Brain tumour segmentation using multimodal MRI

Qinqing Li
q.li@degould.com
Zixian Zhao
zz541@exeter.ac.uk
Xingchen Zhang*
x.zhang12@exeter.ac.uk

Fusion Intelligence Laboratory
Department of Computer Science
University of Exeter
Exeter, EX4 4RN, United Kingdom

Abstract

Accurate brain tumour segmentation from multimodal MRI is critical for diagnosis and treatment planning. While recent approaches have explored explicit fusion modules to integrate complementary information across modalities, the relative effectiveness of different fusion strategies, especially between decoupled and joint training, remains unclear. This work provides a systematic comparison of three architectural paradigms: segmentation model without fusion module using UNet, decoupled fusion with sequential training stages, and joint optimisation with combined segmentation and fusion objectives. Our empirical evaluation reveals that neither fusion approach consistently outperforms the baseline UNet, with decoupled training showing marginal improvements on certain tumour regions while joint optimisation with dynamic loss weighting fails to converge effectively. These findings suggest that incorporating explicit fusion modules does not guarantee improved segmentation performance, and that the design of joint training objectives for multi-modal medical image analysis requires careful consideration of loss balancing and optimisation dynamics.

1 Introduction

Accurate segmentation of brain tumours from multimodal Magnetic Resonance Imaging (MRI) is critical for treatment planning, surgical guidance, and monitoring disease progression. However, manual delineation of tumour regions is time-consuming and subject to inter-observer variability, motivating the development of automated segmentation methods. Clinical MRI protocols typically acquire multiple complementary modalities, including T1-weighted (T1), contrast-enhanced T1-weighted (T1-CE), T2-weighted (T2), and T2-weighted Fluid-Attenuated Inversion Recovery (T2-FLAIR), each highlighting different tissue properties and pathological features (see the left panel of Fig. 1). For instance, tumours exhibit distinct contrast patterns in T1 sequences, while pathological changes such as edema are more prominent in T2 sequences (see the right panel of Fig. 1). Effectively integrating information from these modalities is therefore essential for robust and reliable tumour segmentation.

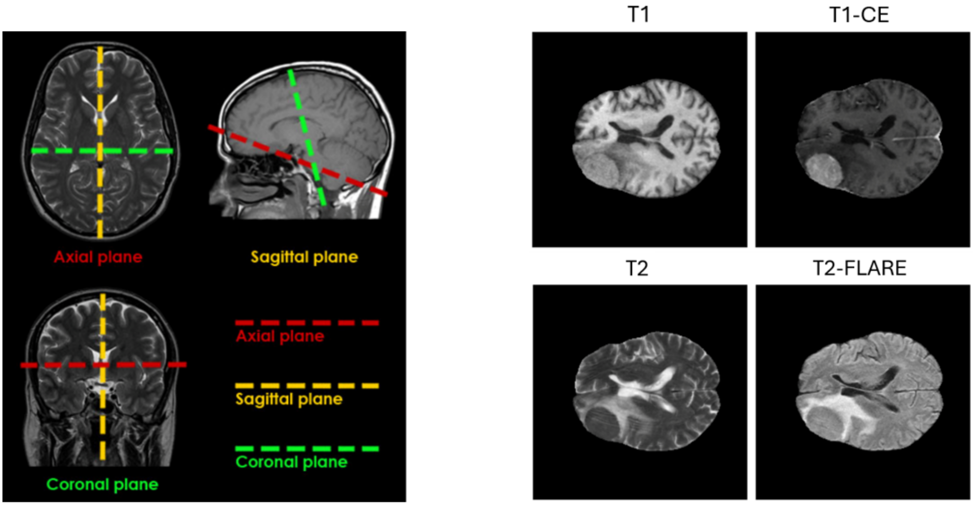


Figure 1: MRI fundamentals for brain tumour segmentation. Left: The three standard anatomical planes used in MRI acquisition (axial, coronal, and sagittal), corresponding to the horizontal, frontal, and side views, respectively. Right: Four MRI modalities from the BraTS-MEN dataset [17], illustrating complementary tissue contrasts: T1-weighted (T1), contrast-enhanced T1-weighted (T1CE), T2-weighted (T2), and T2-FLAIR, each emphasizing different anatomical structures and pathological features.

