Automatic Vestibular Schwannoma Segmentation

Sonja Poe Sam Wu Eli Conlin Gaston Longhitano
Boston University
{svpoe,samwu,econlin,gastonl}@bu.edu

Abstract

Manual delineation of vestibular schwannoma (VS) on magnetic-resonance imaging is slow and prone to inter-observer bias, yet remains the clinical norm for tracking tumour growth. We revisit the task from two complementary angles. First, we mitigate the pronounced size bias of the public VS-SEG dataset by synthesising 3-D T1 and T2 volumes with Med-DDPM, conditioning on automatically generated masks that emulate sub-voxel tumours and atypical anatomic locations. Second, we refine a 2.5-D nnU–Net with (i) a single transformer bottleneck (T), (ii) the same bottleneck modulated by a learnable residual gate (TR), and (iii) an early cross-modality fusion decoder driven by a pre-trained Swin encoder. On the public VS-SEG benchmark our gated bottleneck attains a Dice of 0.906 *on native T2 and* 0.560 *on the size-balanced T1 + T2 set—reducing the gap to the current state-of-the-art. Diffusion-based augmentation consistently boosts all configurations, whereas simple early fusion yields unstable training and lower accuracy, highlighting the challenge of aligning heterogeneous contrasts. Our findings advocate for size-aware data augmentation and lightweight global context as key ingredients for deployable VS segmentation.*

1. Introduction

Tumor segmentation is standard practice for evaluating growth and disease progression in various pathologies, including vestibular schwannomas (VS), a benign tumor on the vestibulocochlear nerve (cranial nerve VIII) [5]. CN VIII transmits information about hearing and balance from the inner ear to the brainstem. VS may occur incidentally or may be associated with genetic conditions such as neurofibromatosis type II (NFII) [24]. While benign, the growth of a vestibular schwannoma is essential to track due to the sensitive location of the growth [14, 34]. The CN VIII is directly next to the facial nerve CN VII and in close contact with other critical structures [2].

Tracking the growth of the tumor is essential for evaluating disease progression and determining whether there is a need to operate. In the case of a larger tumor where surgery may be necessary, long-term postoperative tracking is essential [26]. The current gold standard for evaluating tumor size is through manual segmentation [21, 22]. This process is time-consuming, as each individual slice of the scan is manually processed. Long-term monitoring requires consistency in measurement methods. Differences in segmentation and measurement methods between raters may falsely suggest changes in tumor size. Given the limitations of manual segmentation, it is important to develop methods to standardize and automate the segmentation process [22, 34].

Previous work shows that an automated segmentation model developed using a restrictive set of siloed institutional data can be successfully adapted for data from different imaging systems and patient populations [28]. However, significant shortcomings remain that likely reflect bias and limitations of the data used to train the model [22, 34]. A primary limitation is the relatively small internal dataset used for external validation, which constrains the model's applicability across the full spectrum of tumor sizes and imaging protocols [9, 17]. As the training data consisted primarily of larger tumors undergoing Gamma Knife radiosurgery, very small lesions were underrepresented, resulting in failures to detect these tumors in the validation set [28]. In addition, only T1 post-contrast sequences were studied, leaving the performance of the model on T2 and other sequences untested [28]. These factors highlight the need for more comprehensive datasets that capture a wider range of tumor characteristics and diverse MRI protocols to ensure broader generalization [23, 35].

Our main contributions are twofold: (i) we incorporate data augmentation strategies to expose the model to a broader array of tumors with variable sizes and intensities, and (ii) we introduce targeted modifications to the underlying architecture of the state-of-the-art model used in previous work. We propose a transformer bottleneck between the encoder and decoder of the UNet with and without residual connection to better accommodate heterogeneous magnetic resonance protocols and contrast levels [23]. We systematically address the shortcomings identified in previous studies

