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

Meng Zhou[1,*] and Farzad Khalvati[2,3]

[1]TD Bank Group, Toronto, Canada
[2]Department of Computer Science, University of Toronto
[3]Department of Medical Imaging, University of Toronto
*Work done while at University of Toronto

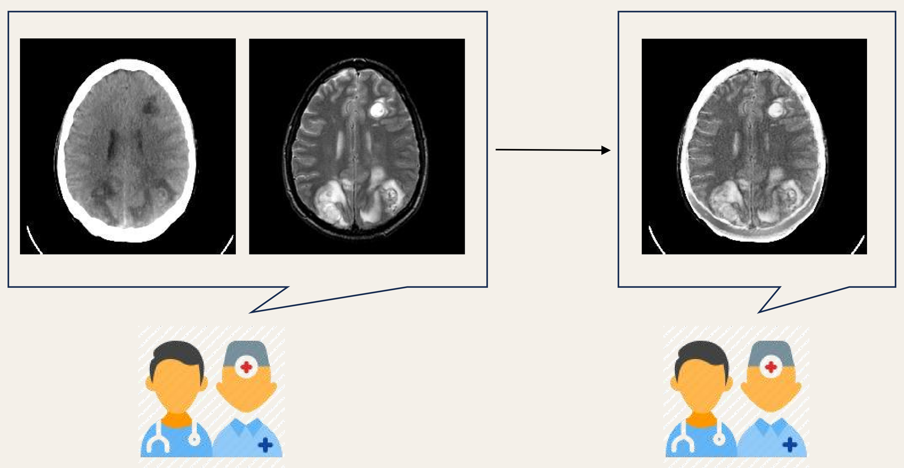
MLMI 2025 Workshop

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

INTODUCTION

Context:

- ➢ Multimodal image fusion (MMIF) plays an increasingly prominent role in clinical diagnosis.
- ➢ Aggregate **complementary information** from different image modalities to produce higher-quality fused images (e.g., anatomical and functional images)



Challenges:

- ➢ CNN based methods [1] are limited by their inherent local receptive fields, which restrict their ability to capture long-range spatial dependencies.
- ➢ Transformer based methods [2,3] create prohibitive costs for clinical applications with large images due to its self-attention mechanisms, limiting the practical deployment capabilities.
- ➢ 3D medical image fusion and the clinical applicability of fused images are still underexplored.

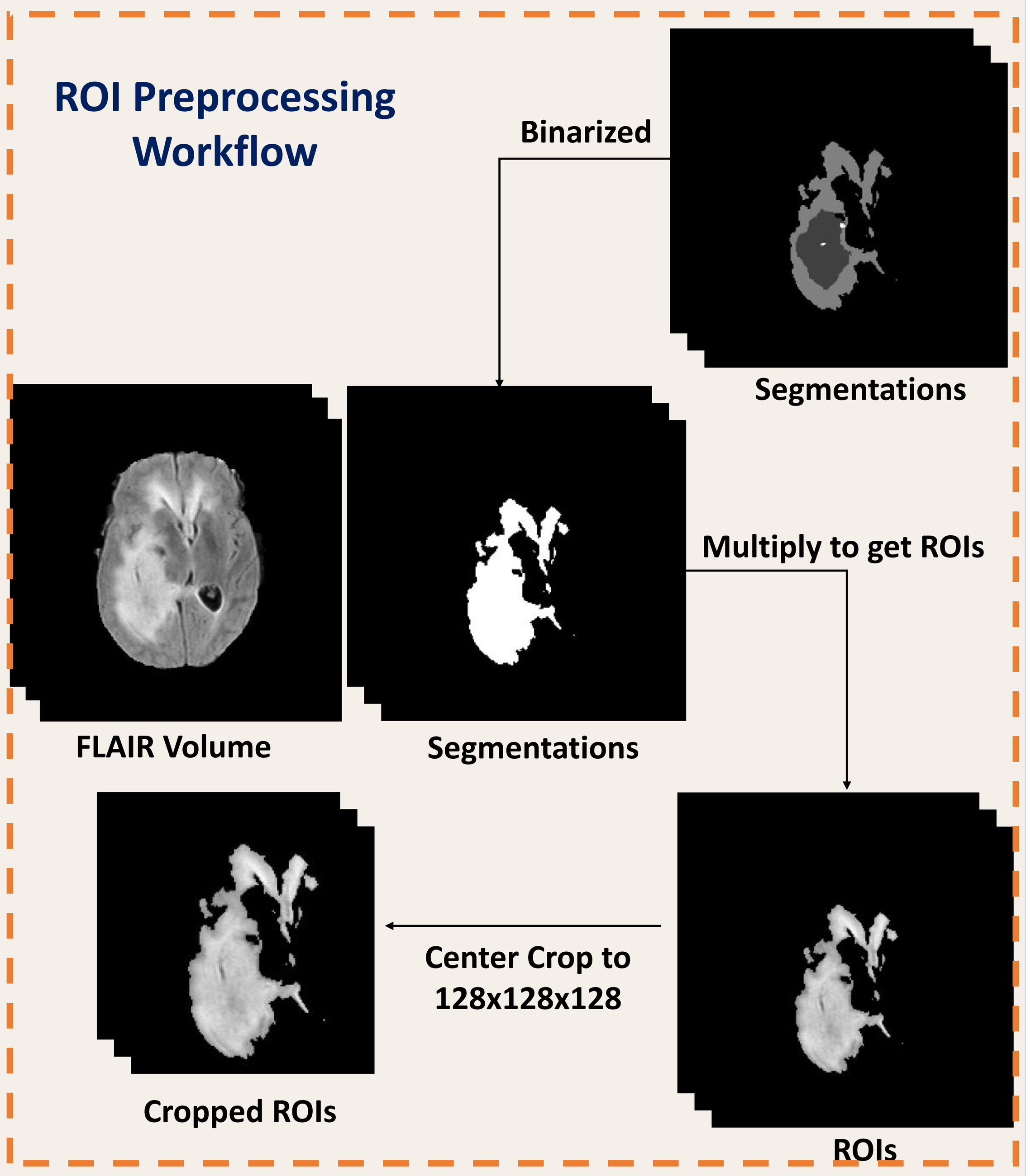
Contributions:

