Pre-Trained Generative Adversarial Network for Limited-Labeled Brain MRI Segmentation

Mehdi Karami[a,✉] karami.mehdi.scholar@gmail.com

Betsabeh Tanoori[a] betsatanoori@gmail.com

[a]Department of Computer Engineering, Zand Institute of Higher Education, Shiraz, Iran

Abstract

Accurate brain MRI segmentation is challenging due to subtle tissue boundaries, inter-subject variability, and the limited availability of annotated data. To address these challenges, we propose a multi-class brain MRI segmentation framework that integrates transfer learning with generative adversarial networks (GANs) to improve robustness and accuracy in low-data settings. The proposed model employs a GAN architecture with a ResNet-50 generator pre-trained on ImageNet, enabling effective feature transfer while stabilizing adversarial training. Brain MRI volumes are processed in a slice-wise 2D manner for computational efficiency, and grayscale slices are mapped to three-channel representations to ensure compatibility with pre-trained backbones. Adversarial learning further enforces local anatomical plausibility in the predicted segmentation maps. Experimental results using six-fold cross-validation demonstrate that the proposed approach consistently outperforms state-of-the-art segmentation models, achieving an accuracy of 98.41%, a Dice Average (excluding background) of 93.43%, and a mean IoU of 90.14%. These results highlight the effectiveness of combining transfer learning and adversarial regularization for reliable brain MRI segmentation under limited data conditions.

Keywords: Brain MRI segmentation, Transfer learning, Generative adversarial network, Deep learning, Low-data regime

1. Introduction

Accurate segmentation of brain tissues from magnetic resonance imaging (MRI) is a critical step in numerous neuroimaging applications, including disease diagnosis, treatment planning, and longitudinal monitoring. In particular, precise delineation of gray and white matter regions is essential due to their subtle intensity differences and complex anatomical boundaries. Despite significant progress, automated segmentation remains challenging, especially in scenarios with limited annotated data or constrained computational resources.

Recent advances in medical image segmentation have been largely driven by deep learning, particularly convolutional neural network (CNN)–based architectures. Contemporary approaches for brain MRI segmentation include CNN–based architectures such as U-Net [1] and its variants (e.g., nnU-Net [2]), as well as transformer-based models including Swin-UNet [3] and TransUNet [4]. While these methods achieve high performance on well-annotated datasets, they often degrade in low-data regimes and rely solely on supervised learning without exploiting additional latent structures in the data.

Generative adversarial network (GAN)-based approaches have shown promise in capturing higher-order spatial consistency and generating richer feature representations through adversarial learning. The adversarial component encourages the network to capture additional latent structures in the data, such as anatomical consistency and realistic spatial relationships, thereby improving segmentation plausibility beyond pixel-wise supervision.

In addition, transfer learning provides strong initialization from large-scale natural image datasets, improving generalization and helping to mitigate overfitting when annotations are scarce.

Despite these advances, existing approaches have not yet established a unified framework that effectively integrates transfer learning with adversarial learning for robust brain MRI segmentation in low-data regimes. This gap motivates the development of a framework that simultaneously leverages prior knowledge and enforces anatomically plausible representations through adversarial regularization.

We propose a novel slice-wise 2D brain MRI segmentation framework that integrates a pre-trained generator within a GAN architecture. The pre-trained generator leverages vi-

sual priors learned from a large-scale natural image dataset, while the adversarial discriminator refines feature representations to produce anatomically coherent segmentation masks. This combination enables robust performance in low-data settings, while maintaining computational efficiency through 2D slice-wise image processing, making the framework suitable for resource-constrained environments.

The key contributions of this study are summarized as follows:

1. We propose TL-GAN, a slice-wise 2D GAN-based framework for multi-class brain MRI segmentation, in which the generator is implemented as a U-Net–style architecture with a ResNet-50 encoder pre-trained on ImageNet, enabling effective transfer learning under limited annotated data.
2. By integrating transfer learning with adversarial regularization, the proposed method achieves stable and robust segmentation performance in low-data settings without requiring large-scale medical annotations.
3. Extensive experimental evaluations, including robustness analysis across different RGB representations of grayscale MRI slices and comparisons with state-of-the-art models, demonstrate the effectiveness and generalization capability of the proposed framework.

The remainder of this paper is organized as follows. Section 2 reviews related works. Section 3 presents the proposed methodology, followed by experimental results in Section 4. Section 5 provides an in-depth discussion of the findings, while Section 6 summarizes key insights, limitations, and potential future directions. Finally, Section 7 concludes the study.

2. Related works

Brain MRI segmentation has been extensively studied using CNNs, transformer-based models, GANs, and transfer learning techniques. U-Nets and their variants remain dominant due to effective local feature extraction capabilities [1, 2], while transformer-based models capture long-range contextual information [3, 4]. GAN-based approaches have been employed to improve spatial consistency and mask realism through adversarial learning [5, 6], and transfer learning has proven effective in accelerating convergence and improving generalization in low-data regimes [7, 8]. Hybrid approaches combining these paradigms leverage complementary strengths [9]. Since the proposed framework integrates transfer learning with a GAN-based architecture using a pre-trained ResNet-50 encoder, the following sections primarily focus on state-of-the-art approaches related to CNN-based segmentation, adversarial learning, and transfer learning in brain MRI analysis.

2.1. U-Net–based segmentation

U-Net [1] and its variants remain the backbone of brain MRI segmentation due to their encoder–decoder structure with skip connections, which enables effective integration of low-level spatial details with high-level semantic features. Numerous architectures, including DeepLabV3 [10], DeepLabV3+ [11], and SegResNet [12], extend this paradigm by enhancing feature extraction through deeper convolutional blocks, residual connections, or multi-scale context aggregation [1, 13]. These models demonstrate strong performance when sufficient annotated data are available, as convolutional hierarchies can robustly learn task-specific MRI representations.

Despite their success, U-Net–based architectures exhibit notable limitations, particularly in low-data regimes. The large number of trainable parameters increases susceptibility to overfitting, resulting in reduced generalization to unseen scans or acquisition settings [14]. Furthermore, repeated downsampling operations can lead to loss of fine-grained spatial information, while the encoder–decoder design may inadequately preserve global contextual cues, yielding fragmented or anatomically inconsistent segmentations [15].

To address these challenges, several studies have incorporated ResNet-based encoders into U-Net–style architectures [16–18], leveraging residual learning to facilitate deeper networks and improved gradient propagation. Such designs also enable the reuse of pre-trained weights from large-scale natural image datasets, forming a foundation for transfer learning in medical image segmentation. Nevertheless, when used in isolation, these architectures remain constrained by limited supervision and lack explicit mechanisms to enforce anatomical plausibility.

2.2. GAN-based segmentation

Generative Adversarial Networks (GANs) are a class of generative models that synthesize data by learning the probability distribution of real data [19]. GANs have also been leveraged in medical image segmentation, where adversarial learning frameworks refine segmentation boundaries or generate auxiliary training masks to improve segmentation accuracy [6]. GANs are composed of a generator and a discriminator. In adversarial segmentation, a discriminator evaluates the masks predicted by the generator (segmenter), thereby encouraging the segmenter to produce outputs that are indistinguishable from the ground-truth labels. This mechanism promotes spatial consistency and preserves the global structure of the segmented regions through the resulting adversarial loss. Integrating an adversarial loss into a U-Net generator has been shown to balance local and global information, yielding more accurate and coherent segmentation of brain tumors [5].

Despite their benefits, GAN-based models face persistent drawbacks. Training GANs is often unstable and highly sen-

sitive to hyperparameters, requiring careful balancing of generator and discriminator updates to avoid non-convergence, vanishing gradients, or mode collapse [20]. GANs also typically require a large number of labeled images. In medical imaging, however, this challenge is further exacerbated by the scarcity of annotated data. This scarcity forces most GAN-based models to be trained from scratch on domain-specific datasets that are often local or publicly inaccessible, resulting in limited reusability and reachability [21].

Overall, adversarial components enhance spatial realism and structural consistency in segmentation maps, but they require extensive data augmentation or large annotated cohorts and impose higher computational and training demands [20, 21].

2.3. Transfer learning approaches

Transfer learning (TL) from natural-image datasets, such as ImageNet, is commonly used to bootstrap segmentation models when annotated data are scarce [22], particularly in medical imaging. Pre-trained encoders, when integrated into architectures like U-Net, provide rich feature representations, accelerate convergence, and improve segmentation performance [7]. More recently, foundation models pre-trained directly on medical scans (e.g., RadiologyNET) have been proposed, often matching or slightly surpassing pre-trained ImageNet models in highly data-constrained settings [8].

Despite these benefits, applying transfer learning to brain MRI segmentation faces several challenges. The principal challenge is the domain gap: natural-image pre-training captures textures, edges, and color distributions that differ substantially from MRI intensities, which are shaped by modality, pulse sequences, and acquisition hardware. Consequently, while TL reduces the need for large annotated datasets for downstream tasks, it may still require fine-tuning to adapt to MRI-specific features [23–26]. For instance, studies have shown that pre-training on in-domain brain tumor MRI data can outperform natural-image pre-training, achieving faster convergence and improved performance even with fewer samples [27].

In summary, transfer learning is a potent strategy for bootstrapping models in low-data regimes; however, its effectiveness in brain MRI segmentation depends on careful adaptation to modality-specific characteristics, an insight that motivates the use of pre-trained ResNet encoders in combination with adversarial training in the proposed framework.

2.4. Hybrid approaches

Recent research has combined multiple paradigms to leverage their complementary strengths. For example, some methods fuse transformers and CNNs to capture both global context and local details for brain tumor segmentation [28]. Other works employ GAN transformer-based models for combined super-resolution and modality translation of MRI [29], or integrate adversarial training into U-Net architectures. One line of work uses a transformer-enhanced U-Net with an adversarial loss for brain tumor segmentation, balancing self-attention with GAN consistency to improve mask quality [5]. Another study combined transfer learning with GANs for brain MRI segmentation, demonstrating that pre-trained backbones can effectively be embedded into adversarial frameworks [9]; however, this work focused primarily on the combination itself and did not address challenges arising from limited annotated data.

Overall, these strategies emphasize technical combinations of multiple paradigms and applications in various medical imaging tasks, illustrating how transfer learning via pre-trained backbones can be integrated into GAN frameworks to stabilize training and enhance realism.

To our knowledge, no prior work has jointly applied transfer learning and adversarial training for 2D slice-wise brain MRI segmentation under limited-data conditions. A key novelty of our proposed method is the integration of a pre-trained ImageNet encoder backbone with a GAN discriminator to segment each axial slice in data-scarce settings. This addresses the gap in existing literature by uniting the sample efficiency of transfer learning with the spatial consistency incentives of adversarial learning, achieving a stable and efficient framework.

3. Methodology

Accurate multi-class segmentation of brain MRI remains challenging due to subtle intensity differences between tissues, inter-subject anatomical variability, and the limited availability of annotated medical data. These challenges are further exacerbated in practical scenarios where computational resources are constrained, limiting the applicability of large-scale 3D models. To address these issues, we propose an efficient and robust segmentation framework, termed **TL-GAN**, which integrates transfer learning with adversarial regularization within a slice-wise 2D formulation.