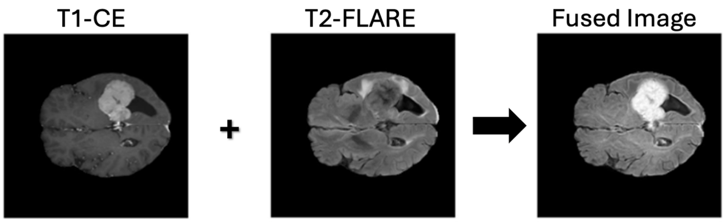


Figure 2: Example of multimodal MRI fusion. T1-CE and T2-FLARE modalities are combined to generate a fused image that preserves complementary information from both input modalities for downstream segmentation.

Most existing medical image segmentation approaches concatenate multiple modalities as multi-channel inputs to segmentation networks [10, 21]. In contrast, recent task-driven fusion methods adopt either decoupled training [5] or joint optimisation strategies [11]. The latter typically employ explicit fusion modules that combine complementary modalities into a unified representation before segmentation, as shown in Fig. 2. However, these fusion strategies have not been systematically compared under controlled experimental settings, and it remains unclear whether explicit fusion modules offer tangible advantages over standard concatenation for brain tumour segmentation.

In this work, we conduct a systematic comparison of three architectural strategies for multimodal brain tumour segmentation using the BraTS-MEN 2023 dataset [17] (Fig. 3): (1) baseline UNet with concatenated inputs, (2) decoupled fusion with separate training stages, and (3) joint optimisation with combined loss $\mathcal{L} = \mathcal{L}_{seg} + \lambda \mathcal{L}_{fusion}$. We evaluate performance using Dice scores across three tumour regions: enhanced tumour (ET), tumour core (TC), and whole tumour (WT), providing empirical evidence for fusion architecture design decisions in medical image segmentation.

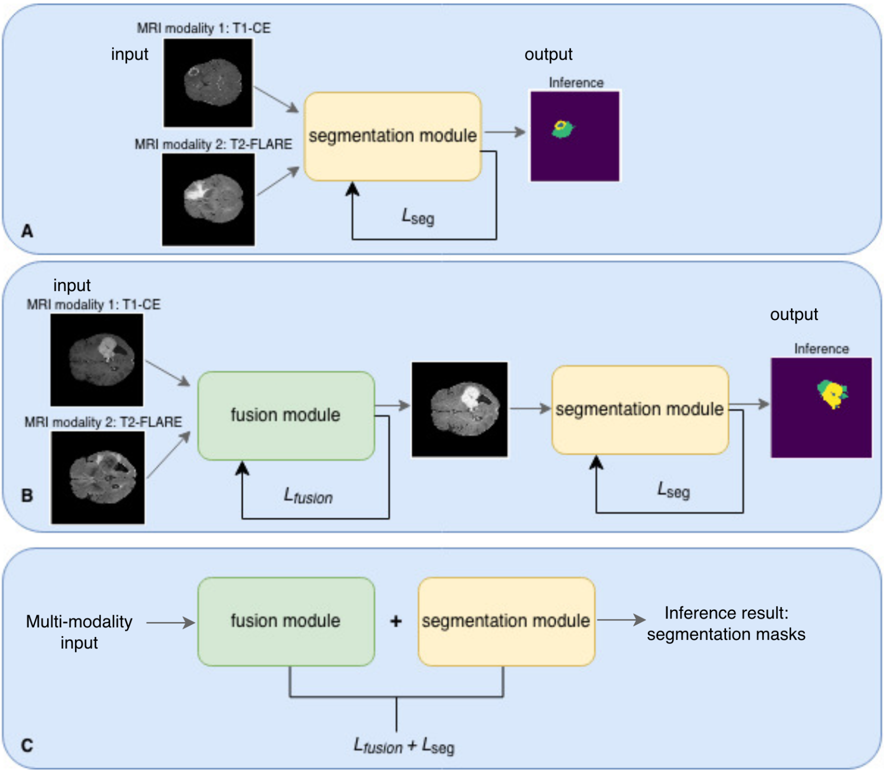


Figure 3: Three architectural strategies compared in this study: (a) segmentation model without fusion module, (b) decoupled fusion, and (c) joint optimisation.

