ClinicalFMamba: Mamba-based Multimodal Medical Image Fusion for Enhanced Clinical Diagnosis

Meng Zhou [1] **Farzad Khalvati** [1 2]

Abstract

Multimodal medical image fusion integrates complementary information from different imaging modalities to enhance diagnostic accuracy and treatment planning. While deep learning methods have advanced fusion performance, existing approaches face critical limitations: CNNs excel at local feature extraction but struggle to model global context effectively, while Transformers achieve superior long-range modeling at the cost of quadratic computational complexity $O(N^2)$ in self-attention mechanisms, limiting clinical deployment. Recent State Space Models (SSMs) offer a promising alternative, enabling efficient long-range dependency modeling in linear time through selective mechanisms. Despite these advances, clinical validation of fused images remains underexplored. In this work, we propose ClinicalFMamba, a novel end-to-end CNN-Mamba hybrid architecture that synergistically combines local and global feature modeling. Our approach introduces: Dilated Gated Convolution Blocks for hierarchical multiscale feature extraction, and a latent Mamba module that efficiently captures long-range spatial dependencies between feature regions and enabling cross-modal fusion in latent space. Comprehensive evaluations on three datasets demonstrate the superior fusion performance across multiple quantitative metrics while achieving **real-time fusion**. Notably, we validate the clinical utility of our approach on the downstream brain tumor classification, achieving up to 7% improvements on the AUC score. Our method establishes a new paradigm for efficient multimodal medical image fusion suitable for real-time clinical deployment.

[1]Department of Computer Science, University of Toronto, Toronto, Canada [2]Department of Medical Imaging, University of Toronto, Toronto, Canada. Correspondence to: Meng Zhou <simonzhou@cs.toronto.edu>.

Presented in the 42^{nd} International Conference on Machine Learning Affinity Workshop New In ML, Vancouver, Canada.

1. Introduction

Medical image fusion (MMIF) aggregates complementary information from multiple modalities (e.g., CT, MRI, PET, SPECT) to produce higher-quality fused images that combine anatomical and functional details (Xie et al., 2023; Zhou et al., 2024). By integrating modality-specific strengths, such as soft-tissue contrast from MRI, bone delineation from CT, metabolic activity from PET, and blood flow information from SPECT, MMIF reveals subtle anatomical structures and pathological features that may be missed when examining individual modalities in isolation. This enhanced visualization capability significantly improves clinical applications, including tumor boundary localization (Chen et al., 2024) and radiotherapy treatment planning (Safari et al., 2023; Xie et al., 2023). The clinical necessity for MMIF arises from the inherent limitations of single-modal imaging systems. Due to hardware constraints and physical imaging principles (Xie et al., 2023), individual modalities can only capture specific aspects of tissue characteristics (Safari et al., 2023), leading to incomplete diagnostic information. Consequently, physicians must analyze multiple images from different modalities separately to obtain a comprehensive understanding, creating a time-consuming workflow that may lead to information fragmentation and potential diagnostic errors. MMIF addresses this challenge by integrating complementary information into a single, comprehensive image that preserves the most relevant features from each modality, thereby supporting more accurate and efficient diagnosis (Wang et al., 2022).

In recent years, deep learning models have significantly improved multimodal fusion performance through their powerful representation capabilities, with researchers primarily utilizing convolutional neural networks (CNNs) for image fusion tasks. Early CNN-based approaches focused on extracting hierarchical features from multiple modalities to generate comprehensive fused representations. DenseFuse (Li & Wu, 2019) introduced an infrared-visible fusion framework using dense blocks with a CNN backbone. MSDNet (Song et al., 2019) captures multi-scale features through various convolutional kernel sizes. (Fu et al., 2021) proposed a residual pyramid attention network for MRI-CT, MRI-PET, and MRI-SPECT fusion using a Feature Energy

Ratio Strategy for latent space fusion that selectively emphasizes informative features from each modality. Similarly, (Li et al., 2022) introduced a double residual attention network to capture detailed features while avoiding gradient issues. However, CNN-based models remain limited by their inherent local receptive fields, which restrict their ability to capture long-range spatial dependencies.

Transformer models (Vaswani et al., 2017) have recently attracted increasing attention by addressing CNNs' limitations in global feature extraction through their powerful self-attention mechanisms. For example, (Ma et al., 2022) proposed SwinFusion, which combines CNN and Transformer models to capture local information while integrating global complementary features from both domains. Similarly, (Xie et al., 2023) proposed a multiscale CNN with residual Swin Transformer layers for effective feature learning from both domains. However, the quadratic computational complexity $O(N^2)$ of self-attention mechanisms creates prohibitive costs for clinical applications with large images, limiting the practical deployment of transformer-based methods despite their superior performance over CNN approaches. Recently, the improved selective structured state space models (Mamba) (Gu & Dao, 2023) provide a novel solution by outperforming Transformers in long-term dependency modeling through selective global information learning with linear complexity ($O(N)$). Several studies have already leveraged Mamba in medical vision tasks, including classification (Yue & Li, 2024), segmentation (Xing et al., 2024; Li et al., 2025; Ma et al., 2024), and multimodal fusion (Li et al., 2024; Xie et al., 2024), achieving superior performance over CNN and Transformer counterparts. Despite these advancements, existing Mamba-based approaches primarily utilize Mamba blocks for feature learning while neglecting discriminative fine-grained local features that CNNs excel at capturing. Moreover, most fusion methods lack validation on clinical downstream tasks, limiting their real-world applicability. Therefore, developing a computationally efficient model that achieves superior fusion performance and demonstrates clinical effectiveness remains crucial.

