SwinSR-UNETR: Multi-Task Learning of Brain Tumor Segmentation and Image Super-Resolution via Shared Swin Transformer Encoding

Jo Vianto[1], Sun-ja Yeom[2], Myeong-Eun Lee[1], Hyung-Jeong Yang[1*]
[1]Dept of AI Convergence, Chonnam National University
[2]University of Tasmania
[*]Corresponding Author

sjovianto@jnu.ac.kr, soonja.yeom@utas.edu.au, myungeun07@gmail.com, [*]hjyang@jnu.ac.kr

Abstract

Brain tumour diagnosis using Magnetic Resonance Imaging (MRI) is often challenged by the limited spatial resolution of clinically acquired scans. This limitation hampers accurate tumour boundary delineation, crucial for treatment planning and monitoring. While separate deep learning models for brain tumour segmentation and super-resolution (SR) have shown progress, handling both tasks independently leads to inefficiencies. To address this, we propose SwinSR-UNETR, a novel multi-task learning framework that jointly performs segmentation and SR on 3D MRI volumes. Built on a modified Swin UNETR backbone, our model shares a transformer-based encoder while employing task-specific decoders, enhancing both spatial detail and semantic understanding. A task-aware uncertainty loss balances optimization between segmentation and reconstruction objectives. Extensive experiments on the BraTS-GLI 2023 dataset demonstrate that SwinSR-UNETR improves both segmentation accuracy and super-resolution quality, outperforming traditional baselines on low-resolution inputs. Our results highlight the effectiveness of unified learning in addressing the challenges of low-resolution MRI-based tumour assessment.

1. Introduction

Magnetic Resonance Imaging (MRI) is widely used in brain tumour diagnosis due to its non-invasive nature and high soft-tissue contrast. However, many clinically acquired MRI scans suffer from limited spatial resolution, primarily due to hardware constraints and the need to minimize scan times. Low-resolution images can hinder accurate tumour boundary delineation, which is critical for diagnosis, treatment planning, and follow-up assessments.

While recent deep learning models have achieved notable success in brain tumour segmentation and super resolution (SR), most are designed for single-task learning, addressing each objective independently. This separation often leads to inefficiencies: segmentation models are limited by poor input quality, and SR models lack semantic awareness of tumour structures.

Integrating both tasks into a unified model poses technical challenges. High-resolution volumetric data increases memory and computation requirements, while balancing the learning objectives of segmentation and reconstruction demands careful architectural and optimization design. Moreover, ensuring that enhanced resolution contributes meaningfully to semantic accuracy requires effective cross-task feature sharing.

To address these issues, we propose SwinSR-UNETR, a multi-task learning framework based on Swin UNETR that jointly performs super resolution and tumour segmentation on brain MRI scans. SwinSR-UNETR leverages the transformer-based encoder of Swin UNETR for shared representation learning, with dual decoders specialized for each task. This design allows both tasks to benefit from global contextual information and localized features simultaneously.

Our contributions are as follows:

(1) We design an end-to-end architecture that combines super resolution and segmentation in a single unified model;

(2) We introduce a task-aware joint loss function to effectively balance reconstruction fidelity and segmentation performance;

(3) We demonstrate improved performance over single-task baselines on the BRATS dataset, both quantitatively and qualitatively.

2. Related Work

Deep learning has driven significant advances in automated brain tumour segmentation, with most approaches leveraging encoder-decoder architectures. U-Net[1] and its 3D variants are widely adopted due to their ability to preserve spatial details via skip connections. Extensions such as Attention U-Net and nnU-Net have improved performance by incorporating attention mechanisms and automated architecture tuning, respectively.