2 Related work

2.1 Image Fusion

Image fusion [19, 26] aims to fuse complementary visual information present in different image modalities into a single image, thereby achieving visual information fusion. Various deep learning-based methods have been proposed in recent years, such as CNN-based methods [15], autoencoder-based methods [8], GAN-based methods [13], transformer-based methods [18], and diffusion model-based methods [22]. Common image fusion applications include visible-infrared image fusion [27, 28], multi-focus image fusion [25], multi-exposure image fusion [24], and multimodal medical image fusion [1]. Image fusion plays a crucial role in various fields, including computer vision, surveillance, autonomous driving, and robotics. In particular, task-driven image fusion has attracted increasing attention in recent years, especially in visible–infrared image fusion.

2.2 Medical image fusion

Medical image fusion models have gained significant attention in recent years. Methods such as EMFusion [20], TFS-Diff [7], and SIMFusion [23] focus on optimising the quality of fused images with respect to traditional image fusion metrics, including peak signal-to-noise ratio (PSNR) and visual information fidelity (VIF). Notably, SIMFusion employs a pre-trained segmentation module to guide the fusion process, achieving improved performance in terms of PSNR and VIF.

However, these methods optimise image fusion quality independently from downstream clinical tasks, which may not guarantee optimal output for subsequent analysis such as segmentation. This motivated task-driven fusion frameworks for medical images that jointly optimise downstream task performance and image quality, rather than relying solely on intermediate image-quality metrics.

2.3 Medical image segmentation

Medical image segmentation has been extensively studied using deep learning models that rely solely on segmentation networks without incorporating fusion modules [14, 23]. In contrast, an emerging research direction integrates fusion as a learnable component optimised jointly for fused-image quality and downstream segmentation performance. These task-driven approaches combine segmentation-specific losses (Dice, cross-entropy) with image-quality objectives, rather than optimising for either in isolation. Notable methods include GAN-based fusion frameworks [3, 12], the sequential approach Fuse4Seg [5], which optimises fusion and segmentation modules in separate stages, and the joint optimisation framework SF-Net [11].

3 Method

This work compares three segmentation strategies for multimodal brain tumour segmentation, as illustrated in Fig. 3: (A) baseline segmentation without explicit fusion, (B) decoupled fusion with sequential training, and (C) joint optimisation combining fusion and segmentation objectives. All three architectures are evaluated to determine the most effective approach for integrating multimodal MRI information. Architectures B and C employ LRRNet [9] as the fusion module and UNet [16] as the segmentation module, which are described below. The UNet is also utilized in Architecture A.

3.1 Fusion Module

The fusion module we used in architecture B and C is LRRNet [9], which employs two parallel LISTA [4] algorithms to extract sparse features S and low-rank features L from each input modality x. Each LISTA algorithm performs iterative soft-thresholding over K iterations, updating features via $Z^{(k+1)} = \eta_{l1}(\lambda_z Z^{(k)} + W_d^T(x - W_d Z^{(k)}), \lambda_\theta)$, where the step size λ_z and thresholding parameter λ_θ are learned through backpropagation, enabling much faster convergence than traditional ISTA where $\lambda_z, \lambda_\theta$ are fixed hyperparameters. The extracted features S_1, S_2, L_1, L_2 from both modalities are then fused into S_f and L_f, producing the final fused image $I_f = S_f + L_f$.

3.2 Segmentation Module

The segmentation module we used in architecture A, B and C is a standard UNet [16] with 3 input channels (T1-CE, T2-FLAIR, and fused image) and a 4-class output. Class 0 represents background, while classes 1, 2, and 3 correspond to specific tumour regions: necrotic tumour core (NCR), peritumoral edema (ED), and enhancing tumour (ET), respectively [17]. Both the encoder and decoder consist of two stages, where each stage comprises a double convolution block applying two consecutive 3×3 convolutions with ReLU activations, followed by max-pooling (encoder) or transposed convolution (decoder). Skip connections concatenate encoder and decoder features at corresponding spatial resolutions.