To this end, we propose a novel end-to-end framework combining CNN and Mamba for multimodal medical image fusion. Our approach introduces: (1) Dilated Gated Convolution Blocks (DGCB) for multi-scale discriminative feature extraction, (2) a latent Mamba model for global feature interactions in the latent space, and (3) cross-channel attention for decoding fused features to image space. The method achieves superior texture preservation with reduced information loss, outperforming state-of-the-art fusion methods both qualitatively and quantitatively. Notably, we take a step further to validate our method for clinical applicability on high-grade glioma(HGG) vs. Low-grade glioma(LGG) brain tumor classification, a critical task for precision diagnosis and prognosis, followed (Zhou & Khalvati, 2024). To summarize, our contributions are as follows:

1. We introduce an end-to-end hybrid framework combining CNNs and Mamba to effectively model both local spatial features and long-range dependencies in medical images.

2. We propose Dilated Gated Convolution Blocks for multiscale feature learning, integrated with latent Mamba and cross-modal channel attention for seamless cross-modal information fusion.

3. To the best of our knowledge, we provide the first benchmark evaluation of Mamba-based fusion methods on the clinical downstream task for LGG/HGG brain tumor pathology type classification.

4. Experiments show our proposed method outperforms several baselines in both quantitative fusion metrics and qualitative image fidelity, as well as the performance on the downstream classification task.

2. Materials and Method

In this section, we present ClinicalFMamba, a novel architecture for multi-modal medical image fusion (Figure 1). The model comprises three key components (also detailed in Section 2.1): (1) a hybrid feature encoder utilizing stacked dilated gated convolution layers to capture multi-scale local features while preserving spatial resolution, (2) a latent Mamba module that models long-range spatial dependencies and performs cross-modal feature fusion in the latent space, and (3) a convolutional decoder that reconstructs the fused representations back to image space.

2.1. Feature Extraction and Image Reconstruction

Hybrid Feature Encoder. The core module in the feature encoder is the **Dilated Gated Convolution Block** (DGCB), designed to efficiently learn local spatial features. The gated mechanism enables cross-region interactions over feature maps and controls information transmission between layers, similar to (Liu et al., 2021). The DGCB processes input feature maps through two parallel convolution blocks, each containing normalization, convolution, and activation layers with 3×3 and 1×1 kernels, respectively. Element-wise multiplication between these features enables cross-region interactions. Subsequently, we combine dilated convolutions (Yu & Koltun, 2015) and pyramid convolutions (Li et al., 2018) to capture multi-scale features and enhance discriminative feature extraction capabilities. Dilated convolutions expand the receptive field while preserving spatial resolution, enabling better capture of local information and fine details. The pyramid convolutions, placed after dilated convolutions, follow (Zhou et al., 2024) by stacking one