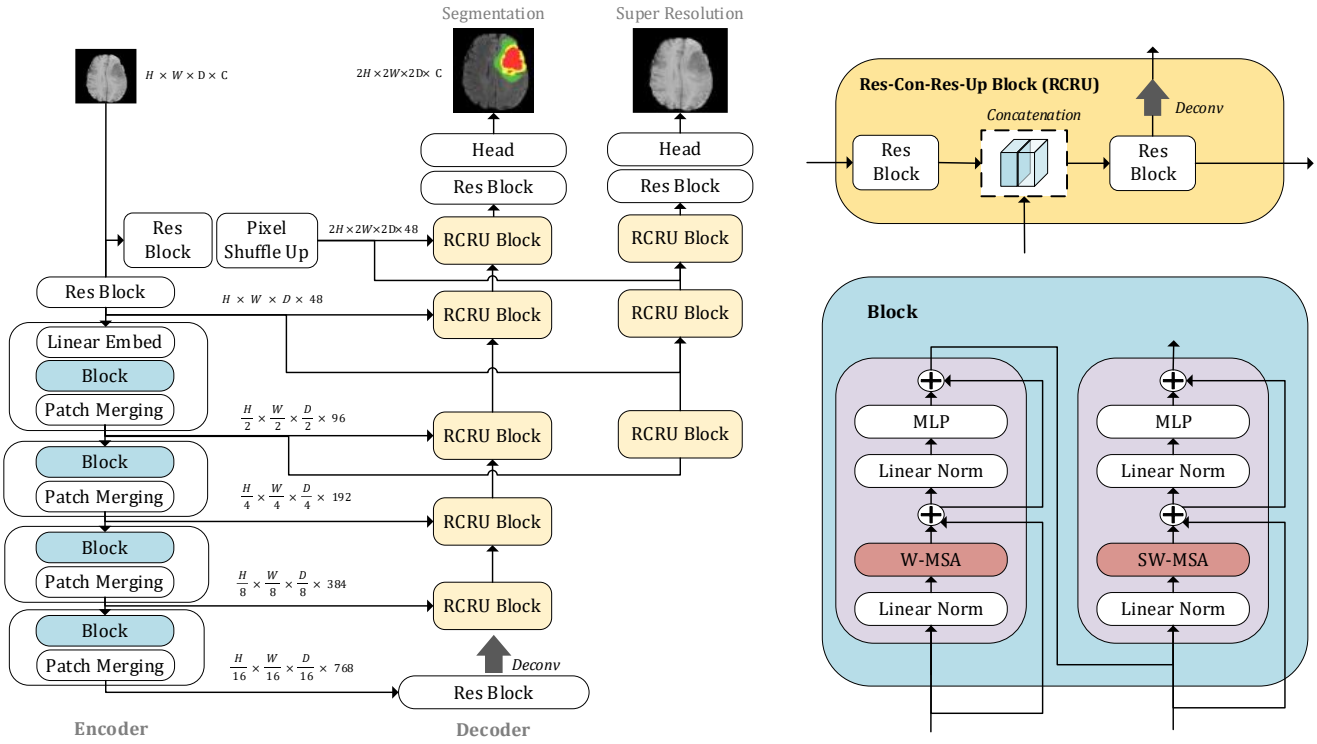

Figure 1. Overview of the proposed SwinSR-UNETR architecture for joint brain MRI segmentation and super-resolution. A shared Swin UNETR encoder extracts hierarchical features using window-based and shifted window self-attention, while the decoder employs partially shared RCRU blocks before splitting into task-specific branches.

More recently, transformer-based models have gained traction. Swin UNETR combines the Swin Transformer [2] encoder with a CNN-based decoder in a U-Net-like structure. It captures long-range dependencies and hierarchical features efficiently, achieving state-of-the-art results on 3D medical imaging benchmarks like BRATS. However, Swin UNETR [3] and similar models are primarily optimized for segmentation only, without consideration for image enhancement or resolution improvement.

Image super resolution (SR) aims to reconstruct high-resolution (HR) images from low-resolution (LR) inputs. Early approaches relied on interpolation and sparse coding, while recent deep learning-based models such as SRCNN, VDSR, and SRGAN have demonstrated superior performance in both natural and medical images. For 3D MRI, several volumetric SR models have been proposed to enhance anatomical clarity.

Few studies have attempted to combine super resolution and segmentation within a unified learning framework. Zhao et al. (2019) introduced a cascaded model where a super resolution network enhances the input before passing it to a segmentation network. Similarly, Jiang et al. (2020) proposed a joint learning model for 2D medical images, showing that shared features can benefit both SR and segmentation. However, neither approach targets brain tumour segmentation on 3D MRI, and both rely on CNN-based architectures without transformer-based encoding.

While segmentation and super resolution have been explored independently in brain MRI, joint learning of these tasks—particularly using a transformer-based architecture on 3D data—remains largely unexplored. Existing methods either decouple the tasks or are limited to 2D data and simpler architectures. To the best of our knowledge, SwinSR-UNETR is the first model to integrate super resolution and brain tumour segmentation into an end-to-end transformer-based framework, leveraging shared volumetric features for improved performance on both tasks.

3. Method

We proposed SwinSR-UNETR, a multi-task transformer-based framework designed to jointly perform brain tumour segmentation and super-resolution on low-resolution 3D MRI volumes. The model leveraged a shared encoder to extract contextual features and used two task-specific decoders for segmentation and image reconstruction. This joint training enabled the model to enhance spatial details while improving semantic understanding, benefiting both tasks.