The proposed TL-GAN framework is designed to achieve high segmentation accuracy under limited-data conditions while maintaining computational efficiency. By operating on 2D slices extracted from volumetric MRI scans, the model significantly reduces memory and hardware requirements, while adversarial learning compensates for reduced contextual information by enforcing local anatomical plausibility in the predicted segmentation maps. The workflow comprises dataset preparation (§3.1), preprocessing (§3.2), and network architecture design (§3.3). Preprocessing ensures spatial and intensity consistency across subjects, while the core GAN-based architecture combines a U-Net–style generator with a pre-trained ResNet-50 encoder and a patch-based discriminator. Model training is guided by a joint supervised and adversarial loss formulation, enabling accurate and anatom-

ically consistent multi-class segmentation. The complete workflow of the proposed TL-GAN framework is illustrated in Figure 1.

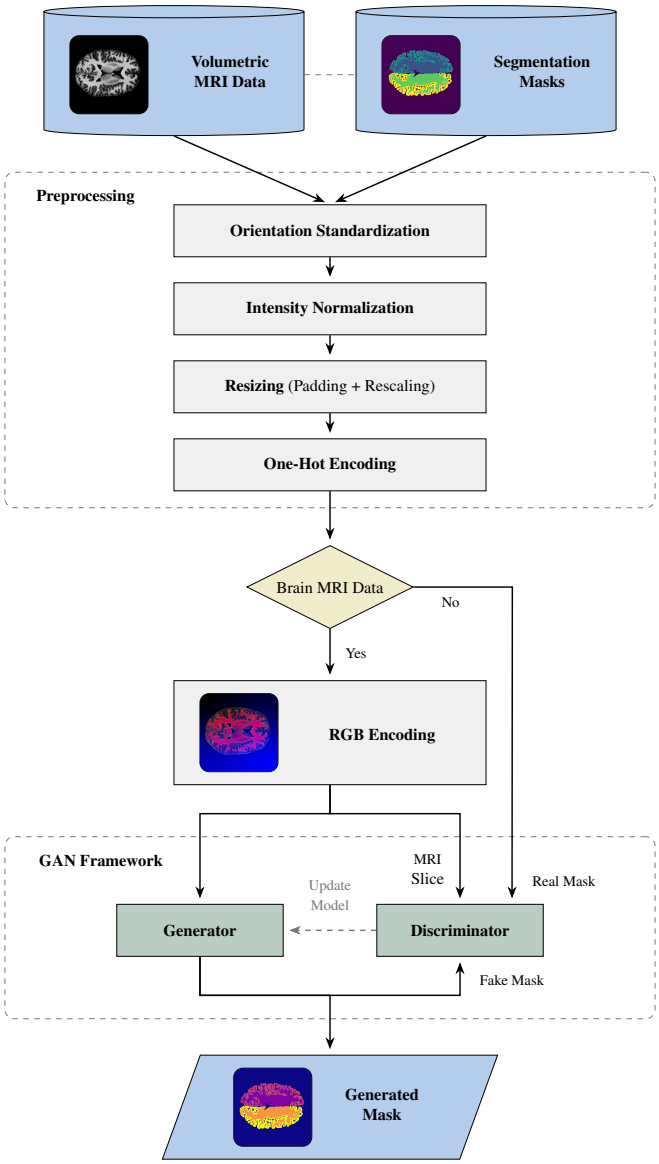

Figure 1. Flowchart of the proposed TL-GAN framework, highlighting the slice-wise input, preprocessing pipeline, generator–discriminator interaction, and multi-class segmentation output.

3.1. Dataset

In this study, we utilize the UltraCortex 9.4 T brain MRI dataset [30], a publicly available collection of high-resolution T1-weighted scans of healthy adult brains. The dataset includes expert-annotated manual segmentations of **gray and white matter** for a subset of **12 subjects**, with each volume meticulously delineated into five anatomical classes: background (non-brain), left cerebral cortex (LC_C), left cerebral white matter (LC_WM), right cerebral cortex (RC_C), and right cerebral white matter (RC_WM). The skull-stripped MR volumes provided by the dataset were utilized in place of the raw T1-weighted scans. The manual segmentations have been validated by two expert neuroradiologists, ensuring high reliability [30].

Given the dataset's small number of annotated brain MRIs, it is well-suited for investigating segmentation methods in low-data regimes. This dataset contains **12 volumetric MR images**, with each image comprising **259 $\pm$ 25 axial slices** (ranging from 192 to 288 slices per subject). In total, the dataset includes **3,104 2D slices** used for segmentation. We adopt a slice-by-slice 2D approach, treating each slice as an independent training instance. This strategy is computationally efficient and aligns with practical clinical imaging workflows, where 2D segmentation is widely used due to lower computational and memory requirements [31].

The statistical characteristics of the utilized dataset after preprocessing are summarized in Table 1. Detailed preprocessing steps are described in the following section (§3.2).

3.2. Preprocessing

In this study, we applied a multi-stage preprocessing pipeline to ensure spatial and intensity consistency across our volumetric MRI dataset, while preserving anatomical integrity. The preprocessing steps were designed to mitigate inter-subject variability, standardize intensity values, and prepare the dataset for deep learning models. These preprocessing steps are as follows:

1. Orientation standardization
2. Intensity normalization
3. Resizing
4. One-hot encoding of segmentation masks

Each of these steps is described in detail in the following subsections, which explain the rationale and specific techniques used in the preprocessing pipeline. The full workflow is schematically summarized in Figure 2.

3.2.1. Orientation standardization

The volumetric MRI data utilized in this study initially exhibited substantial spatial orientation heterogeneity, with scans acquired in various anatomical configurations (like ('L', 'A', 'S'), ('R', 'A', 'S'), ('P', 'S', 'R'), and ('R', 'P', 'I')). To reduce orientation-induced variability and ensure consistent anatomical representation across all subjects, we standardized all scans to the Right–Anterior–Superior (RAS) orientation. This step was crucial for minimizing inter-subject spatial variability, particularly given the limited subject size in our dataset. Standardizing the orientation facilitates downstream image processing and ensures that anatomical structures are consistently aligned across subjects.

Table 1. Statistical summary of the brain MRI dataset after preprocessing.

Statistic	Value
Number of subjects	12
Total number of 2D slices	3,104
Average number of slices per subject	259±25
Slice dimensions	352×352
Number of annotated classes	5 (BG, LC_C, LC_WM, RC_C, RC_WM)
Class distribution	[85.8% BG, 3.60% LC_C, 3.50% LC_WM, 3.60% RC_C, 3.51% RC_WM]

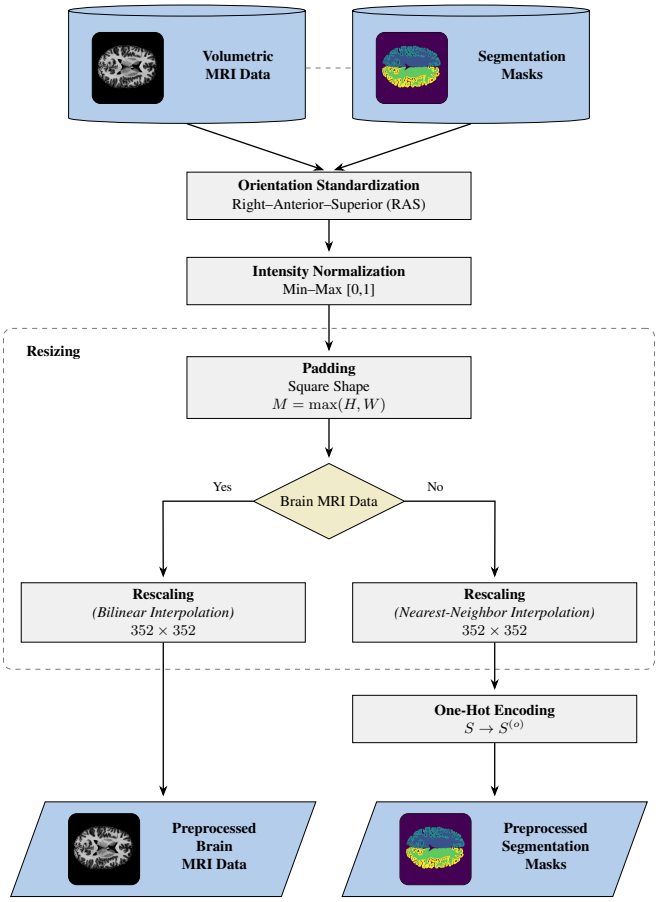

Figure 2. Flowchart of the preprocessing pipeline.

3.2.2. Intensity normalization

To ensure consistent intensity values across the dataset and reduce scale differences, we applied *min–max* normalization to each 2D slice. This maps all intensities to the range $[0, 1]$, making the data less sensitive to acquisition variability and promoting gradient stability, which facilitates more reliable and faster convergence during training.

3.2.3. Resizing

In order to ensure compatibility with the fixed input size expected by our convolutional neural network (CNN) architecture, we standardized the dimensions of each 2D MRI slice. Initially, each slice was transformed into a square format through **conditional zero-padding**, followed by rescaling to a consistent resolution of $\mathbf{352 \times 352}$ **pixels**. This resizing step ensures that all slices have the same input size, which is crucial for training deep neural networks.

To handle the continuous-valued intensity profiles in the skull-stripped MRI slices, we applied *bilinear interpolation* during rescaling. This method estimates pixel values based on the weighted average of neighboring pixels, preserving local structural gradients and mitigating aliasing artifacts, making it well-suited for anatomical imagery.

For the categorical segmentation masks, we employed *nearest-neighbor interpolation*, which assigns each rescaled pixel the label of the nearest corresponding pixel from the original mask. This method preserves anatomical boundaries and prevents interpolation-induced artifacts, ensuring that the labels align correctly with the resized images.

Together, these preprocessing steps standardize the dimensions and intensity resolution of the MRI slices and segmentation masks, enabling stable training for the deep learning model.

3.2.4. One-hot encoding of segmentation masks

As the final preprocessing step, each categorical segmentation mask was converted into a one-hot encoded representation. This transformation generates a separate binary channel for each anatomical class, allowing the segmentation labels to be represented in a format compatible with the network's multi-class output. One-hot encoding aligns the ground-truth masks with the *softmax* activation used in the output layer of the segmentation network and enables **class-wise loss computation** during training. This representation is particularly important for multi-class brain MRI segmentation, as it allows the model to learn distinct decision boundaries for each anatomical structure.

3.3. Model architecture

The segmentation method proposed in this study is based on a Generative Adversarial Network (GAN) architecture. Unlike general GAN-based segmentation models, our framework incorporates a **ResNet-50** encoder in the **generator network** $\boldsymbol{G}$, which is pre-trained on ImageNet to capture rich feature representations. This enables the model to handle complex anatomical structures more effectively, as detailed in the following sections. The **discriminator network** $\boldsymbol{D}$,

in contrast, is CNN-based and evaluates the anatomical plausibility of the generated segmentations by comparing them against expert-annotated MRI slices. As shown in Fig. 3, the generator takes preprocessed MRI slices as input and outputs dense multi-class segmentation maps, while the discriminator assesses these predictions alongside the corresponding input slices. Specific architectural components, training objectives, and input representations are described in detail in the ensuing subsections.