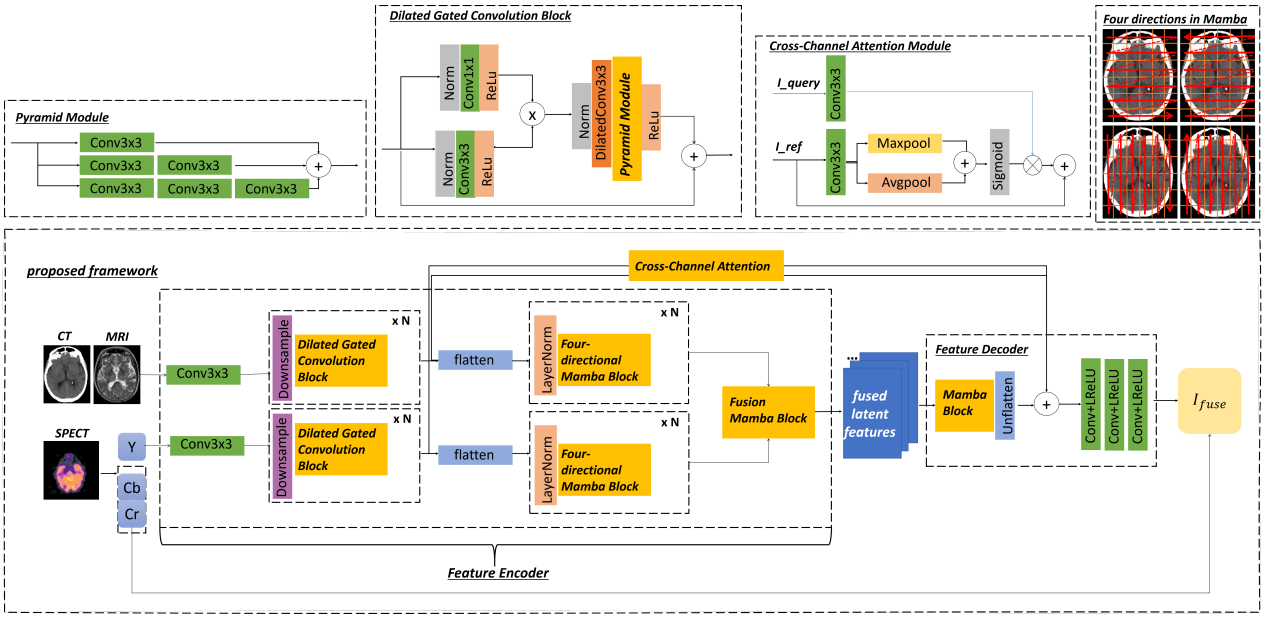

Figure 1. An overview of the proposed framework. `DConv3x3` represents dilated convolution with kernel size 3×3. We use $N = 3$ for DGCB, with dilated rate = 1,3,5, respectively, and $N = 5$ for the latent Mamba model. All `Conv+LReLU` layers in the decoder have 3×3 kernel followed by Leaky-ReLU. Note that the YCbCr conversion only applies to SPECT images. Zoom in for a better view.

3×3 convolution, two 3×3 convolutions, and three 3×3 convolutions for multi-scale feature learning process.

Latent Mamba for Feature Fusion. After learning the multiscale local features, we leverage Mamba's powerful global information modeling capabilities to learn long-term dependencies of latent feature regions. Mamba (Gu & Dao, 2023) introduces the selective scan mechanism (SSM) to filter out irrelevant and redundant information while retaining relevant information. In this work, we adopt the four-directional Mamba and Fusion Mamba block (Peng et al., 2024) for global feature modeling. **Four-directional Mamba block:** given an input feature map $\mathbf{f_{feat}} \in \mathcal{R}^{H\times W\times C}$, we first apply layer normalization and generate two feature maps $\mathbf{f^1_{feat}}, \mathbf{f^2_{feat}} \in \mathcal{R}^{H\times W\times C}$ through parallel 1×1 convolutions. Then, $\mathbf{f^1_{feat}}$ is flattened along four spatial directions (shown in upper right of Figure 1), producing $\mathbf{f^{11}_{feat}}, \mathbf{f^{12}_{feat}}, \mathbf{f^{13}_{feat}}, \mathbf{f^{14}_{feat}}$ in $HW \times C$. These sequences are processed separately by SSM blocks, yielding four outputs, $\mathbf{y^1_{feat}}, \mathbf{y^2_{feat}}, \mathbf{y^3_{feat}}, \mathbf{y^4_{feat}}$. Finally, we unflatten and combine using element-wise addition to obtain $\mathbf{Y} \in \mathcal{R}^{H\times W\times C}$. $\mathbf{Y}$ is then gated by $\mathbf{f^2_{feat}}$ (e.g., $\mathbf{Y} \cdot SiLU(\mathbf{f^2_{feat}})$), and apply another 1×1 convolution to process the final output. **Fusion Mamba block:** The Fusion Mamba block enables cross-modal feature integration by processing dual inputs asymmetrically. Specifically, the first modality generates the projection matrices and timescale parameters, while the second modality provides the input sequence for selective state-space processing. The process is the same as above, except two outputs are expected at the end, $\mathbf{Y^1}, \mathbf{Y^2} \in \mathcal{R}^{H\times W\times C}$. The gating operation is applied to both, $\mathbf{Y^1}, \mathbf{Y^2}$ separately. Finally, another 1×1 convolution is applied on the combined output, $\mathbf{Y^1} + \mathbf{Y^2}$, to process the final feature (More details can be found in Appendix A). The complete latent processing pipeline first applies five Four-Directional Mamba blocks for intra-modal long-range dependency modeling, followed by one Fusion Mamba block for cross-modal integration. The resulting fused features are then passed to the decoder network, detailed below.

Lightweight Image Decoder. The fused latent features are processed through a convolution block containing three upsample-convolutional layers with leaky ReLU activation. We introduce a cross-modal channel attention (CMCA) module to capture inter-channel interactions between fused and original features from both modalities. As shown in the upper right of Figure 1, the cross-modal channel attention module operates with two inputs I_{ref} and I_{query}. Both inputs are processed through 3×3 convolutions. For the reference input I_{ref}, we apply average pooling ($\mathbf{F_{avg}}$) and max pooling ($\mathbf{F_{max}}$) to select important channel-wise representations. The combined channel map, $\mathbf{F_{att}} = sigmoid(\mathbf{F_{avg}} + \mathbf{F_{max}})$, is then applied to the query input to preserve complementary information from both modalities. For bidirectional cross-modal enhancement, we perform this operation twice: first using one modality as reference and the other as query, then swapping their roles. For example, in CT-MRI fusion, we initially use

MRI as I_{ref} and CT as I_{query}, then reverse the assignment. The resulting enhanced features from both operations are element-wise added to the original latent features before decoding back to image space.