3.1. Shared Encoder

We adopted a modified version of the Swin UNETR encoder as the shared backbone. The original Swin UNETR used a patch embedding layer that partitioned the input volume into non-overlapping 3D patches before feeding them into the Swin Transformer blocks. However, in our case, the input MRI volumes were already low-resolution. Applying

patch embedding would have further reduced the spatial resolution drastically, leading to significant loss of information, particularly around small tumour boundaries.

To address this, we removed the patch embedding layer and instead fed the input volume directly into the Swin Transformer blocks after an initial shallow residual convolution layer. This allowed the encoder to maintain a higher effective resolution and better preserve fine-grained details crucial for both segmentation and reconstruction tasks. The encoder outputs were passed to both decoders via skip connections at multiple feature levels.

3.2. Partial Split Shared Decoder

Instead of using two separate decoders, SwinSR-UNETR employed a partially merged decoder to handle both segmentation and super-resolution tasks more efficiently. To reduce computational overhead while maintaining task-specific performance, the decoder began with two shared Residual Convolutional Refinement Units (RCRU). These blocks refined the upsampled features from the transformer encoder and enhanced contextual representations before task separation.

After the shared RCRU layers, the decoder split into two branches, each dedicated to a specific task. The segmentation branch continued with upsampling layers and produced a three-channel output corresponding to the tumour labels. In contrast, the super-resolution branch reconstructed the high-resolution image and generated a four-channel output, which matched the input format. Aside from the shared RCRUs and distinct output heads, the remainder of the decoder architecture followed a standard UNETR-style structure, utilizing transposed convolutions, skip connections, and normalization layers to progressively recover spatial resolution and semantic detail.

3.2. Loss Function

We used Dice loss for the segmentation task and L1 loss for the super-resolution task. To effectively balance the optimization of both objectives, we applied an uncertainty-based loss weighting strategy as introduced by Kendall et al. [4]. In this approach, task-dependent uncertainties are learned during training to adaptively scale each loss term. The final loss function was formulated as:

$$L_{\text{total}} = \frac{1}{2\sigma_1^2} L_{\text{seg}} + \frac{1}{2\sigma_2^2} L_{\text{sr}} + \log(\sigma_1) + \log(\sigma_2)$$

where L_{seg} and L_{sr} denote the segmentation and super-resolution losses, respectively, and σ_1, σ_2 are learnable uncertainty parameters for each task. This formulation enabled the model to learn optimal weighting between tasks during training, resulting in a more stable and effective joint optimization.

4. Experiments & Results

4.1 Experimet Detail & Dataset

We conducted our experiments using the BraTS-GLI 2023[5] dataset, which contains multimodal 3D MRI volumes (T1, T1ce, T2, and FLAIR) along with expert-annotated tumor segmentation masks. This dataset is widely adopted for brain tumor segmentation and is also suitable for evaluating super-resolution tasks due to its high-resolution volumetric structure. Our experimental setup closely follows the Swin UNETR pipeline, including preprocessing, patch-based training, and sliding window inference. The model was trained on a dual-GPU setup with NVIDIA Quadro RTX 8000 GPUs using a patch size of 128×128×128, a batch size of 2, and an initial learning rate of 0.0001, scheduled with cosine annealing. Optimization was performed using the AdamW optimizer, and training was run for 200 epochs. This configuration enables effective utilization of GPU resources and demonstrates the capability of our multi-task model to jointly improve segmentation and super-resolution performance.

4.2. Quantitative Study

Table 1 presents a quantitative comparison among different model variants for both segmentation and super-resolution tasks. The baseline segmentation models, UNETR and Swin UNETR, were trained and evaluated on full-resolution inputs of size 96^3. Swin UNETR outperforms UNETR, achieving a Dice score of 88.00%, compared to 86.62%. For the super-resolution task, the Trilinear interpolation baseline achieves a PSNR of 30.7 and SSIM of 0.925, indicating its reasonable ability to reconstruct structural information from low-resolution inputs. Our proposed multi-task model, Swin-SR, which takes low-resolution inputs, achieves a competitive Dice score of 87.02%, while simultaneously maintaining the same PSNR and SSIM as the Trilinear baseline. This demonstrates the effectiveness of our joint optimization framework in preserving both segmentation accuracy and super-resolution fidelity. Notably, our model bridges the performance gap between single-task segmentation and SR models, while using significantly lower-resolution inputs for more efficient computation. Additionally, an ablation study reveals that removing the uncertainty-based loss weighting results in a noticeable decrease in super-resolution quality, with lower SSIM and PSNR scores, highlighting the importance of adaptive loss balancing in our multi-task framework. Compared to trilinear interpolation, the proposed model achieves a higher PSNR, suggesting improved image sharpness. Nevertheless, it exhibits some difficulty in accurately reconstructing structural details.