3.3.1. Generator network

In the proposed GAN-based brain MRI segmentation framework, the generator G departs from the conventional GAN generator design, which typically relies on shallow convolutional architectures aimed at image synthesis. Instead, we formulate the generator as a **segmentation-oriented network** based on a **U-Net–like architecture with a ResNet-50 encoder**, as illustrated in Fig. 4.

Specifically, the encoder follows the ResNet-50 architecture, which consists of five stages of residual convolutional blocks and has been pre-trained on the ImageNet dataset (1.2 million natural images across 1000 categories) [32]. Leveraging this pre-training enables effective **transfer learning**, allowing the model to exploit rich hierarchical feature representations learned from large-scale data and adapt them to the medical imaging domain, particularly under limited training samples. Moreover, the residual learning mechanism with identity skip connections facilitates stable optimization by alleviating vanishing gradient issues in deep networks.

The decoder is constructed as a symmetrical upsampling path that progressively restores spatial resolution using transpose convolutions (or upsampling followed by convolution). Feature maps from corresponding encoder stages are concatenated via U-Net–style skip connections, enabling the integration of high-level semantic information with fine-grained spatial details. This design effectively transforms the ResNet-50 backbone into a fully convolutional network suitable for dense pixel-wise prediction.

At the final decoding stage, a convolution followed by a softmax activation is applied to generate per-pixel class probabilities for the five anatomical classes. The resulting output constitutes the predicted multi-class segmentation map produced by the generator.

3.3.2. Grayscale to RGB representation

The ResNet-50 backbone employed in the generator network requires three-channel RGB input, whereas brain MRI slices are inherently single-channel grayscale images. To bridge this modality mismatch and enable effective transfer learning from ImageNet-pre-trained weights, we investigated different strategies for constructing three-channel representations from grayscale MRI slices.

Two commonly used encoding approaches were considered as baselines. The **grayscale replication encoder** simply duplicates each MRI slice across the three RGB channels [33]. While straightforward and widely used, this approach produces identical channels and does not exploit the full representational capacity of color-sensitive filters learned during ImageNet pre-training. The **2.5D adjacent-slice encoder** assigns the central slice to the green channel and its immediate neighboring slices to the red and blue channels, respectively [34]. This strategy introduces limited inter-slice context but depends on slice continuity and complicates inference at volume boundaries.

To address these limitations, we propose a novel **IEP-RGB encoder (Intensity + Edges + Position)**, which embeds complementary anatomical information into a three-channel representation using only a single MRI slice. As illustrated in Fig. 5, the three channels encode:

Channel 1 (Intensity): The normalized grayscale image, preserving raw anatomical information.

Channel 2 (Edge magnitude): A Sobel-based gradient magnitude map [35] that emphasizes tissue boundaries and fine structural details, facilitating more accurate delineation of cortical and subcortical regions.

Channel 3 (Positional map): A smooth nonlinear spatial encoding that injects explicit positional priors, reflecting the characteristic spatial organization of brain tissues, such as peripheral cortical regions and centrally located white matter. Given an image of height H and width W, we construct two linearly spaced coordinate grids $X(x, y)$ and $Y(x, y)$ scaled from 0 to 1 across width and height, respectively. These are combined nonlinearly via a cosine transform:

$$\begin{aligned} P(x, y) = 0.25 \cdot \big(1 - \cos(\pi X(x, y))\big) \\ + 0.75 \cdot \big(1 - \cos(\pi Y(x, y))\big) \end{aligned} \tag{1}$$

The weights (0.25 for X and 0.75 for Y) balance horizontal and vertical positional contributions. The map is subsequently normalized to $[0, 1]$.

By integrating intensity, boundary, and spatial information in a parameter-free manner, the IEP-RGB encoder provides a richer input representation while remaining compatible with pre-trained RGB backbones.

3.3.3. Discriminator network

In our GAN-based segmentation framework, the discriminator network D is implemented as a convolutional neural network (CNN) designed to assess the anatomical plausibility of predicted segmentation maps conditioned on the corresponding MRI slices.

As illustrated in Fig. 6, D receives as input a concatenation of a preprocessed three-channel MRI slice and its associated segmentation mask, represented in one-hot form across five channels. This results in an eight-channel input tensor, enabling the discriminator to jointly evaluate image appearance and label consistency.

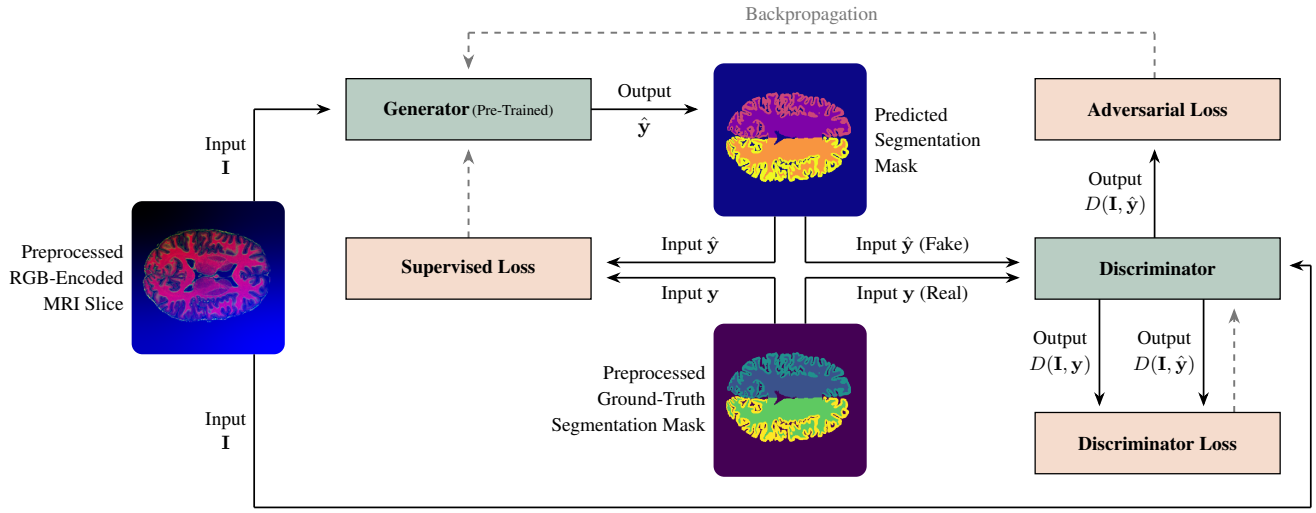

Figure 3. Proposed GAN framework with generator–discriminator interaction and loss components.

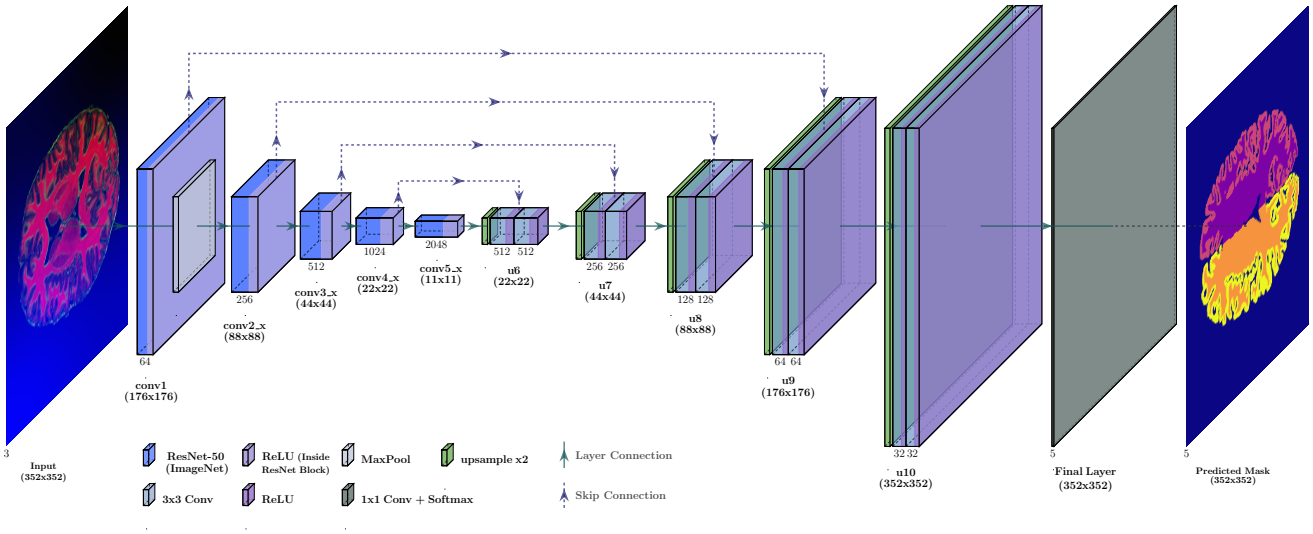

Figure 4. Generator network architecture.

The discriminator is formulated as a patch-based CNN, following the PatchGAN design commonly adopted in image-to-image translation tasks such as Pix2Pix [36]. Rather than producing a single global real/fake decision, D consists of three strided convolutional layers that progressively downsample the input and output a 44×44 grid of logits. Each output element corresponds to a local receptive field of approximately 46×46 pixels in the input, allowing the discriminator to focus on local structural realism, such as boundary smoothness, tissue continuity, and the absence of spurious holes or artifacts.

A sigmoid activation is applied to the logits to obtain patch-wise real/fake probabilities. During training, D is optimized to classify image–mask pairs derived from expert annotations as real and those generated by the generator G as fake. Simultaneously, G is trained to produce segmentation outputs that induce D to predict them as real. This adversarial formulation is particularly well-suited to medical image segmentation, where global anatomical structure is largely constrained by anatomy, while local inconsistencies are common failure modes that benefit from explicit adversarial supervision.

3.3.4. Loss functions

In the proposed GAN-based brain MRI segmentation framework, the generator and discriminator are optimized using complementary objective functions. The generator is guided by a combination of supervised segmentation losses and

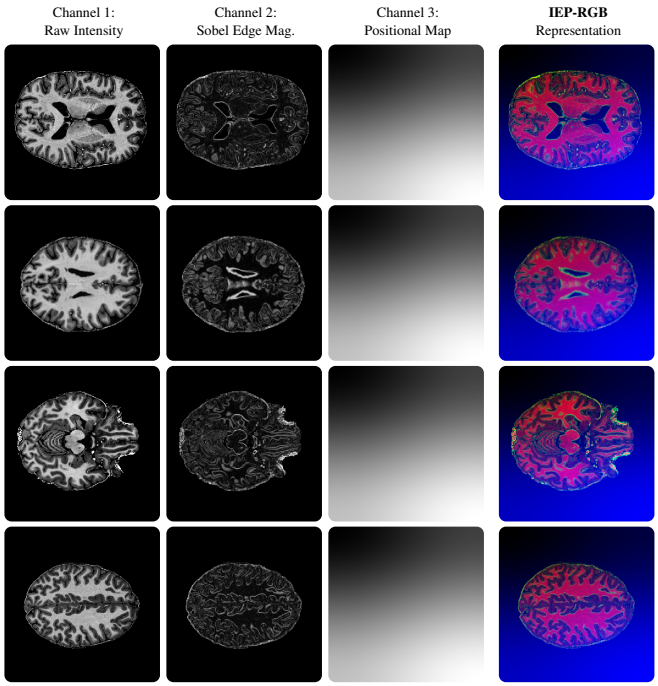

Figure 5. Visualization of the proposed IEP-RGB decomposition for representative brain MRI slices.