3.3 Loss Function

3.3.1 Fusion Loss

For fusion loss, we adopt the multi-level perceptual loss function from LRRNet, which combines pixel-level and feature-level objectives extracted from a pre-trained VGG-16 network:

$$\mathcal{L}_{fusion} = \gamma_1 \mathcal{L}_{pixel} + \gamma_2 \mathcal{L}_{shallow} + \mathcal{L}_{middle} + \gamma_4 \mathcal{L}_{deep}, \tag{1}$$

where the pixel-level term $\mathcal{L}_{pixel} = ||I_f - I_{T2}||^2$ preserves detail information from T2-FLAIR, while shallow, middle, and deep feature losses $\mathcal{L}_{shallow}, \mathcal{L}_{middle}, \mathcal{L}_{deep}$ compare VGG features at different network depths to balance contributions from both modalities. We refer readers to LRRNet [9] for the complete formulation. Hyperparameter values are provided in Section 4.2.

3.3.2 Segmentation Loss

The segmentation loss combines cross-entropy and Dice loss:

$$\mathcal{L}_{seg} = \mathcal{L}_{CE} + \mathcal{L}_{Dice}, \tag{2}$$

where $\mathcal{L}_{CE}$ applies class weights, and $\mathcal{L}_{Dice}$ directly optimises the Dice coefficient used for evaluation.

3.3.3 Joint Optimisation

For joint training, the combined objective is:

$$\mathcal{L}_{total} = \lambda_f \mathcal{L}_{fusion} + \lambda_s \mathcal{L}_{seg}, \tag{3}$$

where λ_f and λ_s balance fusion quality and segmentation accuracy. For decoupled training, the fusion module is pre-trained using only $\mathcal{L}_{fusion}$, then fixed during segmentation training with $\mathcal{L}_{seg}$.

4 Experiments

4.1 Dataset and Evaluation Metrics

We train our model on BraTS-MEN 2023 dataset (Brain Tumour Segmentation-Meningioma Challenge) [17], totaling 1,000 patient cases. Each case includes four MRI modalities (T1,

Table 1: Dice score comparison across tumour regions. Rows 1-2: A and B. Rows 3-10: joint training (C) with dynamic weighting (row 3), static weighting variations (rows 4-9), and segmentation-only ablation (row 10). Rows 11-12: state-of-the-art methods on the same dataset for reference.

	Dice (%)			Lesion-wise Dice (%)		
Method	ET	TC	WT	ET	TC	WT
Baseline UNet	77.98	77.69	72.44	-	-	-
LRRNet then UNet	65.15	75.61	75.90	75.70	75.41	64.74
dynamic λ_f, λ_s	29.55	27.52	23.73	24.19	21.11	11.84
$\lambda_f = 5 \times 10^{-7}, \lambda_s = 1$	21.16	19.39	19.84	17.19	15.69	12.06
$\lambda_f = 1 \times 10^{-6}, \lambda_s = 1$	49.17	48.75	37.87	45.76	45.01	28.40
$\lambda_f = 5 \times 10^{-6}, \lambda_s = 1$	58.53	57.90	32.60	54.87	53.82	24.49
$\lambda_f = 1 \times 10^{-5}, \lambda_s = 1$	55.63	51.36	44.85	52.20	46.88	35.82
$\lambda_f = 5 \times 10^{-5}, \lambda_s = 1$	71.73	69.93	60.90	67.77	66.16	52.26
$\lambda_f = 1 \times 10^{-4}, \lambda_s = 1$	49.55	43.78	37.84	44.13	37.99	26.58
$\mathcal{L} = \mathcal{L}_{seg}$	50.70	48.06	30.49	47.45	43.97	21.35
STU-Net [6]	-	-	-	86.98	87.86	84.46
Ensemble model [2]	-	-	-	87.6	86.7	84.9

T1-CE, T2, T2-FLAIR) and a ground truth segmentation mask. Images are 240×240 pixels with 155 slices per modality. Wwe use only the central 10 slices (72-81) where tumours are most prevalent to train the model. The dataset was split into 80% training, 10% validation, and 10% test sets.