Loss Function. We employ a multi-component loss function combining structural similarity (SSIM), pixel intensity, and gradient difference, following (Zhou et al., 2024; Li et al., 2024; Xie et al., 2024). Unlike previous two-stage approaches that require separate training and fusion phases (Zhou et al., 2024; Li et al., 2018; Fu et al., 2021), our end-to-end framework necessitates a dual-target training strategy where both input modalities serve as reconstruction targets. The total loss is formulated as:

$$\begin{aligned} \mathcal{L}_{pixel} &= \|\hat{x} - max(x_1, x_2)\|_1, \\ \mathcal{L}_{grad} &= \|\nabla\hat{x} - max(\nabla\hat{x_1}, \nabla\hat{x_2})\|_2, \\ \mathcal{L}_{ssim} &= \frac{1}{2}(1 - SSIM(\hat{x}, x_1)) + \frac{1}{2}(1 - SSIM(\hat{x}, x_2)) \end{aligned} \tag{1}$$

$$\mathcal{L}(\theta) = \lambda_1 * \mathcal{L}_{pixel} + \lambda_2 * \mathcal{L}_{grad} + \lambda_3 * \mathcal{L}_{ssim} \tag{2}$$

where $\hat{x}$ is the fused image, x_1, x_2 are the input images from two modalities, $SSIM()$ is the operation to calculate structure similarity for two images, and we use $1-SSIM()$ as the loss to optimize. We empirically set $\lambda_1 = 2, \lambda_2 = 10, \lambda_3 = 5$ for all our fusion models.

2.2. Datasets

In this work, we use four datasets to validate the effectiveness of our proposed approach: MRI-CT (184 pairs) and MRI-SPECT (357 pairs) multi-modality fusion data sets[1], and the BraTS 2019 dataset (Bakas et al., 2017; 2018; Menze et al., 2014) (335 patients). Especially for MRI-SPECT fusion, we converted the SPECT images from the RGB color space to the YCbCr space following (Fu et al., 2021; Li et al., 2022; Zhou et al., 2024), using only the Y-channel images to train the model. All pairs of MRI-CT and MRI-SPECT images were coregistered and preprocessed beforehand so that each pixel intensity is in the range of [0,255]. We further normalized the pixel intensity to [0,1].

For the BraTS dataset, we use the T2 and FLAIR sequences as done in (Zhou et al., 2024; Zhou & Khalvati, 2024). We first obtained the ROIs by multiplying the images with masks, then reshaped the data from 240×240×155 to 128×128×128, and normalized all pixel intensities to the range [0,1]. We converted the 3D data to 2D by slicing over the *Axial plane* for each patient and only considered slices with at least 10% non-zero pixels.

[1] https://www.med.harvard.edu/aanlib/

3. Experiments

All programs were implemented in PyTorch. For both MRI-CT and MRI-SPECT pairs, we trained the autoencoder for 100 epochs with an initial learning rate of 0.0005 and cosine decay to 1e-7, a mini-batch size of 8, and with the Adam optimizer (Kingma & Ba, 2014). We randomly held out 30 image pairs from the MRI-CT dataset and 50 pairs from the MRI-SPECT dataset as the standalone test set. To ensure the robustness of our model, we repeated our experiments three times and ensured that we had different test sets in each run.

To assess the usability of our fusion framework, we further applied our method to a brain tumor classification task between LGG and HGG using the BraTS 2019 data. First, we randomly held out 40 *patients* (20 LGG and 20 HGG patients, 1152 slices in total) as a standalone test set, ensuring patient-level separation to prevent data leakage. The remaining data is used to train our model. We trained our fusion model for 25 epochs with a constant learning rate of 0.001, a mini-batch size of 16. For the classification model, we used ResNet-50 for all experiments and trained with focal loss (Lin et al., 2017) followed by (Zhou & Khalvati, 2024). We trained the model for 50 epochs with a constant learning rate of 0.001 and a mini-batch size of 8. We ran the classification experiment for three trials with different train-validation splits to ensure the robustness and reliability of our findings.

Baseline Model & Comparison. For a comprehensive comparison of image fusion results on MRI-CT and MRI-SPECT data, we evaluated against four state-of-the-art methods spanning different architectural paradigms: one CNN-based method, EH-DRAN (Zhou et al., 2024); two transformer-based methods, SwinFusion (Ma et al., 2022) and MRSCFusion (Xie et al., 2023); and a Mamba-based method MambaDFuse (Li et al., 2024). For quantitative comparisons, we select five commonly used metrics in previous works (Xie et al., 2023; Li et al., 2022; Fu et al., 2021; Chen et al., 2024; Safari et al., 2023): Peak signal-to-noise ratio (PSNR), Structural Similarity (SSIM) (Wang et al., 2004), Feature Mutual Information (Haghighat et al., 2011), Feature SSIM (FSIM) (Zhang et al., 2011), and Information Entropy (EN). For the downstream brain tumor classification task, we assessed clinical utility using Area Under the Curve (AUC), F1-Score, and Accuracy.

4. Results and Discussions

4.1. Image Fusion Results

Main Results. Figure presents qualitative comparisons of fusion results across three representative MRI-CT and MRI-SPECT test pairs. For MRI-CT fusion, we focus on anatomically challenging regions containing rich soft tissue

Table 1. Comparison between different methods on two test datasets, **bold** and underline numbers represent best and second-best values in each dataset, respectively.