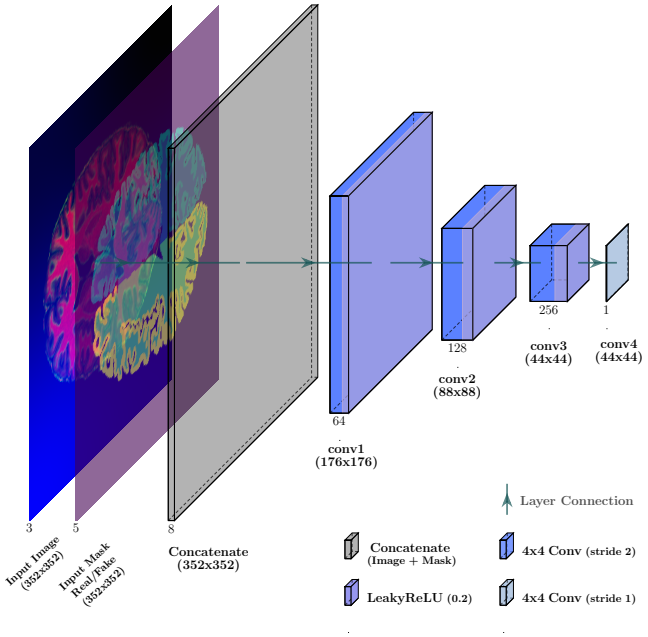

Figure 6. Discriminator network architecture.

adversarial feedback, while the discriminator is trained to distinguish ground-truth segmentation masks from generated predictions conditioned on the corresponding MRI slices. This joint optimization encourages both pixel-wise accuracy and anatomically plausible segmentation outputs, which is particularly important in low-data and class-imbalanced settings.

- *Class weighting:*
  Brain MRI segmentation suffers from severe class imbalance, as background and white matter pixels dominate over thin cortical regions. To address this issue, inverse-frequency class weights are computed from the training masks and median-normalized to stabilize optimization. Specifically, for each class $i \in \{1, \ldots, K\}$, the weight is defined as:

$$w_i \;=\; \frac{N}{K\,(f_i + \varepsilon)} \tag{2}$$

  where K is the number of classes, N is the total number of labeled pixels, f_i is the empirical class frequency of class i, and ε is a small constant for numerical stability. The weights are then normalized by their median value:

$$\tilde{w}_i \;=\; \frac{w_i}{\mathrm{median}(w_1, \ldots, w_K)} \tag{3}$$

  The median-normalized weights are used in all weighted loss terms described below.
- *Generator loss:*
  The generator G is optimized using a composite objective that combines a supervised segmentation loss with an adversarial loss:

$$\mathcal{L}_G \;=\; \mathcal{L}_{\mathrm{sup}} \;+\; \lambda_{\mathrm{adv}}\,\mathcal{L}_{\mathrm{adv}} \tag{4}$$

  where $\lambda_{\mathrm{adv}}{>}0$ controls the contribution of adversarial regularization.

  1. *Supervised loss:*
     The supervised loss is defined as the sum of a weighted categorical cross-entropy loss and a weighted Dice loss:

$$\mathcal{L}_{\mathrm{sup}} \;=\; \mathcal{L}_{\mathrm{WCCE}} \;+\; \mathcal{L}_{\mathrm{WDice}} \tag{5}$$

     • *Weighted categorical cross-entropy (WCCE):*
     Let $Y_{x,y,i} \in \{0,1\}$ denote the one-hot encoded ground-truth label and $\hat{Y}_{x,y,i} \in [0,1]$ the predicted probability for class i at pixel (x,y). The weighted categorical cross-entropy loss is defined as:

$$\mathcal{L}_{\mathrm{WCCE}} \;=\; -\frac{1}{HW}\sum_{x=1}^{H}\sum_{y=1}^{W}\sum_{i=1}^{K}\tilde{w}_i\;Y_{x,y,i}\;\log\left(\hat{Y}_{x,y,i}\right) \tag{6}$$

     where H and W denote the spatial height and width of the input image.
     • *Weighted Dice loss (WDice):*
     To explicitly promote spatial overlap between predicted and ground-truth regions, we employ a weighted Dice loss. The per-class soft Dice score is computed as:

$$\text{Dice}_i = \frac{2\sum_{x,y} Y_{x,y,i}\,\hat{Y}_{x,y,i} + \varepsilon}{\sum_{x,y} Y_{x,y,i} + \sum_{x,y} \hat{Y}_{x,y,i} + \varepsilon} \quad (7)$$

where ε is a small constant to avoid division by zero. The weighted Dice loss is then computed as:

$$\mathcal{L}_{\text{WDice}} = \frac{1}{K}\sum_{i=1}^{K} \tilde{w}_i \left(1 - \text{Dice}_i\right) \quad (8)$$

This formulation emphasizes overlap quality and mitigates class imbalance effects.

2. *Adversarial loss:*
   In addition to supervised learning, the generator receives adversarial feedback from the discriminator to encourage anatomically realistic segmentation masks. Let $D(I, \hat{y})$ denote the discriminator logits for an input MRI slice I and paired with a generated segmentation mask $\hat{y}$. The generator's adversarial loss is defined as a patch-wise *binary cross-entropy* against the "real" label:

$$\mathcal{L}_{\text{adv}} = \frac{1}{HW}\sum_{x=1}^{H}\sum_{y=1}^{W} \text{BCE}\big(1, \sigma\big(D(I, \hat{y})_{x,y}\big)\big) \quad (9)$$

   where σ is the logistic sigmoid. This term encourages the generator to produce segmentation masks that are indistinguishable from expert annotations at the local patch level.

- *Discriminator loss:*
  The discriminator D is trained to classify ground-truth pairs (I, y) as real and generated pairs $(I, \hat{y})$ as fake using a patch-wise *binary cross-entropy* loss:

$$\begin{aligned}\mathcal{L}_D = &\frac{1}{HW}\sum_{x=1}^{H}\sum_{y=1}^{W} \text{BCE}\Big(\mathbf{1}, \sigma(D(I, y)_{x,y})\Big) \\ &+ \frac{1}{HW}\sum_{x=1}^{H}\sum_{y=1}^{W} \text{BCE}\Big(\mathbf{0}, \sigma(D(I, \hat{y})_{x,y})\Big)\end{aligned} \quad (10)$$

Minimizing $\mathcal{L}_D$ improves the discriminator's ability to detect unrealistic segmentation patterns, while minimizing $\mathcal{L}_G$ encourages the generator to produce label-consistent and anatomically plausible segmentations.

4. Experimental results

This section evaluates the proposed TL-GAN segmentation framework using six-fold cross-validation. We first describe the evaluation criteria used to quantify segmentation performance (§4.1), followed by details of the data splitting strategy and training–testing protocol (§4.2). The optimization setup, hyperparameter configuration, and implementation environment are then summarized to ensure reproducibility (§4.3).

Quantitative comparisons with state-of-the-art brain MRI segmentation models are subsequently presented to position the proposed method within the existing literature (§4.4). To better understand the contribution of individual design choices, we conduct ablation studies analyzing the effects of transfer learning and adversarial training (§4.5). The effectiveness of different RGB encoding strategies, including grayscale replication, 2.5D adjacent-slice encoding, and the proposed IEP-RGB encoder, is examined within the same experimental framework (§4.6). Finally, qualitative segmentation results are provided to visually assess anatomical plausibility and boundary consistency.

Unless otherwise stated, all reported results correspond to the test split of each fold and are expressed as the mean ± standard deviation across folds.

4.1. Evaluation metrics

Segmentation performance was primarily evaluated using the Dice coefficient, computed on a per-class basis. For each class c, let $y_{\text{true}}^{(c)}$ and $y_{\text{pred}}^{(c)}$ denote the binary ground-truth and predicted segmentation masks, respectively. The Dice coefficient was defined as:

$$\text{Dice}_c = \frac{2\,|y_{\text{true}}^{(c)} \cap y_{\text{pred}}^{(c)}|}{|y_{\text{true}}^{(c)}| + |y_{\text{pred}}^{(c)}|} \quad (11)$$

where $|\cdot|$ denotes the cardinality of the voxel set. For numerical stability, a small constant $\epsilon = 10^{-6}$ was added to both the numerator and denominator.

Dice scores were reported separately for each tissue class (i.e., background, left cortical white matter, left cortical cortex, right cortical white matter, right cortical cortex).

The *Dice Average* was used as the primary evaluation metric, obtained by averaging Dice scores across the four target tissue classes while excluding the background class:

$$\text{Dice}_{\text{average}} = \frac{1}{C}\sum_{c=1}^{C} \text{Dice}_c, \quad C = 4 \quad (12)$$

The background class was excluded from this average because its large volumetric dominance can artificially inflate performance estimates. This exclusion provides a more informative and clinically meaningful assessment of segmentation quality for anatomically relevant structures.

As a secondary metric, the *Mean Intersection over Union (Mean IoU)* was computed following the standard definition. For each class c:

$$\text{IoU}_c = \frac{|y_{\text{true}}^{(c)} \cap y_{\text{pred}}^{(c)}|}{|y_{\text{true}}^{(c)} \cup y_{\text{pred}}^{(c)}|} \quad (13)$$

and the mean was obtained as:

$$\text{Mean IoU} = \frac{1}{C+1} \sum_{c=0}^{C} \text{IoU}_c \tag{14}$$

where $c = 0$ corresponds to the background class. Unlike Dice Average, Mean IoU was computed over all classes, including background, to ensure consistency with the implementation of the `tf.keras.metrics.MeanIoU` metric.

Both metrics were aggregated across slices and subjects within each fold of the six-fold cross-validation, and results were reported as the mean $\pm$ standard deviation across folds.

4.2. Training strategy

This section describes the data partitioning scheme, cross-validation protocol, and training procedure used for the proposed GAN-based segmentation model. To ensure subject-level independence and prevent information leakage across subsets, the dataset, comprising twelve volumetric brain MRI scans, was evaluated using a six-fold cross-validation protocol. Subjects, rather than individual slices, were treated as the unit of partitioning, such that all slices from a given subject were assigned exclusively to a single subset within each fold.

At each fold, the twelve subjects were divided into three disjoint subsets: eight subjects for training, two for validation, and two for testing. This fixed 8:2:2 partitioning was repeated across six folds so that each subject appeared exactly once in the test set while serving as training or validation data in the remaining folds. This strategy ensured that performance evaluation reflected true generalization to unseen subjects rather than benefiting from slice-level correlations.

