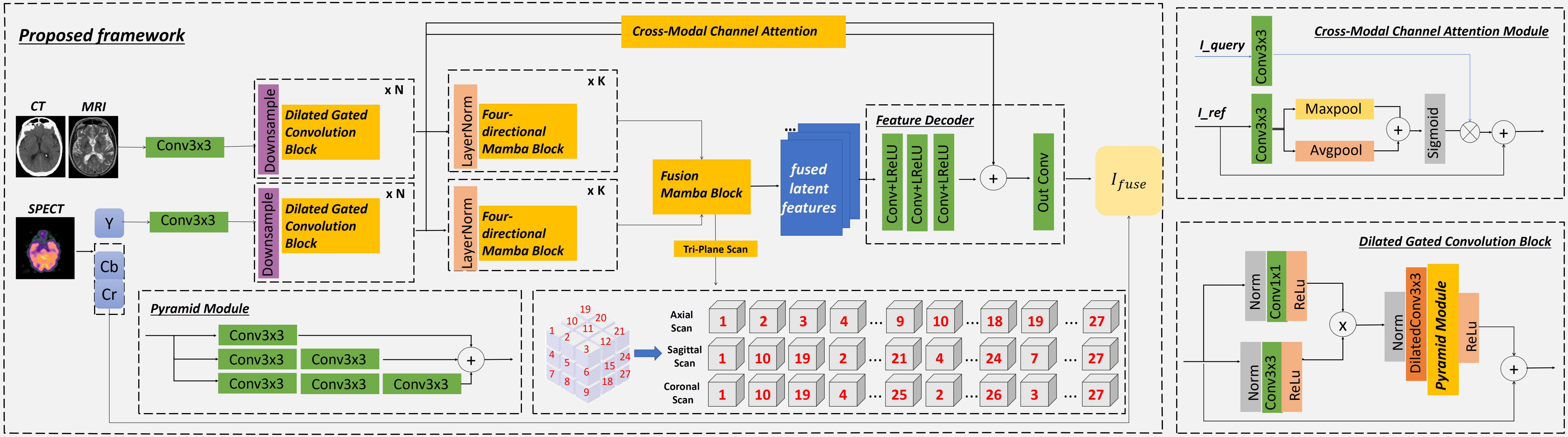