We evaluate segmentation performance using four standard metrics: Dice score, HD95 (95th percentile Hausdorff Distance), and their lesion-wise variants. Lesion-wise Dice and HD95 are defined as:

$$\text{Lesion-wise Dice} = \frac{\sum_{i=1}^{L} \text{Dice}(l_i)}{|\text{TP}| + |\text{FN}| + |\text{FP}|}, \tag{4}$$

$$\text{Lesion-wise HD}_{95} = \frac{\sum_{i=1}^{L} \text{HD}_{95}(l_i)}{|\text{TP}| + |\text{FN}| + |\text{FP}|}, \tag{5}$$

where L is the number of ground truth lesions and (true positive (TP) + false negative (FN)) is equal to L. A ground truth lesion is counted as TP if at least 1 predicted pixel overlaps with that ground truth lesion's mask, and is counted as FN if the model predicts zero pixels within that ground truth lesion's mask.

4.2 Implementation Details

Fusion loss hyperparameters were set to $\gamma_1 = 10$, $\gamma_2 = 1.5$, and $\gamma_4 = 2000$. Cross-entropy class weights were set to $[0.5, 2.0, 5.0, 4.0]$ for classes 0 to 3. Training employed Adam optimiser with batch size 4 and early stopping (patience 5). For sequential training, UNet used learning rate 10^{-3} while LRRNet used 10^{-4}. For joint training, both modules used learning rate 10^{-4}. Models were trained for up to 50 epochs, with early stopping applied based on validation Dice score.

For joint optimisation, we explored fixed and dynamic weights for balancing fusion and segmentation objectives. The fixed weight evaluated are $\lambda_f \in \{5\times10^{-7}, 1\times10^{-6}, 5\times10^{-6}, 1\times10^{-5}, 5\times10^{-5}, 1\times10^{-4}\}$ with $\lambda_s = 1$. The dynamic weight approach adjusts weights each epoch defined by:

$$\lambda_f = \frac{\eta_f \exp(r_f/T)}{\eta_f \exp(r_f/T) + \eta_s \exp(r_s/T)}, \quad \lambda_s = 1 - \lambda_f \tag{6}$$

where $r_f = \mathcal{L}_{fusion}^{(t)} / \mathcal{L}_{fusion}^{(t-1)}$ measures the relative change in fusion loss between consecutive epochs, with $\eta_f = \eta_s = 1.0$ and $T = 1.0$. All models were implemented in PyTorch, with training on a single NVIDIA A100 GPU (40GB).

4.3 Results

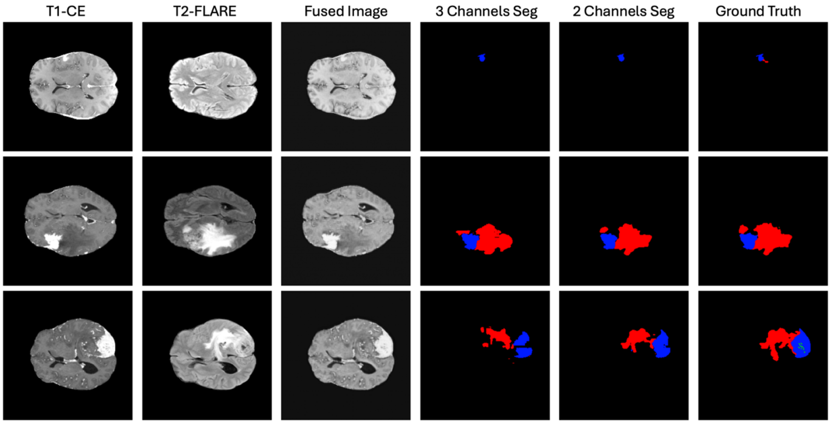
Performance across different model architectures and weighting schemes for joint fusion is summarised in Tables 1 and 2. The baseline UNet (A) achieved robust performance; the decoupled approach (B) demonstrated modest improvement on the WT region but yielded no significant overall gains. Among joint training (C) configurations, the dynamic weighting scheme underperformed relative to baseline, while the static weighting showed improvements. Segmentation visual results is shown in Fig. 4, in which the static weighting also showed closer segmentation prediction compared to ground truth.