Within each subset, all slices from the selected subjects were concatenated to form the corresponding training, validation, and test datasets, along with their associated manual segmentation masks. No subject overlap occurred between subsets at any fold.

Training was conducted in an adversarial manner, with the generator and discriminator optimized jointly using alternating updates. In each iteration, one generator update was followed by one discriminator update.

The generator was trained using a composite objective consisting of a supervised segmentation loss (weighted categorical cross-entropy plus weighted Dice loss) and an adversarial loss term. The relative contribution of the adversarial component was controlled by a fixed weighting factor $\lambda_{\text{adv}} = 0.01$.

The discriminator was trained to distinguish between real image–mask pairs (expert annotations) and fake pairs produced by the generator, using a patch-wise binary cross-entropy loss. This alternating optimization scheme encouraged the generator to produce anatomically plausible and spatially consistent segmentation maps while maintaining pixel-wise accuracy.

4.3. Optimization and implementation details

Optimization of both the generator and discriminator networks was performed using the Adam optimizer [37] with an initial learning rate of 1×10^{-4} and default momentum parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$. To adaptively adjust the learning rate during training, a *ReduceLROnPlateau* scheduler was employed, monitoring the validation supervised loss. The learning rate was reduced by a factor of 0.2 after two consecutive epochs without improvement, with a lower bound of 1×10^{-6}.

All models were trained for 50 epochs with a batch size of 2. To improve training stability in the adversarial setting and prevent exploding gradients, global gradient clipping was applied with an ℓ_2-norm threshold of 5.0.

The proposed framework was implemented in Python using TensorFlow 2.19 [38, 39] and Keras 3.10 [40]. OpenCV 4.11.0.86 [41] was used for image processing, scikit-learn 1.6.1 [42] for dataset partitioning, and NiBabel 5.3.1 [43] for medical image input/output. All experiments were conducted on Google Colab Pro using a Python 3 Google Compute Engine environment equipped with an NVIDIA A100-SXM4 GPU (40 GB memory).

4.4. Quantitative comparison with state-of-the-art segmentation models

To assess the effectiveness of the proposed TL-GAN framework, we conducted a comprehensive comparison against five widely adopted state-of-the-art segmentation models: DeepLabV3+ [11], Swin-UNet [3], nnU-Net [2], SegResNet [12], and TransUNet [4].

All models were trained and evaluated under identical experimental conditions using the same six-fold subject-level cross-validation protocol to ensure a fair comparison. All competing architectures were implemented in a consistent 2D slice-wise configuration, following their commonly adopted default pipelines reported in the literature. This setting aligns with the slice-based nature of the proposed approach and is particularly suitable for the limited size of the available annotated dataset.

Model performance was assessed using Accuracy, per-class Dice coefficients, Dice Average (excluding the background class), and Mean IoU, with results reported as mean $\pm$ standard deviation across folds.

As summarized in Table 2, the proposed TL-GAN consistently achieves superior performance across all evaluated metrics when compared with the competing models. Notably, TL-GAN attains the highest Dice Average (0.9343), reflecting improved overall segmentation accuracy across the four clinically relevant tissue classes. The proposed model also yields the best Mean IoU (0.9014), indicating more pre-

cise region-wise overlap between predicted and ground-truth segmentations.

Performance gains are particularly evident in the cortical gray matter classes (Dice 2 and Dice 4), which are known to pose significant challenges due to their thin structures and ambiguous boundaries. In these classes, TL-GAN demonstrates clear improvements over strong transformer-based and convolutional baselines, including TransUNet and nnU-Net. This suggests that the integration of transfer learning with adversarial supervision effectively enhances both local boundary delineation and global anatomical consistency, especially in low-data settings.

The comparative Dice Average results across models are illustrated in Figure 7, while additional analyses, including Mean IoU comparisons (Figure 11), fold-wise Dice stability (Figure 12), and per-class Dice distributions (Figure 13), are provided in the Supplementary Material.

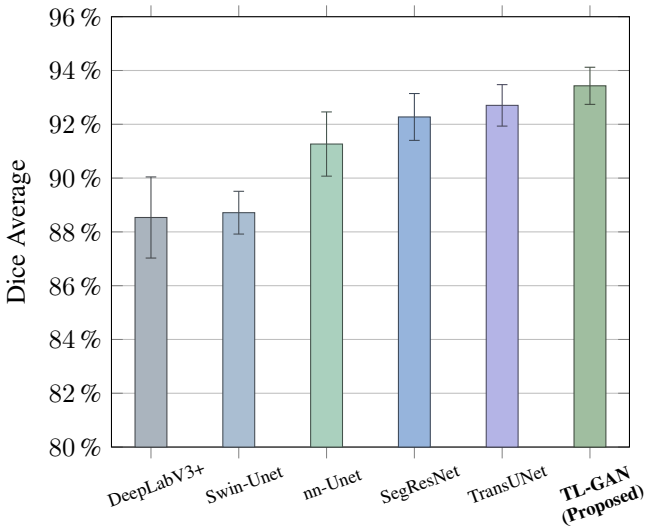

Figure 7. Dice Average (background excluded) for the proposed and state-of-the-art models (mean ± standard deviation across six folds).

4.5. Ablation study

To quantify the individual contributions of transfer learning and adversarial training in the proposed framework, a comprehensive ablation study was conducted. All configurations shared the same backbone architecture and employed the IEP-RGB encoder as the three-channel input representation, ensuring a fair and controlled comparison.

Four training configurations were evaluated:

1. Supervised only: The segmentation network was trained from scratch using supervised loss, without transfer learning and without adversarial training.
2. TL-Supervised: The network was initialized using a pre-trained generator (transfer learning) and trained solely with supervised segmentation loss, without the GAN architecture.
3. GAN only: The network was trained with adversarial supervision, but without transfer learning initialization.
4. TL-GAN (Proposed): The full model, combining both transfer learning and adversarial training.

All configurations were trained using the six-fold cross-validation strategy described in Section 4.2. Performance was evaluated using Accuracy, Dice Average, Mean IoU, and per-class Dice scores, reported as mean ± standard deviation across folds. Quantitative results are summarized in Table 3.

As shown in Table 3, both transfer learning and adversarial training independently improve segmentation performance over the purely supervised baseline. The TL-Supervised configuration yields a substantial gain in Dice Average (+1.7%), highlighting the effectiveness of pre-trained representations. Similarly, adversarial training alone improves boundary consistency and regional coherence, as reflected in higher Dice scores for anatomically challenging classes.

The proposed TL-GAN configuration consistently achieves the best performance across all evaluation metrics, demonstrating that transfer learning and adversarial supervision provide complementary benefits. This improvement is particularly pronounced in cortical regions (Dice 2 and Dice 4), which are more sensitive to structural inconsistencies.

Figure 8 further illustrates the per-fold Dice Average (background excluded), confirming the robustness and stability of the proposed approach across all validation folds.

4.6. RGB Encoder Comparison within the Proposed Framework

To analyze the impact of input representation on segmentation performance, we evaluated three encoder configurations within the proposed TL-GAN framework: (i) the IEP-RGB encoder (proposed), (ii) a 2.5D adjacent-slice encoder, and (iii) a grayscale replication encoder. All configurations shared the same generator–discriminator architecture and training protocol and were evaluated under identical six-fold cross-validation settings.

Quantitative results are summarized in Table 4, reporting Accuracy, Dice Average (excluding background), Mean IoU, and per-class Dice scores as mean ± standard deviation across folds. Per-class Dice distributions and Dice Average are illustrated in Figure 9.

The grayscale replication encoder, which duplicates a single slice across three channels, provides no additional semantic information beyond compatibility with ImageNet-pre-trained weights. Consequently, it yields the lowest overall performance among the evaluated encoders. In contrast, the 2.5D adjacent-slice encoder achieves the highest Dice Average and Mean IoU, benefiting from explicit through-plane anatomical context provided by neighboring slices. This

Table 2. Quantitative comparison of the proposed TL-GAN model with state-of-the-art segmentation models (mean ± standard deviation across six folds).

SOTA Models	DeepLabV3+	Swin-UNet	nnU-Net	SegResNet	TransUNet	**TL-GAN (Proposed)**
Accuracy	0.9791 ± 0.2184	0.9780 ± 0.2459	0.9834 ± 0.0978	0.9853 ± 0.1664	0.9861 ± 0.1563	**0.9871 ± 0.1647**
Dice Average (BG excluded)	0.8853 ± 1.5076	0.8871 ± 0.7952	0.9127 ± 1.1939	0.9227 ± 0.8707	0.9270 ± 0.7707	**0.9343 ± 0.6898**
Dice 0 (BG)	0.9916 ± 0.1106	0.9906 ± 0.1169	0.9927 ± 0.0576	0.9937 ± 0.0898	0.9941 ± 0.0720	**0.9949 ± 0.0789**
Dice 1 (LC WM)	0.9276 ± 1.3364	0.9299 ± 0.8101	0.9469 ± 0.8734	0.9461 ± 1.0785	0.9533 ± 0.6538	**0.9610 ± 0.5424**
Dice 2 (LC C)	0.8472 ± 1.6552	0.8471 ± 0.8236	0.8889 ± 1.2352	0.8936 ± 1.2499	0.8996 ± 1.1024	**0.9077 ± 1.2047**
Dice 3 (RC WM)	0.9216 ± 1.7901	0.9272 ± 1.0620	0.9310 ± 2.8479	0.9531 ± 0.7599	0.9540 ± 0.8080	**0.9590 ± 0.4729**
Dice 4 (RC C)	0.8450 ± 1.5432	0.8443 ± 0.8299	0.8837 ± 0.9567	0.8982 ± 1.0038	0.9012 ± 0.8847	**0.9095 ± 0.8701**
Mean IoU	0.8466 ± 1.8365	0.8478 ± 1.0191	0.8789 ± 1.5851	0.8914 ± 1.1762	0.8970 ± 1.1102	**0.9014 ± 0.9419**

Table 3. Ablation study results evaluating key components of the proposed framework (mean ± standard deviation across six folds).