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**

information from MRI and dense bone structures from CT (highlighted in red boxes). Effective fusion should preserve both the high-contrast skeletal features from CT and detailed soft tissue boundaries from MRI. Our qualitative analysis reveals distinct limitations across baseline methods. EH-DRAN exhibits significant contrast degradation, producing smoothed intensity distributions that compromise both MRI tissue detail and CT structural information. MRSCFusion demonstrates inconsistent performance, generating artifacts with undesirable pixel intensities while failing to preserve critical anatomical details such as brain contours and tissue boundaries, particularly evident in the middle sample (Figure). Although SwinFusion better preserves MRI tissue characteristics compared to MRSCFusion, it fails to maintain adequate contrast differentiation between modalities, resulting in washed-out structural boundaries. MambaDFuse suffers from substantial detail loss in MRI-derived regions and exhibits severe contrast distortion. In contrast, our proposed method demonstrates superior preservation of both modality-specific features and inter-modal contrast. The fused images exhibit edge definition between tissue types while maintaining fine-grained anatomical details from both input modalities (the first and last sample in Figure). Our approach achieves more natural-appearing intensity distributions with improved overall contrast that facilitates better visual interpretation of anatomical structures.

For the MRI-SPECT fusion task, we focus on regions exhibiting rich functional information from SPECT and complementary structural details from MRI (highlighted in red box) for a better comparison. Effective MRI-SPECT fusion requires a balance between preserving SPECT's functional metabolic information and MRI's high-resolution tissue information. Consistent with the MRI-CT results, EH-DRAN demonstrates poor contrast preservation, failing to maintain the distinctive functional signatures present in SPECT imaging. MRSCFusion introduces significant intensity artifacts and exhibits substantial loss of MRI texture information, potentially obscuring critical anatomical boundaries. While SwinFusion achieves reasonable overall fusion quality, the high-intensity functional regions from SPECT tend to overwhelm and blur fine-grained MRI structural details. MambaDFuse shows improved functional information preservation from SPECT compared to other baselines, but continues to suffer from detail loss in MRI-derived regions. Our proposed method demonstrates superior performance by maintaining an optimal balance between functional and structural information. The fused images successfully preserve SPECT's functional characteristics while retaining MRI's detailed anatomical structure and tissue contrast.

The quantitative metrics, computed over three distinct test sets, are reported with mean values and standard deviations in Table 1. Our proposed method achieves the best performance in terms of SSIM, FMI, FSIM, and Information Entropy for the MRI-CT fusion task. The high FMI, FSIM, and Information Entropy scores indicate that our fused images maintain superior structural similarity and contain richer information. Although our method shows a slightly lower PSNR score compared to EH-DRAN, this may be attributed to our method's emphasis on preserving complementary information rather than pixel-level reconstruction fidelity. In contrast, our approach balances the contributions from both MRI and CT images. For the MRI-SPECT fusion task, our method consistently surpasses baseline methods in PSNR, SSIM, FSIM, and Information Entropy. Despite a slightly lower FMI than EH-DRAN, all other metrics demonstrate that our approach effectively preserves more functional and morphological information from MRI and SPECT images. Also, the superior Information Entropy score further confirms enhanced information retention from both modalities. These quantitative results strongly corroborate the qualitative fusion improvements observed in our visual analysis.

Fusion Time. Next, we assess the model complexity by examining the total number of trainable parameters and the image fusion time for each image pair using the MRI-SPECT dataset, as detailed in Table 2. The MRI-SPECT dataset is