These results contradict our initial hypothesis that joint training would improve segmentation performance. Our findings diverge notably from SF-Net [11], which employs a comparable joint training strategy and achieves SOTA performance on brain MRI segmentation using the same dataset. We hypothesize that this discrepancy may stem from several factors: (1) architectural incompatibilities between our fusion and segmentation modules, or (2) training instabilities that were not effectively mitigated in our implementation.

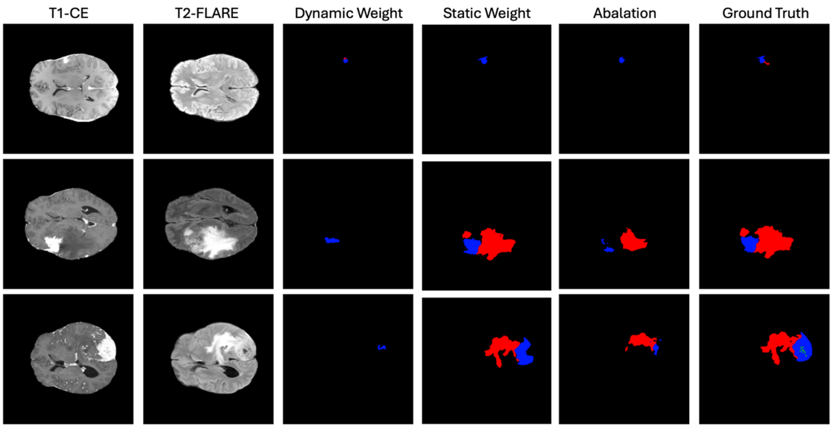
Table 2: HD95 comparison across tumour regions

	HD95 (mm)			Lesion-wise HD95 (mm)		
Method	ET	TC	WT	ET	TC	WT
Baseline UNet	77.98	77.69	72.44	-	-	-
dynamic λ_f, λ_s	17.98	21.21	44.25	248.86	259.80	294.88
$\lambda_f = 5\times10^{-7}, \lambda_s = 1$	11.72	16.82	23.40	161.54	193.76	212.74
$\lambda_f = 1\times10^{-6}, \lambda_s = 1$	35.41	38.82	42.80	279.05	282.98	304.45
$\lambda_f = 5\times10^{-6}, \lambda_s = 1$	5.59	5.94	24.79	142.56	145.98	249.96
$\lambda_f = 1\times10^{-5}, \lambda_s = 1$	7.35	9.92	19.58	153.44	170.55	212.52
$\lambda_f = 5\times10^{-5}, \lambda_s = 1$	4.84	6.98	13.72	101.39	106.75	156.89
$\lambda_f = 1\times10^{-4}, \lambda_s = 1$	9.94	16.35	24.79	183.75	202.80	244.11
$\mathcal{L} = \mathcal{L}_{seg}$	6.46	8.96	29.26	169.89	182.03	266.48
STU-Net [6]	-	-	-	33.45	29.88	41.20
Ensemble model [2]	-	-	-	37.98	38.68	42.9

However, SF-Net's strong performance on the identical dataset demonstrates that joint training remains a viable approach when carefully designed. In contrast, architectures A and B provide more stable and predictable performance. Future work should investigate more appropriate weighting strategies, alternative fusion mechanisms, and training stabilisation techniques to bridge the performance gap between our approach and SOTA joint training methods.



(a)



(b)

Figure 4: Qualitative segmentation results comparing different approaches. (a) Baseline comparison showing input modalities (T1CE, T2-FLAIR), LRRNet fusion output (fused images), joint model prediction, baseline UNet prediction, and ground truth. (b) Joint training variants showing fragmented predictions across all loss configurations, with segmentation-only ablation ($\mathcal{L} = \mathcal{L}_{seg}$) showing partial improvement but still underperforming compared to ground truth.