Configuration	Supervised only	TL-Supervised	GAN only	**TL-GAN (Proposed)**
Transfer Learning	✗	✓	✗	✓
GAN	✗	✗	✓	✓
Accuracy	0.983±0.002	0.987±0.002	0.986±0.001	**0.987±0.002**
Dice Average (BG excluded)	0.912±0.012	0.929±0.007	0.928±0.007	**0.934±0.007**
Dice 0 (BG)	0.993±0.001	0.994±0.001	0.994±0.001	**0.995±0.001**
Dice 1 (LC WM)	0.943±0.014	0.957±0.006	0.955±0.009	**0.961±0.005**
Dice 2 (LC C)	0.882±0.015	0.902±0.010	0.900±0.011	**0.908±0.012**
Dice 3 (RC WM)	0.941±0.010	0.955±0.006	0.955±0.006	**0.959±0.005**
Dice 4 (RC C)	0.882±0.012	0.901±0.008	0.900±0.011	**0.909±0.009**
Mean IoU	0.880±0.014	0.900±0.010	0.892±0.010	**0.901±0.009**

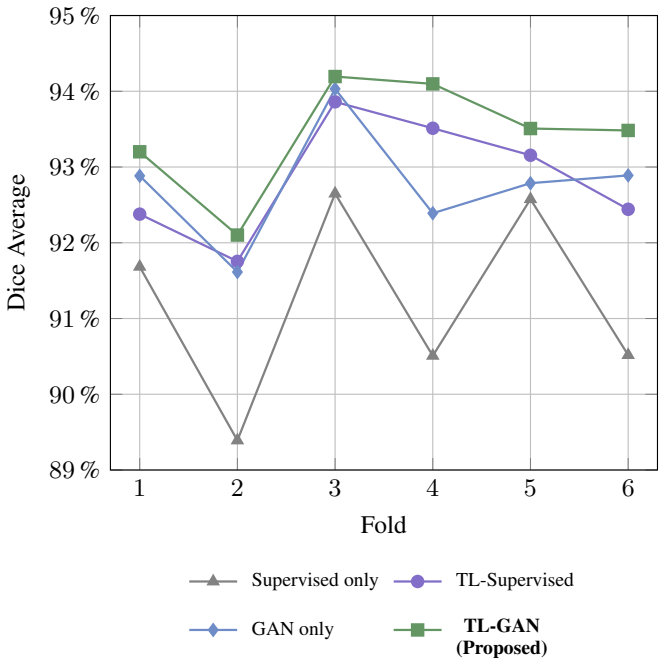

Figure 8. Per-fold Dice Average (background excluded) for comparative ablation setups.

additional spatial continuity is particularly advantageous for cortical structures, whose boundaries evolve smoothly across adjacent slices.

The proposed IEP-RGB encoder, while operating strictly on a single slice, consistently outperforms grayscale replication and achieves competitive performance relative to the 2.5D formulation. By decomposing each slice into complementary intensity, edge-preserving, and positional channels, the IEP-RGB encoder enriches in-plane structural representation and improves boundary delineation without relying on inter-slice information. Although its Dice Average is marginally lower than that of the 2.5D encoder, the difference remains small and consistent across folds.

Importantly, the IEP-RGB encoder offers practical advantages over adjacent-slice approaches. It does not require access to neighboring slices at inference time and is therefore robust to anisotropic resolution, variable slice thickness, and missing or corrupted slices, conditions commonly encountered in real-world clinical MRI datasets. This design choice enhances the general applicability and deployment flexibility of the proposed model, particularly in settings where volumetric consistency cannot be guaranteed.

Overall, these results demonstrate that the IEP-RGB encoder provides a favorable trade-off between segmentation accuracy and robustness. By embedding informative structural priors directly into the channel representation, it enables strong performance while maintaining slice-wise independence, making it well-suited for transfer-learning-based adversarial segmentation frameworks.

Table 4. Encoder comparison results within the proposed framework (mean ± standard deviation across six folds).

Configuration	Grayscale Replication Encoder	2.5D Adjacent-Slice Encoder	IEP-RGB Encoder (Proposed Encoder)
Accuracy	0.9869 ± 0.0015	0.9879 ± 0.0015	0.9871 ± 0.0016
Dice Average (BG excluded)	0.9331 ± 0.0073	0.9375 ± 0.0069	0.9343 ± 0.0069
Dice 0 (BG)	0.9948 ± 0.0007	0.9954 ± 0.0007	0.9949 ± 0.0008
Dice 1 (LC WM)	0.9603 ± 0.0052	0.9619 ± 0.0048	0.9610 ± 0.0054
Dice 2 (LC C)	0.9040 ± 0.0131	0.9142 ± 0.0102	0.9077 ± 0.0120
Dice 3 (RC WM)	0.9600 ± 0.0057	0.9612 ± 0.0046	0.9590 ± 0.0047
Dice 4 (RC C)	0.9081 ± 0.0108	0.9128 ± 0.0097	0.9095 ± 0.0087
Mean IoU	0.9000 ± 0.0101	0.9057 ± 0.0099	0.9014 ± 0.0094

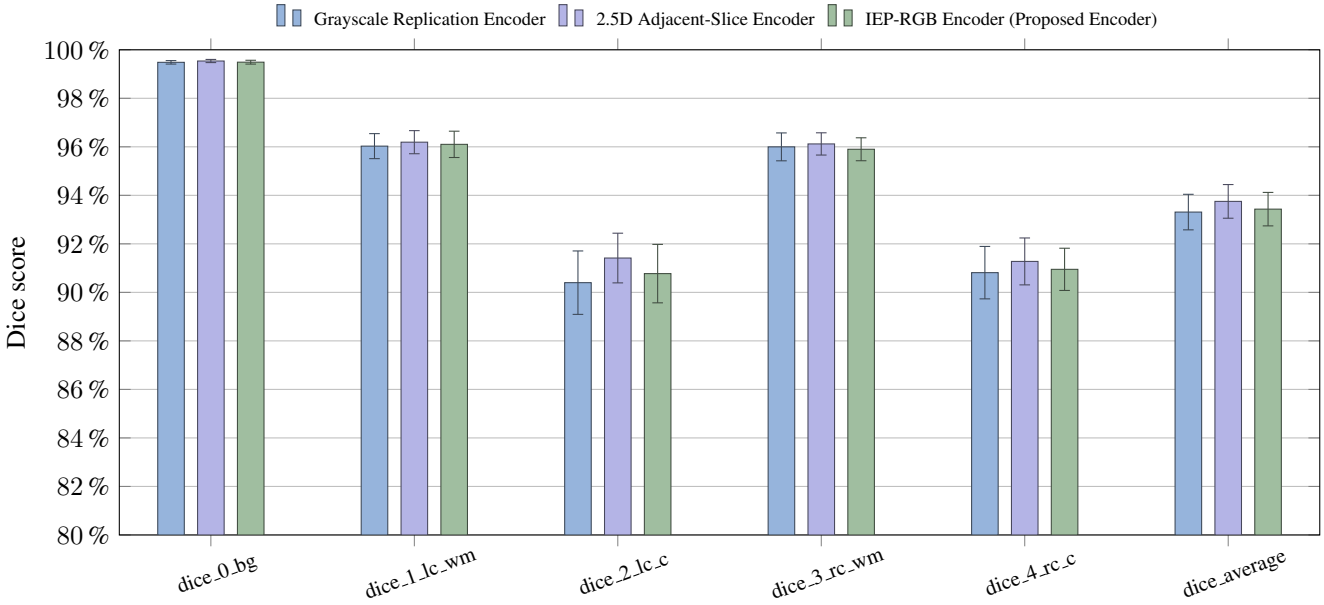

Figure 9. Per-class Dice scores and Dice Average (background excluded) for different encoder configurations evaluated within the proposed model framework (mean ± standard deviation across six folds).

4.7. Qualitative results

Figure 10 illustrates representative qualitative segmentation results produced by the proposed TL-GAN model on test subjects from the six-fold cross-validation. For each example, the input brain MR slice, the predicted multi-class segmentation, and the corresponding ground-truth annotation are shown side by side.

Overall, the predicted masks closely match the ground-truth segmentations across a wide range of anatomical variations and slice positions. The model accurately delineates cortical and subcortical structures, capturing fine-grained boundaries between white matter and cortical regions in both hemispheres. Notably, the predicted segmentations exhibit smooth and anatomically coherent contours, with minimal spurious regions or discontinuities, reflecting the effectiveness of the adversarial training in enforcing local realism.

The qualitative results further demonstrate robustness to variations in brain size, shape, and contrast, including slices with partial brain coverage and reduced tissue visibility. In challenging regions, such as thin cortical areas and tissue interfaces near ventricular boundaries, the proposed model preserves structural consistency and avoids the jagged edges or fragmented predictions commonly observed in purely supervised segmentation approaches.

Taken together, these visual examples corroborate the quantitative findings reported in Secs. 4.4 to 4.6, confirming that the proposed TL-GAN framework not only improves numerical performance metrics but also yields anatomically plausible and visually coherent segmentations suitable for downstream neuroimaging analysis.

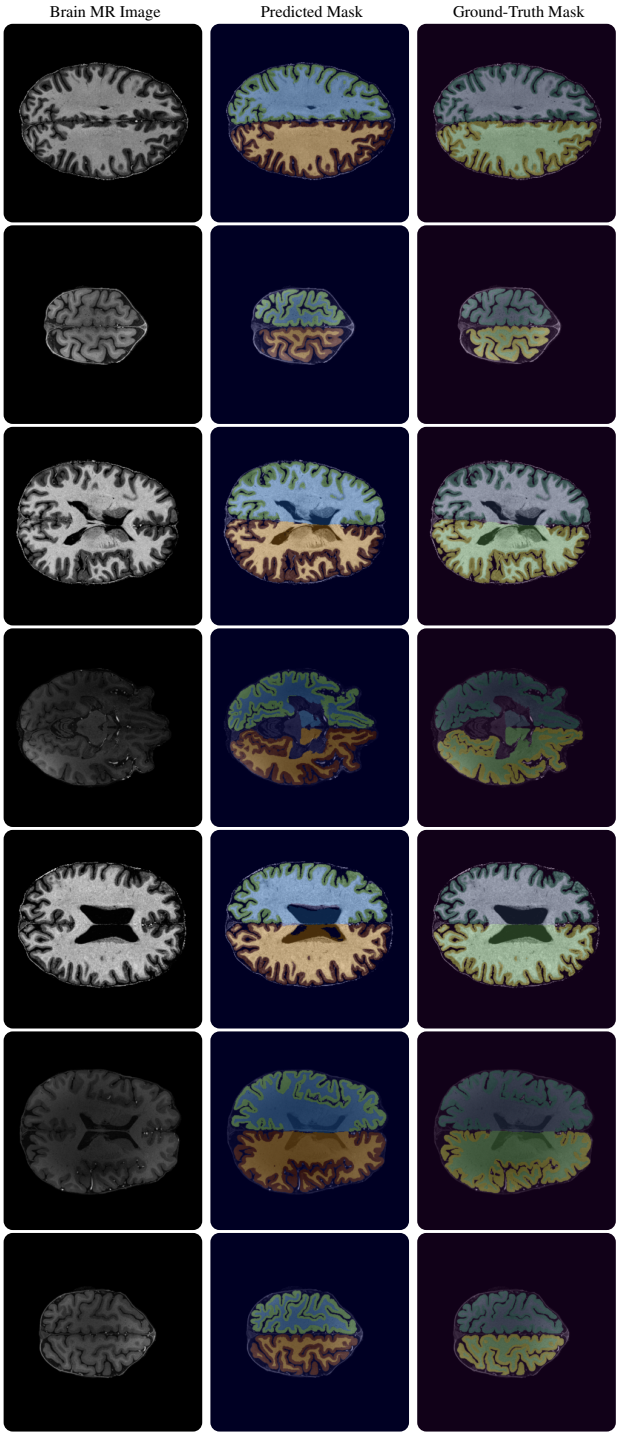

Figure 10. Qualitative segmentation results of the proposed TL-GAN model. From left to right: input brain MR image, predicted multi-class segmentation, and ground-truth annotation for representative test subjects.