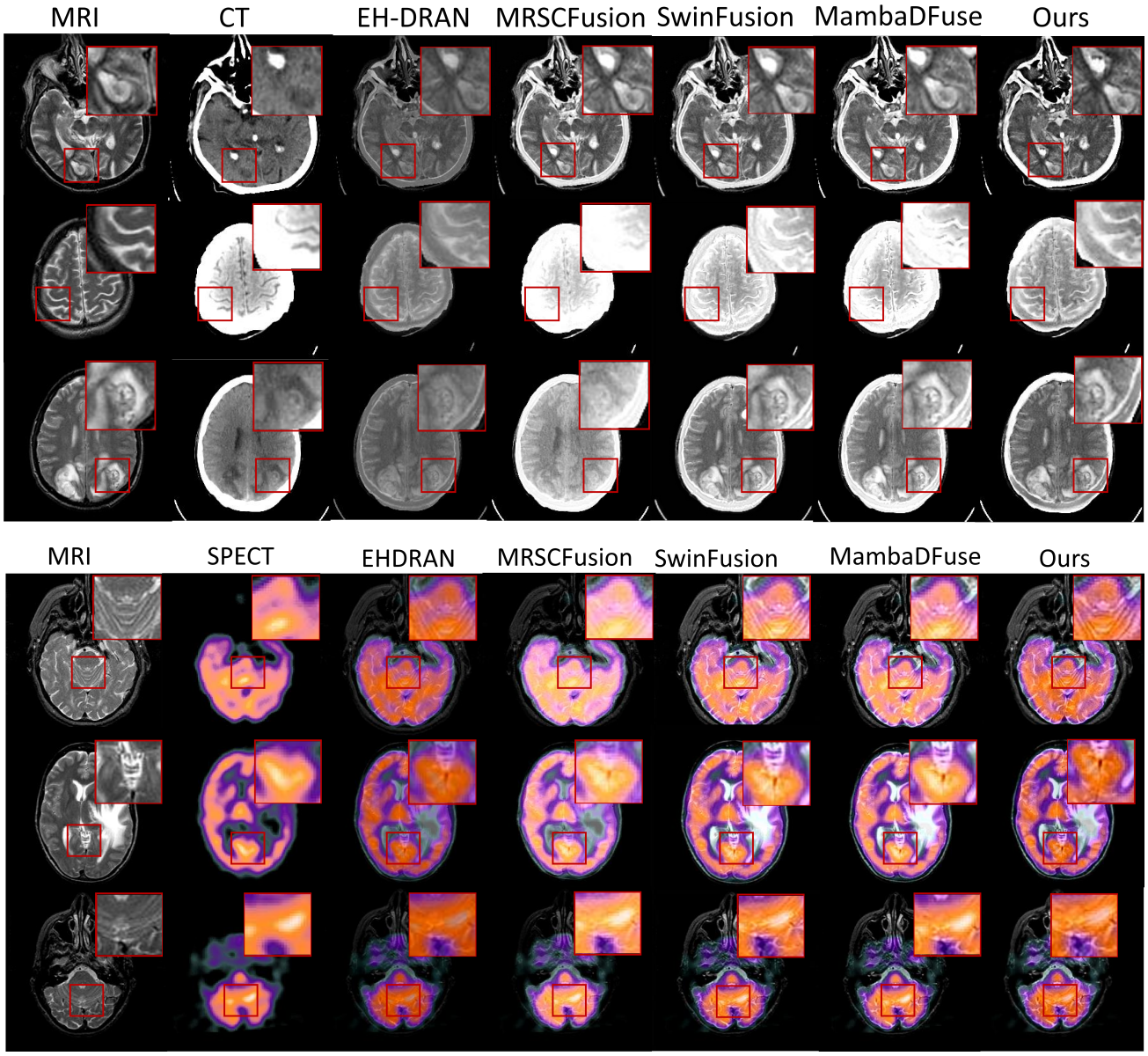

Figure 2. Qualitative results for MRI-CT (top three rows) and MRI-SPECT (bottom three rows) fusion task. We randomly select three sample pairs from both test sets and show the fusion results across different methods. Zoom in for a better view.

Table 2. Comparison between different methods in average inference time on MRI-SPECT dataset

	EH-DRAN	SwinFusion	MRSCFusion	MambaDFuse	Ours
Params(M)	0.50	0.97	23.00	1.34	4.05
Time(s)	1.26	1.31	2.85	0.66	0.96

selected due to its larger number of image pairs and its representation of a more complex task, which closely mirrors real-world scenarios. Our proposed method achieves efficient performance with 4.05M parameters and 0.96 seconds per image pair, demonstrating a favorable balance between model complexity and computational efficiency. While MambaDFuse achieves the fastest inference time (0.66s), our method maintains competitive speed while providing superior fusion quality, as demonstrated in our quantitative results. The results suggest strong potential for **real-time** clinical applications.

Ablation Study. We conducted comprehensive ablation experiments to validate the effectiveness of our Cross-Modal Channel Attention (CMCA) module across MRI-CT and MRI-SPECT fusion tasks. Our hypothesis is that CMCA en-

Table 3. Ablations on the Cross-Modal Channel Attention (CMCA) module on both datasets. Ours w/o represents the model without CMCA, Ours represents the model described in Section 2

Dataset	Method	PSNR↑	SSIM↑	FMI↑	FSIM↑	EN↑
MRI-CT	Ours w/o CMCA	15.967±0.211	0.761±0.006	0.876±0.004	0.813±0.002	14.621±0.527
	Ours	**16.519±0.352**	**0.783±0.005**	**0.883±0.003**	**0.820±0.001**	**15.213±0.069**
MRI-SPECT	Ours w/o CMCA	19.693±0.162	0.748±0.007	0.853±0.005	0.839±0.003	14.691±0.323
	Ours	**21.561±0.067**	**0.759±0.009**	**0.856±0.003**	**0.848±0.002**	**14.871±0.334**

hances modality-specific features by leveraging cross-modal channel importance, enabling the model to adaptively select the most informative channels for optimal fusion performance. Table 3 presents quantitative results comparing our full model against the baseline (without CMCA). The integration of CMCA consistently improves performance across all evaluation metrics on both datasets. Specifically, we observe substantial improvements in structural preservation metrics such as SSIM by 0.011 and FSIM by 0.009. These gains demonstrate CMCA's effectiveness in preserving complementary structural information from both source modalities.

4.2. Classification Results

Table 4. Comparison of LGG/HGG brain tumor classification performance between different methods. Values are reported as mean±standard deviation.

	AUC↑	F1-Score↑	Accuracy↑
T2 (1-channel)	0.722±0.021	0.703±0.018	0.604±0.037
FLAIR (1-channel)	0.727±0.024	0.701±0.008	0.611±0.017
T2+FLAIR (2-channel)	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	**0.790±0.013**	**0.778±0.023**	**0.665±0.004**

To validate the clinical utility of our proposed fusion framework, we conducted a downstream brain tumor classification task to distinguish between high-grade glioma (HGG) and low-grade glioma (LGG). Following (Zhou et al., 2024), we utilized T2-weighted and FLAIR sequences for this evaluation. We compared five different input configurations: single-modality approaches using either T2 or FLAIR independently, dual-modality approach using channel-wise concatenation of T2 and FLAIR, fusion using EH-DRAN baseline, and our proposed ClinicalFMamba fusion method. The classification results presented in Table 4 demonstrate the superior performance of our fusion framework across all evaluation metrics. Our ClinicalFMamba method achieves the highest performance with an AUC of 0.790, F1-score of 0.778, and Accuracy of 0.665, representing substantial improvements over single-modality baselines and dual-modality concatenation. Notably, our method also outperforms the EH-DRAN fusion baseline by 2.1% in AUC and 5.5% in F1-Score, demonstrating the effectiveness of our fusion strategy. These results validate that our fusion framework successfully integrates complementary information from T2 and FLAIR modalities, enhancing tumor details and tissue contrast characteristics that are critical for accurate glioma grading. The consistent performance gains suggest strong potential for clinical deployment in computer-aided diagnosis systems for brain tumor assessment.