- ➢ We present a CNN-Mamba hybrid framework to effectively model local and global features in **2D and 3D** medical images.
- ➢ We propose **dilated gated convnets** for multiscale feature learning and **cross-modal channel attention** for cross-modal information fusion.
- ➢ We are the first to extend the Mamba-based fusion method to 3D medical imaging through a novel **tri-plane scanning strategy**

MATERIALS

Dataset and Preprocessing:

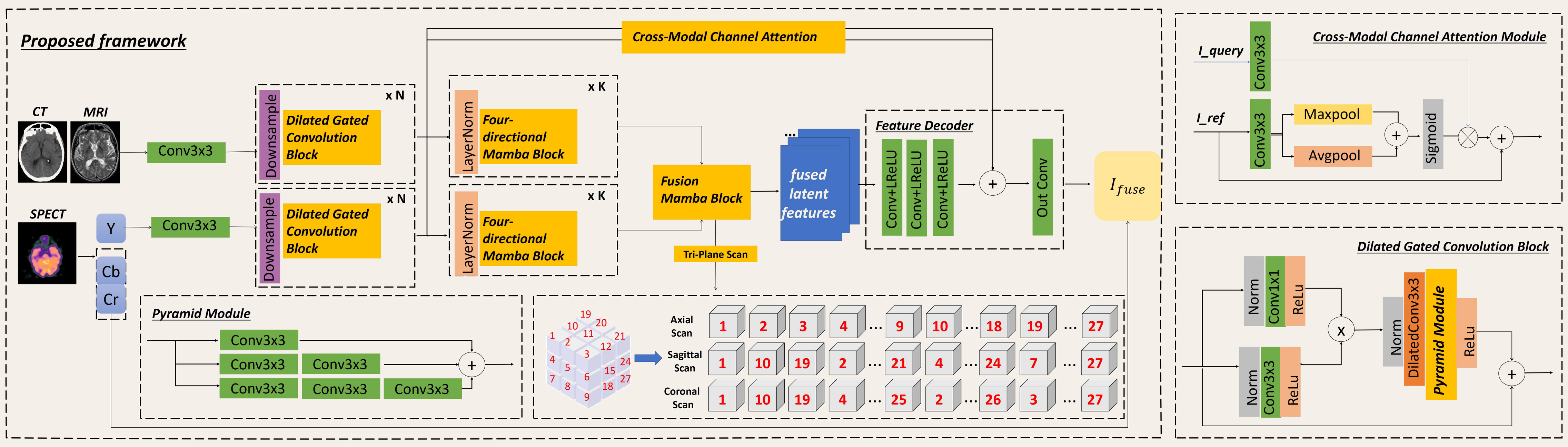
- ➢ **MRI-CT and MRI-SPECT** datasets from The Harvard Whole Brain Atlas dataset, normalized to [0,1]
- ➢ **BraTS 2019** Dataset with **T2 and FLAIR** sequences are used for downstream classification task
- ➢ Reshape from 240x240x155 to ROI-based volume 128x128x128 [5], normalized to [0,1]
- ➢ For 2D, further slice the volume through Axial plane