5. Discussion

The experimental results demonstrate that the proposed TL-GAN framework effectively combines transfer learning and adversarial regularization to achieve robust and consistent brain MRI segmentation under limited-data conditions. The ablation analysis confirms a clear synergy between the pre-trained generator and the adversarial loss: transfer learning provides semantically rich and stable feature representations, while adversarial feedback acts as a structural regularizer that enforces anatomically plausible segmentation boundaries and reduces overfitting. This interaction leads to improved Dice Average and Mean IoU, as well as reduced performance variance across cross-validation folds, indicating enhanced generalization.

Furthermore, the encoder comparison reveals that the framework is largely insensitive to the specific RGB encoding strategy, underscoring its architectural robustness; however, the proposed IEP-RGB encoder offers the best balance between performance and computational efficiency by embedding complementary intensity, boundary, and positional cues without introducing additional parameters or inference complexity.

In comparison with contemporary SOTA models based solely on supervised U-Net or transformer architectures, the proposed method consistently achieves superior segmentation accuracy, particularly for cortical and white matter regions where subtle intensity variations and ambiguous boundaries pose significant challenges.

Despite these advantages, the current framework operates in a 2D slice-wise setting and was evaluated on a relatively small cohort, which may limit its ability to fully exploit 3D contextual information and capture population-level anatomical variability. Future work will therefore focus on extending the model to volumetric formulations and validating its performance on larger, multi-center datasets.

Overall, these findings highlight the effectiveness of integrating pre-trained representations with adversarial learning as a principled and data-efficient strategy for reliable brain MRI segmentation.

6. Future directions

Although the current study validates the effectiveness of the proposed 2D slice-wise framework, several limitations merit consideration. While slice-wise processing substantially reduces computational cost and hardware requirements, consistent with our objective of developing a deployable model for resource-constrained settings, it inherently ignores inter-slice spatial continuity. Consequently, the model may not fully capture higher-order volumetric dependencies and global anatomical context.

Future work will investigate efficient 3D or hybrid 2.5D formulations that preserve volumetric coherence while main-

taining computational feasibility, particularly in environments with greater hardware availability.

In addition, extending the proposed framework to other neuroimaging tasks, such as brain tumor segmentation, lesion detection, and subtle structural abnormality identification under low-data or domain-shift conditions, represents a promising direction for enhancing robustness and clinical applicability.

7. Conclusion

This study presented a novel slice-wise 2D framework for multi-class brain MRI segmentation that integrates transfer learning with adversarial regularization to address the challenges of limited annotated data and constrained computational resources. By embedding a pre-trained generator within a GAN-based segmentation architecture, the proposed model effectively leverages prior knowledge from a large-scale natural image dataset while refining anatomical consistency through adversarial feedback.

Extensive experimental evaluations using six-fold cross-validation demonstrated that the proposed model consistently outperforms state-of-the-art segmentation models across multiple quantitative metrics and tissue classes, particularly in anatomically subtle gray and white matter regions. These results highlight the model's strong generalization capability, robustness, and suitability for low-data neuroimaging scenarios.

Further analysis showed that the framework is resilient to variations in RGB encoding strategies, indicating architectural stability and reduced dependence on specific preprocessing choices. Combined with its computational efficiency, these properties position the proposed method as a practical and deployable solution for real-world neuroimaging pipelines.

Overall, this work provides an effective and data-efficient pathway for advancing automated brain MRI segmentation, bridging the gap between modern deep learning techniques and practical clinical constraints. Future research will explore extensions toward volumetric modeling and broader clinical applications under domain-shift conditions.

Availability of data and implementation

The data used in this study are publicly available at https://openneuro.org/datasets/ds005216. The source code used in this study is archived at https://doi.org/10.5281/zenodo.18678676. The code will be made publicly available upon journal acceptance.

Declaration of Competing Interest

The authors declare that they do not have any known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Supplementary Material

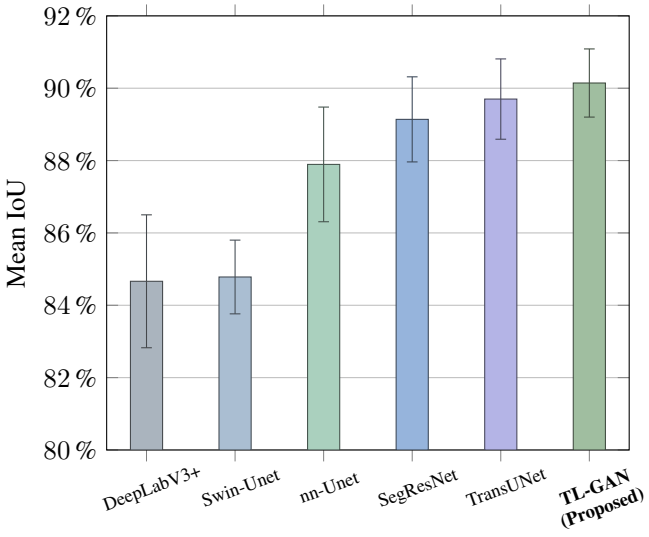

Figure 11. Mean IoU for the proposed and state-of-the-art models (mean ± standard deviation across six folds).

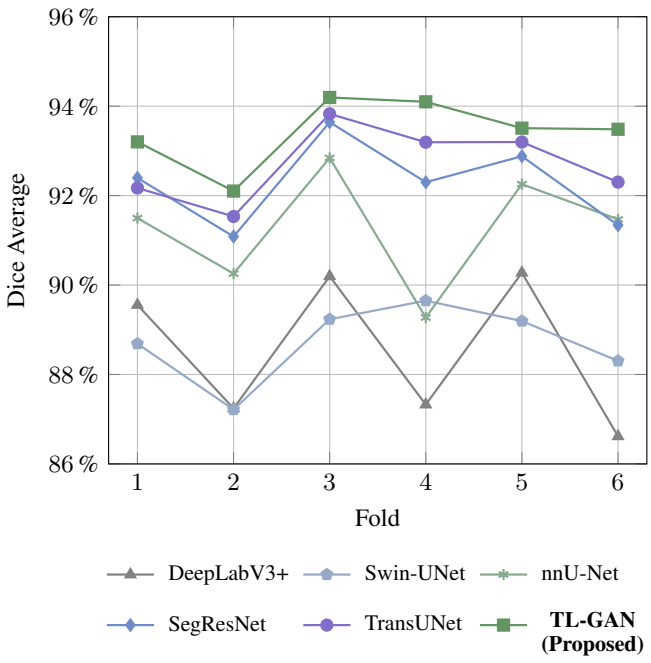

Figure 12. Per-fold Dice Average (background excluded) for the proposed and state-of-the-art models.

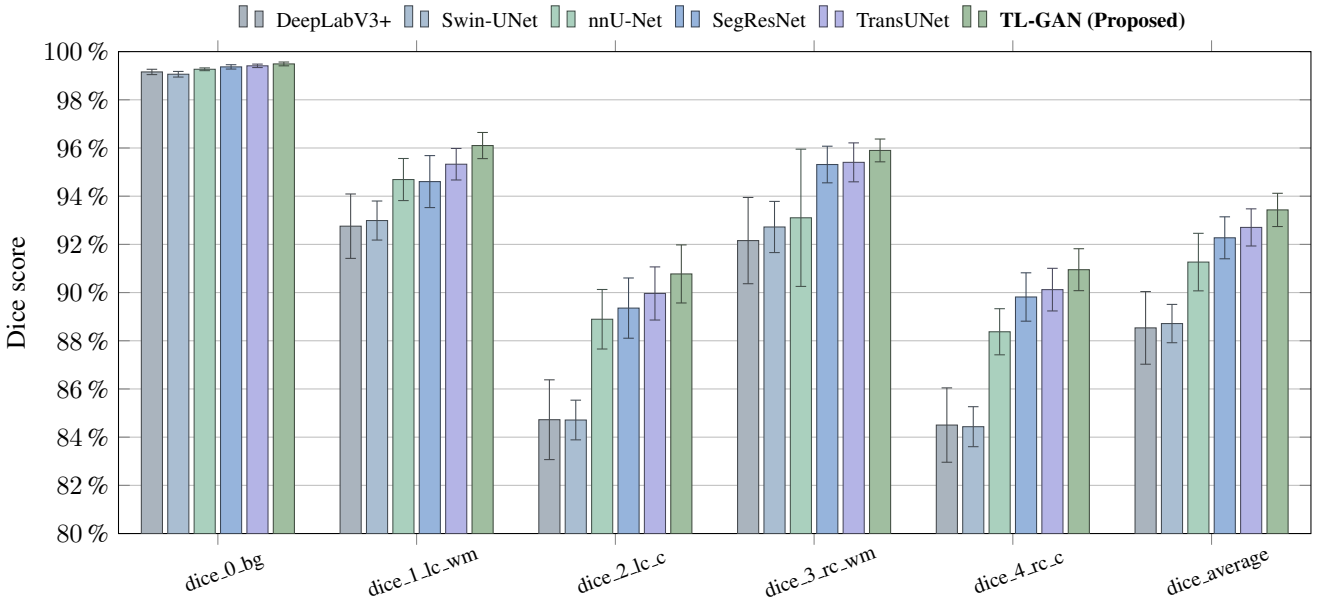

Figure 13. Per-class Dice scores and Dice Average (background excluded) for the proposed and state-of-the-art models (mean ± standard deviation across six folds).

References

[1] O. Ronneberger, P. Fischer, T. Brox, U-Net: Convolutional Networks for Biomedical Image Segmentation, in: N. Navab, J. Hornegger, W. M. Wells, A. F. Frangi (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, Vol. 9351, Springer International Publishing, Cham, 2015, pp. 234–241. doi:10.1007/978-3-319-24574-4_28. 1, 2

[2] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, K. H. Maier-Hein, nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation, Nature Methods 18 (2) (2021) 203–211. doi:10.1038/s41592-020-01008-z. 1, 2, 10

[3] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, M. Wang, Swin-Unet: Unet-Like Pure Transformer for Medical Image Segmentation, in: L. Karlinsky, T. Michaeli, K. Nishino (Eds.), Computer Vision – ECCV 2022 Workshops, Springer Nature Switzerland, Cham, 2023, pp. 205–218. 1, 2, 10

[4] J. Chen, J. Mei, X. Li, Y. Lu, Q. Yu, Q. Wei, X. Luo, Y. Xie, E. Adeli, Y. Wang, M. P. Lungren, S. Zhang, L. Xing, L. Lu, A. Yuille, Y. Zhou, TransUNet: Rethinking the U-Net architecture design for medical image segmentation through the lens of transformers, Medical Image Analysis 97 (2024) 103280. doi:10.1016/j.media.2024.103280. 1, 2, 10

[5] M. Zhang, Q. Sun, Y. Han, M. Zhang, W. Wang, J. Zhang, Generative adversarial DacFormer network for MRI brain tumor segmentation, Scientific Reports 15 (1) (2025) 17840. doi:10.1038/s41598-025-02714-4. 2, 3

[6] S. Xun, D. Li, H. Zhu, M. Chen, J. Wang, J. Li, M. Chen, B. Wu, H. Zhang, X. Chai, Z. Jiang, Y. Zhang, P. Huang, Generative adversarial networks in medical image segmentation: A review, Computers in Biology and Medicine 140 (2022) 105063. doi:10.1016/j.compbiomed.2021.105063. 2