4.3. Qualitative Study

In the qualitative comparison between the trilinear

Table 1. Quantitative comparison of segmentation and super-resolution performance across different model variants. UNETR and Swin UNETR are single-task segmentation baselines. Trilinear interpolation serves as a non-learned SR baseline. Our proposed SwinSR-UNET.

Model Variant	Input Size	Output Size	Dice ↑	PSNR ↑	SSIM ↑
UNETR (Seg only) [6]	96^3	96^3	86.62%	-	-
Swin UNETR (Seg only) [3]	96^3	96^3	**88.00%**	-	-
Trilinear(SR only)	48^3	96^3	-	33.36	**0.960**
Ours SWIN-SR (Multi-task, w/o uncertainty loss)	48^3	96^3	87.05%	32.98	0.924
Ours SWIN-SR (Multi-task)	48^3	96^3	87.02%	**33.64**	0.935

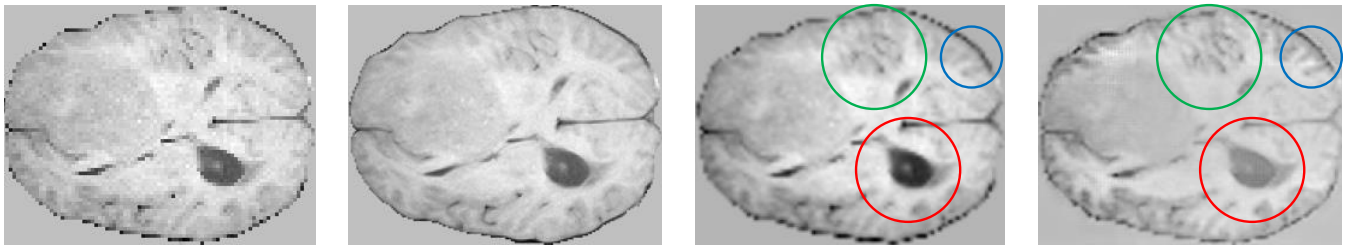

Figure 2. Qualitative Study Visualization, Comparisson between Trilinier and SwinSR-Unetr (Ours)

interpolation baseline and our proposed SwinSR-UNETR model, several notable differences were observed. Based on our study, we found that while our model occasionally struggles to maintain perfect color accuracy in certain regions (highlighted with red circles), it consistently preserves finer structural details (highlighted with green circles) compared to trilinear interpolation. Additionally, the SwinSR-UNETR output exhibits significantly smoother and more natural edges (highlighted with blue circles), whereas the trilinear method often produces blurrier or less defined boundaries. These observations suggest that our model offers better spatial fidelity and visual sharpness, even though minor intensity deviations can occur.

5. Conclusion

We propose SwinSR-UNETR, a multi-task transformer model for joint brain tumour segmentation and MRI super-resolution. Leveraging a shared Swin Transformer encoder and partially split decoders, the model efficiently learns both spatial and semantic features. An uncertainty-weighted loss further stabilizes multi-task optimization. On the BraTS-GLI 2023 dataset, SwinSR-UNETR achieves competitive Dice, PSNR, and SSIM scores compared to single-task baselines, while preserving fine structures and smooth edges. These results highlight the potential of unified segmentation and super-resolution frameworks for brain tumour imaging.

Acknowledgement

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) under the Artificial Intelligence Convergence Innovation Human Resources Development (IITP-2023-RS-2023-00256629) grant funded by the Korea government(MSIT)

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT). (RS-2023-00208397)

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2025-RS-2024-00437718) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation)

References

[1] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," 2015, pp. 234–241. doi: 10.1007/978-3-319-24574-4_28.

[2] Z. Liu *et al.*, "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," Mar. 2021.

[3] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. R. Roth, and D. Xu, "Swin UNETR: Swin Transformers for Semantic Segmentation of Brain Tumors in MRI Images," 2022, pp. 272–284. doi: 10.1007/978-3-031-08999-2_22.

[4] A. Kendall, Y. Gal, and R. Cipolla, "Multi-Task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics," May 2017.

[5] H. B. Li *et al.*, "The Brain Tumor Segmentation (BraTS) Challenge 2023: Brain MR Image Synthesis for Tumor Segmentation (BraSyn)," May 2023.

[6] A. Hatamizadeh *et al.*, "UNETR: Transformers for 3D Medical Image Segmentation," Mar. 2021.