5 Conclusion

This study investigated the effectiveness of joint training for image fusion and segmentation tasks using a combined LRRNet-UNet architecture. Our experimental results reveal several key insights into multi-task learning for medical image analysis.

The decoupled training strategy, where fusion and segmentation models were trained sequentially, achieved Dice scores of 65.15%, 75.61%, and 75.90% for ET, TC, and WT regions, respectively. In contrast, our best joint training configuration yielded 74.41%, 70.84%, and 55.95% for the same regions. While the joint approach showed competitive performance on ET regions, it underperformed on WT and TC segmentation, resulting in an overall inferior performance compared to the decoupled strategy.

These findings suggest that simultaneous optimisation of fusion and segmentation objectives presents greater optimisation challenges than sequential single-objective optimisation. This outcome aligns with theoretical expectations: joint training introduces additional complexity through dynamic weight parameters, effectively expanding the hyperparameter search space and making convergence more difficult to achieve. The increased dimensionality of the optimisation landscape, combined with the inherently non-convex nature of segmentation loss functions such as Dice loss, compounds the difficulty in reaching global optima.

Acknowledgment

The authors would like to acknowledge the use of the University of Exeter's Advanced Research Computing facilities in carrying out this work.

References

[1] Muhammad Asad Azam, Khurram Bilal Khan, Salman Salahuddin, Ehsan Rehman, Shahzad Ali Khan, Muhammad Awais Khan, Sherif Kadry, and Amir H. Gandomi. 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.

[2] Daniel Capellán-Martín, Zhifan Jiang, Abhijeet Parida, Xinyang Liu, Van Lam, Hareem Nisar, Austin Tapp, Sarah Elsharkawi, Maria J Ledesma-Carbayo, Syed Muhammad Anwar, et al. Model ensemble for brain tumor segmentation in magnetic resonance imaging. In *International Challenge on Cross-Modality Domain Adaptation for Medical Image Segmentation*, pages 221–232. Springer, 2024.

[3] Chao Fan, Hao Lin, and Yingying Qiu. U-patch gan: A medical image fusion method based on gan. *Journal of Digital Imaging*, 36(1):339–355, 2023.

[4] Karol Gregor and Yann LeCun. Learning Fast Approximations of Sparse Coding. In *Proceedings of the 27th International Conference on Machine Learning (ICML)*, pages 399–406, 2010.

[5] Yuchen Guo and Weifeng Su. Fuse4Seg: Image-Level Fusion Based Multi-Modality Medical Image Segmentation. *arXiv preprint arXiv:2409.10328*, 2024.

[6] Zhen Huang, Jing Ye, Hao Wang, Zhiqiang Deng, Yifan Su, Tao Li, Jian Cheng, Jie Chen, Sheng Guo, Yelong Shen, and Jie He. Evaluating STU-Net for Brain Tumor Segmentation. In *Brain Tumor Segmentation, and Cross-Modality Domain Adaptation for Medical Image Segmentation*, pages 140–151, 2024.

[7] Yichen Jie, Fang Zhou, Hao Tan, Guanghui Wang, Xiang Cheng, and Xin Li. Tri-Modal Medical Image Fusion Based on Cartoon-Texture Decomposition and Sparse Representation. *SSRN Electronic Journal*, 2022.

[8] Hui Li and Xiao-Jun Wu. DenseFuse: A fusion approach to infrared and visible images. *IEEE Transactions on Image Processing*, 28(5):2614–2623, 2019.

[9] Hui Li, Tianyang Xu, Xiao-Jun Wu, Jiwen Lu, and Josef Kittler. LRRNet: A novel representation learning guided fusion network for infrared and visible images. *IEEE transactions on pattern analysis and machine intelligence*, 45(9):11040–11052, 2023.

[10] Chenyu Liu, Wangbin Ding, Lei Li, Zhen Zhang, Chenhao Pei, Liqin Huang, and Xiahai Zhuang. Brain tumor segmentation network using attention-based fusion and spatial relationship constraint. In *International MICCAI Brainlesion Workshop*, pages 219–229. Springer, 2020.