[7] A. Pourmahboubi, N. Arsalani Saeed, H. Tabrizchi, A brain tumor segmentation enhancement in MRI images using U-Net and transfer learning, BMC Medical Imaging 25 (1) (2025) 307. doi:10.1186/s12880-025-01837-4. 2, 3

[8] M. Napravnik, F. Hržić, M. Urschler, D. Miletić, I. Štajduhar, Lessons learned from RadiologyNET foundation models for transfer learning in medical radiology, Scientific Reports 15 (1) (2025) 21622. doi:10.1038/s41598-025-05009-w. 2, 3

[9] A. Khaled, Transfer Learning using Generative Adversarial Networks for MRI Brain Image Segmentation (Aug. 2022). doi:10.20944/preprints202208.0192.v1. 2, 3

[10] L.-C. Chen, G. Papandreou, F. Schroff, H. Adam, Rethinking Atrous Convolution for Semantic Image Segmentation (Dec. 2017). arXiv:1706.05587, doi:10.48550/arXiv.1706.05587. 2

[11] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, H. Adam, Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation, in: V. Ferrari, M. Hebert, C. Sminchisescu, Y. Weiss (Eds.), Computer Vision – ECCV 2018, Springer International Publishing, Cham, 2018, pp. 833–851. 2, 10

[12] A. Myronenko, 3D MRI Brain Tumor Segmentation Using Autoencoder Regularization, in: A. Crimi, S. Bakas, H. Kuijf, F. Keyvan, M. Reyes, T. van Walsum (Eds.), Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries, Springer International Publishing, Cham, 2019, pp. 311–320. 2, 10

[13] Z. Xiong, ResSAXU-Net for multimodal brain tumor segmentation from brain MRI, Scientific Reports 15 (1) (2025) 24179. doi:10.1038/s41598-025-09539-1. 2

[14] R. Yousef, S. Khan, G. Gupta, T. Siddiqui, B. M. Albahlal, S. A. Alajlan, M. A. Haq, U-Net-Based Models towards Optimal MR Brain Image Segmentation, Diagnostics 13 (9). doi:10.3390/diagnostics13091624. 2

[15] N. Mu, Z. Lyu, M. Rezaeitaleshmahalleh, C. Bonifas, J. Gosnell, M. Haw, J. Vettukattil, J. Jiang, S-Net: A multiple cross aggregation convolutional architecture for automatic segmentation of small/thin structures for cardiovascular applications, Frontiers in Physiology Volume 14 - 2023. 2

[16] J. Zhu, R. Zhang, H. Zhang, An MRI brain tumor segmentation method based on improved U-Net, Mathematical Biosciences and Engineering 21 (1) (2023) 778–791. doi:10.3934/mbe.2024033. 2

[17] A. Khorasani, Enhanced glioma semantic segmentation using U-net and pre-trained backbone U-net architectures, Scientific Reports 15 (1) (2025) 31821. doi:10.1038/s41598-025-17895-1.

[18] Y. Pamungkas, E. Triandini, W. Yunanto, Y. Thwe, Impact of Hyperparameter Tuning on ResNet-UNet Models for Enhanced Brain Tumor Segmentation in MRI Scans 5 (2). 2

[19] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative Adversarial Networks (Jun. 2014). arXiv:1406.2661, doi:10.48550/arXiv.1406.2661. 2

[20] M. M. Saad, R. O'Reilly, M. H. Rehmani, A survey on training challenges in generative adversarial networks for biomedical image analysis, Artificial Intelligence Review 57 (2) (2024) 19. doi:10.1007/s10462-023-10624-y. 3

[21] A. Iqbal, M. Sharif, M. Yasmin, M. Raza, S. Aftab, Generative adversarial networks and its applications in the biomedical image segmentation: A comprehensive survey, International Journal of Multimedia Information Retrieval 11 (3) (2022) 333–368. doi:10.1007/s13735-022-00240-x. 3

[22] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE. 3

[23] S. Konstantakos, J. Cani, I. Mademlis, D. I. Chalkiadaki, Y. M. Asano, E. Gavves, G. T. Papadopoulos, Self-supervised visual learning in the low-data regime: A comparative evaluation, Neurocomputing 620 (2025) 129199. doi:10.1016/j.neucom.2024.129199. 3

[24] J. M. Valverde, V. Imani, A. Abdollahzadeh, R. De Feo, M. Prakash, R. Ciszek, J. Tohka, Transfer Learning in Magnetic Resonance Brain Imaging: A Systematic Review, Journal of Imaging 7 (4) (2021) 66. doi:10.3390/jimaging7040066.

[25] Z. N. K. Swati, Q. Zhao, M. Kabir, F. Ali, Z. Ali, S. Ahmed, J. Lu, Content-Based Brain Tumor Retrieval for MR Images Using Transfer Learning, IEEE Access 7 (2019) 17809–17822. doi:10.1109/ACCESS.2019.2892455.

[26] H. Jiang, J. Guo, H. Du, J. Xu, B. Qiu, 1 University of Science and Technology of China, Hefei, Anhui 230026, China, 2 School of Electrical Engineering and Automation, Hefei University of Technology, Hefei, Anhui 230009, China, Transfer learning on T1-weighted images for brain age estimation, Mathematical Biosciences and Engineering 16 (5) (2019) 4382–4398. doi:10.3934/mbe.2019218. 3

[27] D. S. Terzi, N. Azginoglu, In-Domain Transfer Learning Strategy for Tumor Detection on Brain MRI, Diagnostics 13 (12) (2023) 2110. doi:10.3390/diagnostics13122110. 3

[28] Z. U. Abidin, R. A. Naqvi, A. Haider, H. S. Kim, D. Jeong, S. W. Lee, Recent deep learning-based brain tumor segmentation models using multi-modality magnetic resonance imaging: A prospective survey, Frontiers in Bioengineering and Biotechnology Volume 12 - 2024. 3

[29] M. Abdollahi, H. Davoudi, M. Ebrahimi, Combined Medical Image Super-Resolution and Modality Translation Using GAN Transformer-Based Model, in: 2023 International Conference on Computational Science and Computational Intelligence (CSCI), IEEE, Las Vegas, NV, USA, 2023, pp. 1133–1138. doi:10.1109/CSCI62032.2023.00186. 3

[30] L. Mahler, J. Steiglechner, B. Bender, T. Lindig, D. Ramadan, J. Bause, F. Birk, R. Heule, E. Charyasz, M. Erb, V. J. Kumar, G. E. Hagberg, P. Martin, G. Lohmann, K. Scheffler, Submillimeter Ultra-High Field 9.4 T Brain MR Image Collection and Manual Cortical Segmentations, Scientific Data 12 (1) (2025) 635. doi:10.1038/s41597-025-04779-2. 4

[31] A. Avesta, S. Hossain, M. Lin, M. Aboian, H. M. Krumholz, S. Aneja, Comparing 3D, 2.5D, and 2D Approaches to Brain Image Auto-Segmentation., Bioengineering (Basel, Switzerland) 10 (2). doi:10.3390/bioengineering10020181. 4

[32] K. He, X. Zhang, S. Ren, J. Sun, Deep Residual Learning for Image Recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Las Vegas, NV, USA, 2016, pp. 770–778. doi:10.1109/CVPR.2016.90. 6

[33] C. Gu, M. Lee, Deep Transfer Learning Using Real-World Image Features for Medical Image Classification, with a Case Study on Pneumonia X-ray Images, Bioengineering 11 (4) (2024) 406. doi:10.3390/bioengineering11040406. 6

[34] H. Zhang, A. M. Valcarcel, R. Bakshi, R. Chu, F. Bagnato, R. T. Shinohara, K. Hett, I. Oguz, Multiple Sclerosis Lesion Segmentation with Tiramisu and 2.5D Stacked Slices, in: D. Shen, T. Liu, T. M. Peters, L. H. Staib, C. Essert, S. Zhou, P.-T. Yap, A. Khan (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2019, Springer International Publishing, Cham, 2019, pp. 338–346. 6

[35] I. Sobel, G. Feldman, A 3×3 isotropic gradient operator for image processing, Pattern Classification and Scene Analysis (1973) 271–272. 6

[36] P. Isola, J. -Y. Zhu, T. Zhou, A. A. Efros, Image-to-Image Translation with Conditional Adversarial Networks, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5967–5976. doi:10.1109/CVPR.2017.632. 7

[37] D. P. Kingma, J. Ba, Adam: A Method for Stochastic Optimization (Jan. 2017). arXiv:1412.6980, doi:10.48550/arXiv.1412.6980. 10

[38] T. Developers, TensorFlow, Zenodo (Mar. 2025). doi:10.5281/zenodo.15009305. 10

[39] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Y. Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, Xiaoqiang Zheng, TensorFlow: Large-scale machine learning on heterogeneous systems (2015). 10

[40] F. Chollet, et al., Keras (2015). 10

[41] G. Bradski, The OpenCV library, Dr. Dobb's Journal of Software Tools. 10

[42] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830. 10

[43] M. Brett, C. J. Markiewicz, M. Hanke, M.-A. Côté, B. Cipollini, D. Papadopoulos Orfanos, P. McCarthy, D. Jarecka, C. P. Cheng, E. Larson, Y. O. Halchenko, M. Cottaar, S. Ghosh, D. Wassermann, S. Gerhard, G. R. Lee, Z. Baratz, B. Moloney, H.-T. Wang, E. Kastman, J. Kaczmarzyk, R. Guidotti, J. Daniel, O. Duek, A. Rokem, M. Scheltienne, C. Madison, A. Sólon, F. C. Morency, M. Goncalves, R. Markello, C. Riddell, C. Burns, J. Millman, A. Gramfort, J. Leppäkangas, J. J. van den Bosch, R. D. Vincent, H. Braun, K. Subramaniam, A. Van, J. H. Legarreta, K. J. Gorgolewski, P. R. Raamana, J. Klug, R. Vos de Wael, B. N. Nichols, E. M. Baker, S. Koudoro, S. Hayashi, B. Pinsard, C. Haselgrove, M. Hymers, O. Esteban, F. Pérez-García, G. Becq, J. Dockès, N. N. Oosterhof, B. Amirbekian, H. Christian, I. Nimmo-Smith, L. Nguyen, P. Suter, S. Reddigari, S. St-Jean, E. Panfilov, E. Garyfallidis, G. Varoquaux, J. Newton, K. S. Hahn, L. Waller, O. P. Hinds, Sandro, B. Fauber, B. Dewey, F. Perez, J. Roberts, J.-B. Poline, J. Stutters, K. Jordan, M. Cieslak, M. E. Moreno, T. Hrnčiar, V. Haenel, Y. Schwartz, B. C. Darwin, B. Thirion, C. Gauthier, I. Solovey, I. Gonzalez, J. Palasubramaniam, J. Lecher, K. Leinweber, K. Raktivan, M. Calábková, P. Fischer, P. Gervais, S. Gadde, T. Ballinger, T. Roos, V. R. Reddam, freec84, Nipy/nibabel: 5.3.1, Zenodo (Oct. 2024). doi:10.5281/zenodo.13936989. 10