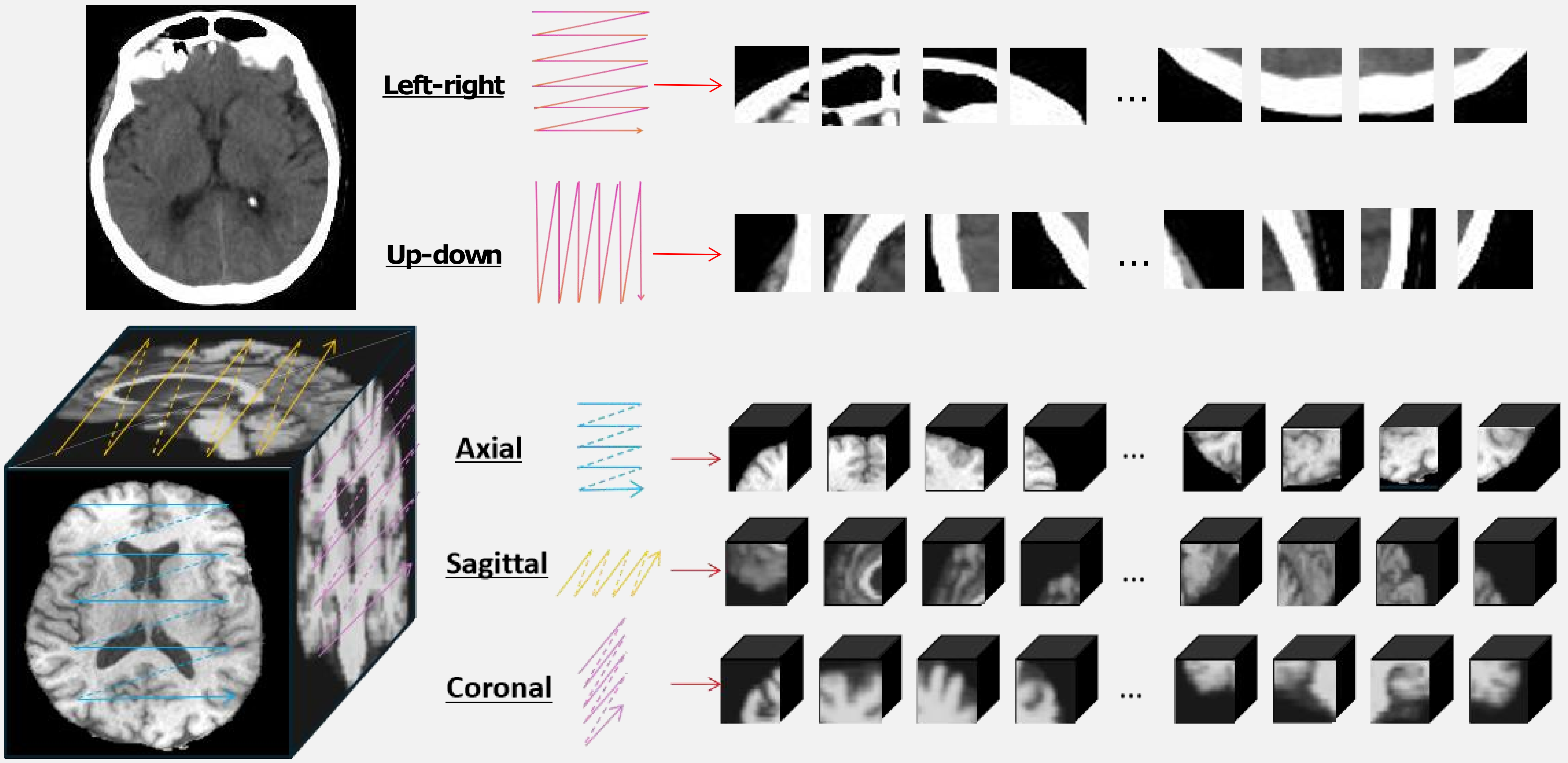