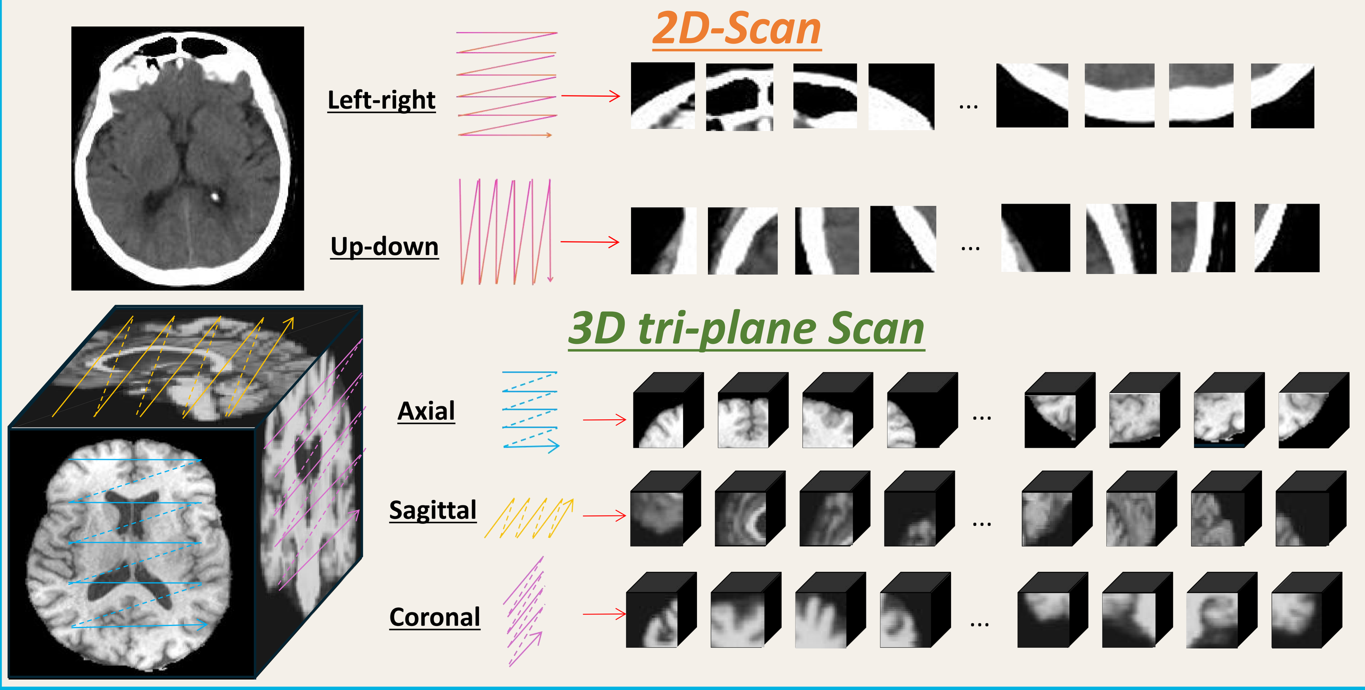
METHODS

Proposed Method:

- ➢ End-to-end training instead of the two-stage approach [1], the output is the final fused image



- ➢ Dilated Gated Conv Block is designed to efficiently learn multi-scale local spatial, descrminative features
- ➢ Mamba blocks with 2D scan for 2D images [4], with 3D *tri-plane scan* for 3D images