5. Conclusions

In this work, we proposed a novel end-to-end CNN-Mamba hybrid architecture for effective multimodal medical image fusion. We integrated Dilated Gated Convolution Blocks to capture multiscale fine-grained details from both modalities and leveraged a latent Mamba model incorporating four-directional and fusion Mamba blocks to learn long-range dependencies and perform cross-modal feature integration in latent space. Unlike the previous two-stage approaches, our framework eliminates the need for separate fusion preprocessing, enabling direct end-to-end optimization. Extensive evaluations demonstrate that our method outperforms several baseline approaches in both subjective visual quality and objective fusion metrics while achieving real-time processing speeds. The significant improvement in downstream brain tumor classification further validates the clinical utility of our fusion framework. We envision our approach being applied to disease localization tasks, radiotherapy treatment planning, and surgical navigation in real-world clinical settings.

For future work, we plan to extend our method to 3D medical image fusion, as volumetric data is more prevalent in clinical practice and would enable more comprehensive multimodal analysis.

Impact Statement

This research contributes to the growing field of AI-assisted medical imaging, potentially reducing diagnostic errors and improving healthcare delivery efficiency. Real-time multimodal image fusion can significantly enhance diagnostic workflows by providing radiologists with comprehensive visualizations that combine complementary anatomical and functional information. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here. While our method improves diagnostic accuracy, we acknowledge that AI-assisted diagno-

sis should complement rather than replace clinical expertise. The improved fusion quality and downstream classification performance must be validated through extensive clinical trials before deployment.

References

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J. S., Freymann, J. B., Farahani, K., and Davatzikos, C. Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. *Scientific data*, 4(1):1–13, 2017.

Bakas, S., Reyes, M., Jakab, A., Bauer, S., Rempfler, M., Crimi, A., Shinohara, R. T., Berger, C., Ha, S. M., Rozycki, M., et al. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge. *arXiv preprint arXiv:1811.02629*, 2018.

Chen, W., Li, Q., Zhang, H., Sun, K., Sun, W., Jiao, Z., and Ni, X. Mr–ct image fusion method of intracranial tumors based on res2net. *BMC Medical Imaging*, 24(1): 169, 2024.

Fu, J., Li, W., Du, J., and Huang, Y. A multiscale residual pyramid attention network for medical image fusion. *Biomedical Signal Processing and Control*, 66:102488, 2021.

Gu, A. and Dao, T. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

Haghighat, M. B. A., Aghagolzadeh, A., and Seyedarabi, H. A non-reference image fusion metric based on mutual information of image features. *Computers & Electrical Engineering*, 37(5):744–756, 2011.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Li, G., Huang, Q., Wang, W., and Liu, L. Selective and multi-scale fusion mamba for medical image segmentation. *Expert Systems with Applications*, 261:125518, 2025.

Li, H. and Wu, X.-J. Densefuse: A fusion approach to infrared and visible images. *IEEE Transactions on Image Processing*, 28(5):2614–2623, 2019. doi: 10.1109/TIP.2018.2887342.

Li, H., Xiong, P., An, J., and Wang, L. Pyramid attention network for semantic segmentation. *arXiv preprint arXiv:1805.10180*, 2018.

Li, W., Peng, X., Fu, J., Wang, G., Huang, Y., and Chao, F. A multiscale double-branch residual attention network for anatomical–functional medical image fusion. *Computers in Biology and Medicine*, 141:105005, 2022.

Li, Z., Pan, H., Zhang, K., Wang, Y., and Yu, F. Mambadfuse: A mamba-based dual-phase model for multi-modality image fusion. *arXiv preprint arXiv:2404.08406*, 2024.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988, 2017.

Liu, H., Dai, Z., So, D., and Le, Q. V. Pay attention to mlps. *Advances in neural information processing systems*, 34: 9204–9215, 2021.

Ma, J., Tang, L., Fan, F., Huang, J., Mei, X., and 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.

Ma, J., Li, F., and Wang, B. U-mamba: Enhancing long-range dependency for biomedical image segmentation. *arXiv preprint arXiv:2401.04722*, 2024.

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.

Peng, S., Zhu, X., Deng, H., Lei, Z., and Deng, L.-J. Fusionmamba: Efficient image fusion with state space model. *arXiv preprint arXiv:2404.07932*, 2024.

Safari, M., Fatemi, A., and Archambault, L. Medfusiongan: multimodal medical image fusion using an unsupervised deep generative adversarial network. *BMC Medical Imaging*, 23(1):203, 2023.

Song, X., Wu, X.-J., and Li, H. Msdnet for medical image fusion. In *International conference on image and graphics*, pp. 278–288. Springer, 2019.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Wang, G., Li, W., Gao, X., Xiao, B., and Du, J. Functional and anatomical image fusion based on gradient enhanced decomposition model. *IEEE Transactions on Instrumentation and Measurement*, 71:1–14, 2022.

Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Xie, X., Zhang, X., Ye, S., Xiong, D., Ouyang, L., Yang, B., Zhou, H., and Wan, Y. Mrscfusion: Joint residual swin transformer and multiscale cnn for unsupervised multimodal medical image fusion. *IEEE Transactions on Instrumentation and Measurement*, 72:1–17, 2023. doi: 10.1109/TIM.2023.3317470.

Xie, X., Cui, Y., Ieong, C.-I., Tan, T., Zhang, X., Zheng, X., and Yu, Z. Fusionmamba: Dynamic feature enhancement for multimodal image fusion with mamba. *arXiv preprint arXiv:2404.09498*, 2024.

Xing, Z., Ye, T., Yang, Y., Liu, G., and Zhu, L. Segmamba: Long-range sequential modeling mamba for 3d medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 578–588. Springer, 2024.

Yu, F. and Koltun, V. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015.

Yue, Y. and Li, Z. Medmamba: Vision mamba for medical image classification. *arXiv preprint arXiv:2403.03849*, 2024.

Zhang, L., Zhang, L., Mou, X., and Zhang, D. Fsim: A feature similarity index for image quality assessment. *IEEE transactions on Image Processing*, 20(8):2378–2386, 2011.

Zhou, M. and Khalvati, F. Conditional generation of 3d brain tumor regions via vqgan and temporal-agnostic masked transformer. In *Medical Imaging with Deep Learning*, 2024.

Zhou, M., Zhang, Y., Xu, X., Wang, J., and Khalvati, F. Edge-enhanced dilated residual attention network for multimodal medical image fusion. In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 4108–4111. IEEE, 2024.

A. Details on Fusion Mamba Block

To effectively integrate heterogeneous multimodal information, (Peng et al., 2024) extended the original Mamba architecture to handle dual inputs, introducing the Fusion State Space Model (FSSM). The FSSM block employs an asymmetric processing strategy where one input modality (x_a or x_b) generates the projection matrices and timescale parameters (B, C, Δ), while the complementary input serves as the sequence for selective scan processing (y_a or y_b = SSM($\bar{A}, \bar{B}, C$)(x_a or x_b)). The complete FusionMamba block incorporates eight FSSM components organized in a symmetric architecture, with four directional variants for each input modality. Using CT-MRI fusion as an illustrative example, given input feature maps $F_a, F_b \in \mathcal{R}^{H\times W\times C}$, representing CT and MRI modalities respectively, the block generates two sets of feature representations for F_a, F_b following the similar procedure done in the four-directional Mamba block:

$$\mathbf{X}^a, \mathbf{Z}^a = \text{Conv}_x^1(\text{Norm}(\mathbf{F}_{in}^a)), \text{Conv}_z^1(\text{Norm}(\mathbf{F}_{in}^a)); \tag{3}$$

$$\mathbf{X}^b, \mathbf{Z}^b = \text{Conv}_x^1(\text{Norm}(\mathbf{F}_{in}^b)), \text{Conv}_z^1(\text{Norm}(\mathbf{F}_{in}^b)). \tag{4}$$

Next, X^a and X^b are flattened along four scanning directions, as shown in the upper right of Figure 1. The resulting 1D sequences are then processed through corresponding FSSM blocks to enable cross-modal information integration and long-range dependency modeling:

$$\begin{cases} \mathbf{x}_i^a, \mathbf{x}_i^b = \text{Flatten}_i(\mathbf{X}^a), \text{Flatten}_i(\mathbf{X}^b), \\ \mathbf{y}_i^a, \mathbf{y}_i^b = \mathbf{FSSM}_i^a(\mathbf{x}_i^a, \mathbf{x}_i^b), \mathbf{FSSM}_i^b(\mathbf{x}_i^b, \mathbf{x}_i^a). \end{cases} \quad i = 1, 2, 3, 4. \tag{5}$$

where $\mathbf{FSSM}^a$ and $\mathbf{FSSM}^b$ represent symmetric processing blocks that handle X^a and X^b respectively. The outputs from these blocks are then processed independently and reshaped back to spatial dimensions, producing two enhanced feature maps $Y^a, Y^b \in \mathcal{R}^{H\times W\times C}$. These complementary representations are subsequently combined to generate the final output F_{out}:

$$\begin{aligned} \mathbf{Y}^a, \mathbf{Y}^b &= \sum_{i=1}^{4} \text{Unflatten}_i(\mathbf{y}_i^a), \sum_{i=1}^{4} \text{Unflatten}_i(\mathbf{y}_i^b), \\ \mathbf{F}_{\text{out}}^a &= \text{Conv}_o^a(\mathbf{Y}^a \cdot \text{SiLU}(\mathbf{Z}^a)) + \mathbf{F}_{\text{in}}^a, \\ \mathbf{F}_{\text{out}}^b &= \text{Conv}_o^b(\mathbf{Y}^b \cdot \text{SiLU}(\mathbf{Z}^b)) + \mathbf{F}_{\text{in}}^b, \\ \mathbf{F}_{\text{out}} &= \text{Conv}_f(\mathbf{F}_{\text{out}}^a + \mathbf{F}_{\text{out}}^b). \end{aligned} \tag{6}$$

where $Conv_o$ here represents a 1×1 convolution layer.