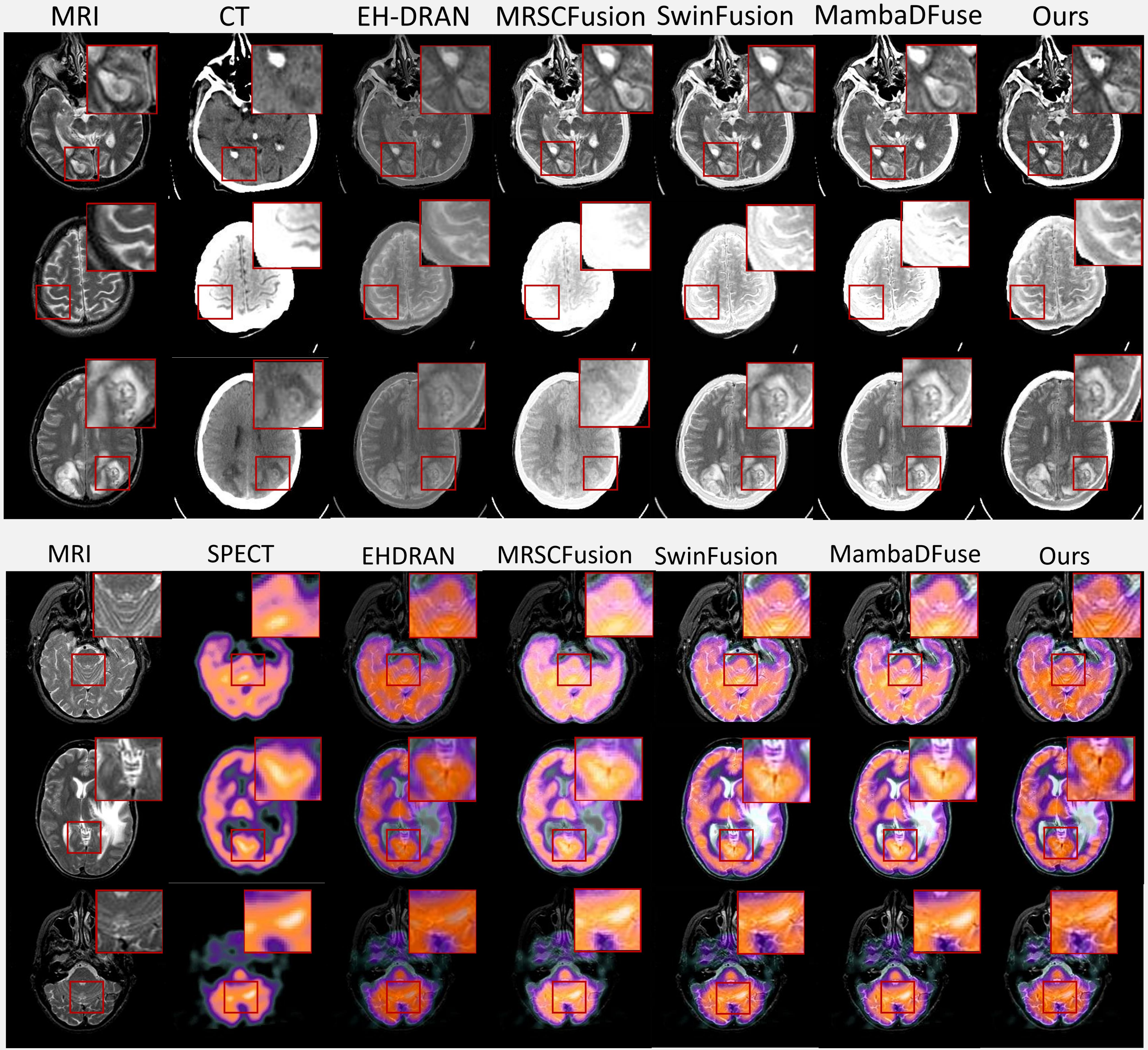
Fig. 1. An overview of the proposed framework. DilatedConv3x3 represents dilated convolution with kernel size 3×3. All Conv+LReLU layers in the decoder have 3×3 convolution followed by Leaky-ReLU. For 3D implementation, all operations are replaced with their corresponding 3D alternatives (i.e., Conv3×3 to Conv3×3×3).