Loss-Function

- ➢ Pixel loss, ensure appropriate intensity information is retained between fused and input images

$$L_{pixel} = \left\| I_f - max(I_1, I_2) \right\|_1$$

- ➢ Gradient loss, preverse fine-grained texture details from input images

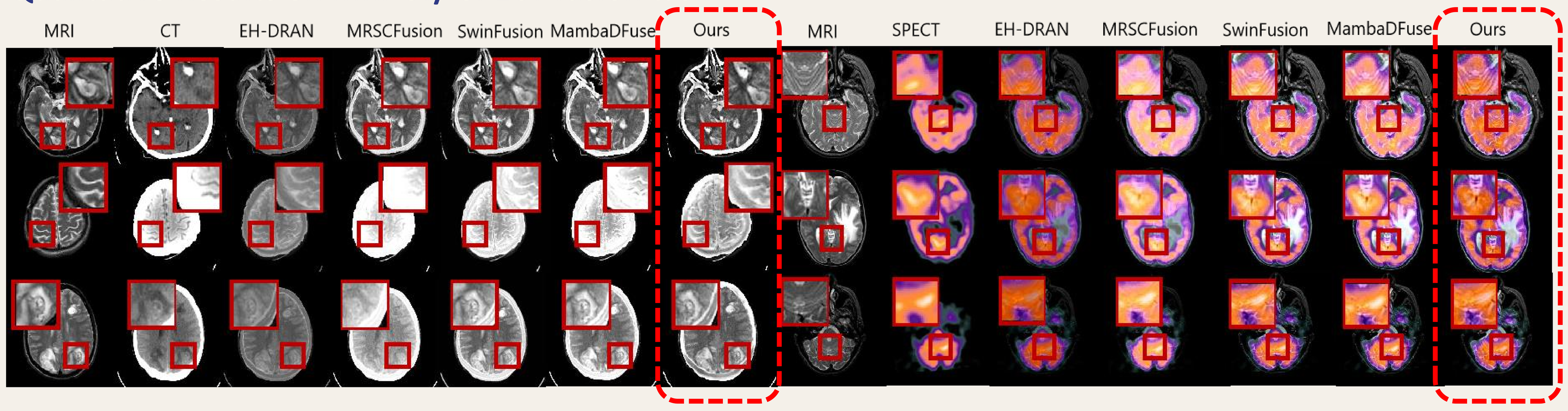
$$L_{grad} = \left\| \nabla I_f - max(\nabla I_1, \nabla I_2) \right\|_2$$

- ➢ Similarity loss, preserve structure similarity between fused and input images

$$L_{ssim} = \frac{1}{2}(1 - SSIM(I_f, I_1)) + \frac{1}{2}(1 - SSIM(I_f, I_2))$$

RESULTS

Qualitative Results on MRI-CT/-SPECT fusion task



Quantitative Results on Downstream Classification task

- ➢ 2D and 3D LGG/HGG Pathology type ROI-based classification

BraTS-2D	AUC	F1-Score	Accuracy
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
Ours	**0.790±0.013**	**0.778±0.023**	**0.665±0.004**

BraTS-3D	AUC	F1-Score	Accuracy
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**
Ours-3D	**0.652±0.038**	**0.584±0.023**	0.647±0.013

Ablations

- ➢ Proposed Cross-modal channel attention module and 3D tri-plane scan strategy

BraTS-2D	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**

Key Takeaways

- ➢ Visually, the fused images from our method preserve both modality-specific features and inter-modal contrast.
- ➢ Using the fused image, we validate its clinical utility on brain tumor classification task by comparing with other methods such as single-modality, dual-modality and EH-DRAN basline. Our method yields the best performance, demonstrating strong potential for clinical usage.

Conclusion

Summary:

- ➢ We propose a novel CNN-Mamba hybrid framework for effective multimodal 2D and 3D medical image fusion. Experiments show our method outperforms several baselines.
- ➢ 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.

Future Work:

- ➢ Extend to other diseases and validation on other tasks such as segmentation.
- ➢ Explore pure Mamba-based method to replace CNN.

References

[1] 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.
[2] Xie X, Zhang X, Ye S, Xiong D, Ouyang L, Yang B, Zhou H, Wan Y. Mrscfusion: Joint residual swin transformer and multiscale cnn for unsupervised multimodal medical image fusion. IEEE Transactions on Instrumentation and Measurement. 2023 Sep 20;72:1-7.
[3] 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. 2022 Jun 30;9(7):1200-17.
[4] Peng S, Zhu X, Deng H, Deng LJ, Lei Z. Fusionmamba: Efficient remote sensing image fusion with state space model. IEEE Transactions on Geoscience and Remote Sensing. 2024 Nov 11.
[5] Zhou M, Khalvati F. Conditional generation of 3d brain tumor regions via vqgan and temporal-agnostic masked transformer. InMedical Imaging with Deep Learning 2024 Dec 23.

Acknowledgements

No funding was received for conducting this study. We thank the computational resources provided by Google Colab