[11] Yu Liu, Fuhao Mu, Yu Shi, and Xun Chen. SF-Net: A Multi-Task Model for Brain Tumor Segmentation in Multimodal MRI via Image Fusion. *IEEE Signal Processing Letters*, 29:1799–1803, 2022.

[12] Yu Liu, Yu Shi, Fuhao Mu, Juan Cheng, and Xun Chen. Glioma Segmentation-Oriented Multi-Modal MR Image Fusion With Adversarial Learning. *IEEE/CAA Journal of Automatica Sinica*, 9(8):1528–1531, 2022.

[13] Jiayi Ma, Han Xu, Junjun Jiang, Xiaoguang Mei, and Xiao-Ping Zhang. DDcGAN: A dual-discriminator conditional generative adversarial network for multi-resolution image fusion. *IEEE Transactions on Image Processing*, 29:4980–4995, 2020.

[14] Fadillah Maani, Anees Ur Rehman Hashmi, Mariam Aljuboory, Numan Saeed, Ikboljon Sobirov, and Mohammad Yaqub. Advanced Tumor Segmentation in Medical Imaging: An Ensemble Approach for BraTS 2023 Adult Glioma and Pediatric Tumor Tasks. In *Lecture Notes in Computer Science*, pages 264–277, 2024.

[15] K Ram Prabhakar, V Sai Srikar, and R Venkatesh Babu. DeepFuse: A Deep Unsupervised Approach for Exposure Fusion with Extreme Exposure Image Pairs. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 4724–4732, 2017.

[16] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[17] Sage Bionetworks. BraTS 2023 Challenge. https://www.synapse.org/Synapse:syn51156910/wiki/622351, 2023. Accessed 23 Oct. 2025.

[18] Vibashan Vs, Jeya Maria Jose Valanarasu, Poojan Oza, and Vishal M Patel. Image fusion transformer. In *2022 IEEE International conference on image processing (ICIP)*, pages 3566–3570. IEEE, 2022.

[19] Gang Xiao, Durga Prasad Bavirisetti, Gang Liu, and Xingchen Zhang. *Image Fusion*. Springer, 2020.

[20] Han Xu and Jiayi Ma. EMFusion: An Unsupervised Enhanced Medical Image Fusion Network. *Information Fusion*, 76:177–186, 2021.

[21] Siwei Yang, Xianhang Li, Jieru Mei, Jieneng Chen, Cihang Xie, and Yuyin Zhou. 3D-TransUNet for brain metastases segmentation in the BraTS2023 challenge. In *International Challenge on Cross-Modality Domain Adaptation for Medical Image Segmentation*, pages 190–199. Springer, 2023.

[22] Jun Yue, Leyuan Fang, Shaobo Xia, Yue Deng, and Jiayi Ma. Dif-fusion: Toward high color fidelity in infrared and visible image fusion with diffusion models. *IEEE Transactions on Image Processing*, 32:5705–5720, 2023.

[23] Xiaowen Zhang, Aiping Liu, Gang Yang, Yu Liu, and Xun Chen. SIMFusion: A semantic information-guided modality-specific fusion network for MR Images. *Information Fusion*, 112:102560, 2024.

[24] Xingchen Zhang. Benchmarking and comparing multi-exposure image fusion algorithms. *Information Fusion*, 74:111–131, 2021.

[25] Xingchen Zhang. Deep learning-based multi-focus image fusion: A survey and a comparative study. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44 (9):4819–4838, 2021.

[26] Xingchen Zhang. Intelligence of Fusion: Deep Learning-Based Image Fusion, 2025. URL https://openreview.net/forum?id=YKorIXQPRd.

[27] Xingchen Zhang and Yiannis Demiris. Visible and Infrared Image Fusion Using Deep Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8): 10535–10554, 2023.

[28] Zixian Zhao, Andrew Howes, and Xingchen Zhang. MultiTaskVIF: Segmentation-oriented visible and infrared image fusion via multi-task learning. *arXiv preprint arXiv:2505.06665*, 2025.