- Pixel loss, ensure appropriate intensity information is retained between fused and input images $L_{pixel} = \|I_f - max(I_1, I_2)\|_1$
- Gradient loss, preverse fine-grained texture details from input images, $L_{grad} = \|\nabla I_f - max(\nabla I_1, \nabla I_2)\|_2$
- Similarity loss, preserve structure similarity between fused and input images, $L_{ssim} = \frac{1}{2}(1 - SSIM(I_f, I_1)) + \frac{1}{2}(1 - SSIM(I_f, I_2))$

2D vs. 3D Scan in Mamba



Result


MRI
CT
EH-DRAN
MRSCFusion
SwinFusion
MambaDFuse
Ours
MRI
SPECT
EHDRAN
MRSCFusion
SwinFusion
MambaDFuse
Ours

Quantitative Results

Dataset	Method	PSNR↑	SSIM↑	FMI↑	FSIM↑	EN↑
MRI-CT	EH-DRAN	**16.830±0.490**	0.753±0.007	0.883±0.005	0.820±0.003	10.727±0.531
	SwinFusion	14.962±0.173	0.768±0.007	0.882±0.002	0.810±0.001	8.445±0.078
	MRSCFusion	14.476±0.205	0.713±0.012	0.877±0.006	0.791±0.010	7.544±0.232
	MambaDFuse	15.873±0.289	0.771±0.007	0.882±0.005	0.817±0.004	15.018±0.167
	ClinicalFMamba (Ours)	16.519±0.352	**0.783±0.005**	**0.883±0.003**	**0.820±0.001**	**15.213±0.069**
MRI-SPECT	EH-DRAN	21.455±0.071	0.736±0.002	**0.876±0.004**	0.843±0.003	11.970±0.538
	SwinFusion	17.557±0.021	0.728±0.004	0.808±0.007	0.819±0.011	13.066±0.428
	MRSCFusion	18.412±0.211	0.734±0.012	0.827±0.009	0.814±0.006	9.87±0.600
	MambaDFuse	21.021±0.034	0.748±0.004	0.845±0.006	0.829±0.002	14.126±0.439
	ClinicalFMamba (Ours)	**21.561±0.067**	**0.759±0.009**	0.856±0.003	**0.848±0.002**	**14.871±0.334**

	PSNR↑	MS-SSIM↑	EN↑
EH-DRAN-3D	30.653±0.554	0.814±0.095	17.402±1.134
ClinicalFMamba-3D (Ours)	**33.937±0.361**	**0.859±0.045**	**20.468±1.541**

Fusion Time Analysis

For 2D model on BraTS-2D dataset, our model has 4.05M parameters and achieves an average time of **0.1s** per image pair with 128x128 resolution

For 3D model on BraTS-3D dataset, our model has 6.01M parameters and achieves an avergae time of **7.3s** per image volume with 128x128x128 resolution

Ablations

MRI-CT	PSNR	SSIM	FMI	FSIM	EN
w/o CMCA	15.967±0.349	0.761±0.007	0.876±0.004	0.813±0.005	14.621±0.075
with CMCA	16.519±0.352	0.783±0.005	0.883±0.003	0.820±0.001	15.213±0.069

BraTS-3D	PSNR	MS-SSIM	EN
w/o 3D-scan	26.882±0.324	0.833±0.073	19.558±1.374
with 3D-can	33.937±0.361	0.859±0.045	20.468±1.541

Quantitative Results on BraTS-classification

Dataset	Method	AUC↑	F1-Score↑	Accuracy↑
BraTS-2D	T2	0.722±0.021	0.703±0.018	0.604±0.037
	FLAIR	0.727±0.024	0.701±0.008	0.611±0.017
	T2+FLAIR	0.723±0.028	0.717±0.012	0.640±0.015
	EH-DRAN	0.769±0.003	0.723±0.006	0.640±0.011
	ClinicalFMamba (Ours)	**0.790±0.013**	**0.778±0.023**	**0.665±0.004**
BraTS-3D	T2-3D	0.647±0.022	0.560±0.029	0.635±0.041
	FLAIR-3D	0.641±0.110	0.529±0.223	0.566±0.010
	T2-3D+FLAIR-3D	0.636±0.072	0.540±0.145	0.630±0.027
	EH-DRAN-3D	0.646±0.015	0.571±0.037	**0.657±0.016**
	ClinicalFMamba-3D (Ours)	**0.652±0.038**	**0.584±0.023**	0.647±0.013

Conclusion

- Novel CNN-Mamba architecture for effective 2D and 3D medical image fusion
- We propose dilated gated convnets for multiscale feature learning and cross-modal channel attention for cross-modal information fusion and decode.
- Introduced a tri-plane scanning strategy for 3D medical image fusion.
- The framework is able to generate fused images in real-time, and we valiate the clinical usage of the fused images through brain tumor classification task.
- Extensive evaluated on three datasets to valid the effectiveness of the proposed approach.

- Future Work: Explore pure Mamba-based methods and extend to other dieases and downstream tasks like (3D) segmentation.

Thank you!

Arxiv preprint:

Appendix