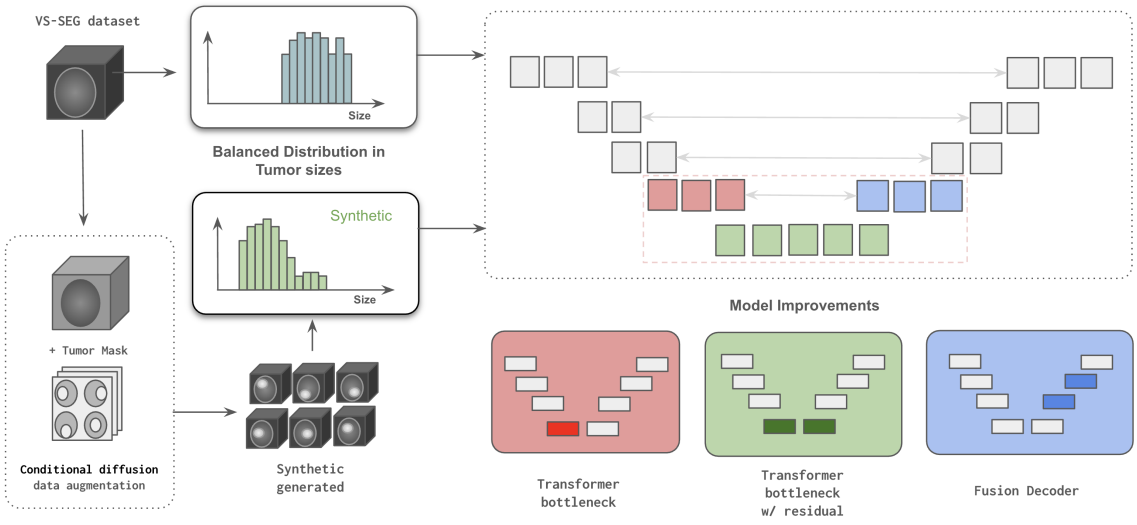

Figure 1. Overview of our Vestibular Schwannoma Segmentation pipeline. Starting from the VS-SEG dataset (top left), we extract real tumor masks and employ conditional diffusion–based augmentation (dashed box, bottom left) to generate synthetic volumes that fill under-represented size bins (green histogram vs. original blue). The combined real + synthetic data achieve a balanced tumor-size distribution and are fed into our segmentation network (large right box). Within the network, we evaluate three architectural enhancements—Transformer bottleneck (red), Transformer bottleneck w/ residual (green), and Fusion Decoder (blue)—each targeting different layers of the V-shaped encoder–decoder backbone.

and we hypothesize that these improvements will open the door for future research on the adaptability of automated vestibular schwannoma segmentation (AVS) in real-world clinical environments. Figure 1 summarizes the proposed architecture pipeline.

2. Methodology

A challenge in the 3D segmentation of medical images is the non-isotropic nature of the scans. 3D convolutions work best in isotropic images, where the voxel dimensions are equal [13]. The 2.5D CNN architecture proposed in previous work [26] combats these issues using 2D convolution on the sharp in-plane view, axial slice. Afterward, these features are downsampled, they are nearly isotropic, and now 3D convolutions are incorporated to get context across slices [26]. The principal advancement to the traditional U-Net is the addition of supervised spatial attention mapping [25]. The feature map is retrieved from each convolutional layer, passed through two convolutional layers reducing to a single channel and creating a map. Each value represents the importance of the given spatial locations [1, 26].

The attention map is multiplied by the input feature map. This is applied with a residual connection to retain the original features. For the supervised portion, the ground truth tumor is downsampled via average pooling in order to match the resolution of the attention map. Loss is computed between the attention map and the downsampled ground truth mask [25].

While good performance in automated segmentation methods has been reported in the literature, in order for automated segmentation methods to be clinically applicable, they need to generalize well with diverse images such as MRIs from different imaging protocols and other patient populations [28]. Suresh et al. validated an automated segmentation method trained on isolated data from one institution (publicly available dataset shared by Shapey et. al) using nnUnet, a modification of UNet, which is a deep learning method for image segmentation that uses a convolutional neural network. While this study demonstrated reasonable performance in different data, there is still a need for improving generalization and focusing on structures that are harder to detect.

We address limitations inherent in the current model architecture and the absence of small tumors within the training dataset by incorporating synthetic data generated via diffusion models. Recent work provides empirical evidence and shows that diffusion-based augmentation effectively covers rare and small-lesion cases and mitigates domain shift [16, 29]. These diffusion models can produce artificial samples that closely resemble underrepresented

cases [7, 9]. To that end, our goal is to broaden the range of tumor sizes and morphological features through synthetic examples, thereby exposing the model to a more diverse set of patterns and challenging scenarios [34]. In addition, we propose targeted modifications to the underlying transformer architecture to enhance performance further. Current work confirms that inserting Transformer blocks between encoder/decoder (or within skip paths) improves global-context capture and accuracy [4, 15, 20].

2.1. Datasets

We leverage the vestibular-schwannoma-SEG database (VS-SEG), an open annotated dataset, and baseline algorithm [26]. This database contains a labeled dataset of MRI images (in DICOM format) collected on 242 consecutive patients with vestibular schwannomas (VS) undergoing radiosurgery [32]. Additionally, registration matrices (.tfm format) and segmentation contour lines (JSON format) are provided and used to identify the tumor region in the images. All tumor structures were manually segmented in consensus by the treating neurosurgeon and physicist [26].

The dataset contains more than 26 GB of images, with 3 for each patient. Using 3DSlicer, a software package for image analysis and scientific visualization of medical images, a 3D reconstruction of the patient's head can be created, along with images identifying the segmented tumor region [18]. We utilized these images to train our model, as they provide images and visualizations of the tumor region.

We split the dataset into training, validation, and test sets, following standard machine learning practices aiming for a robust model evaluation and preventing overfitting. Specifically, 70% of the data is used for training, 15% for validation, and 15% for testing. The training set optimize the CNN's weights, while the validation set will help tune hyperparameter and assess model performance during training. Finally, the unseen test set serves as the final benchmark to evaluate the generalization capability of the model on new, unexposed patient data.

We acquire the imaging data for our analysis from The Cancer Imaging Archive (TCIA) via the National Biomedical Imaging Archive (NBIA) platform. Following dataset identification, patient images of interest are added to the data basket within the NBIA web interface [6]. We download the data using the NBIA Data Retriever, a specialized tool that streamlines bulk data acquisition by managing standardized directory structures and DICOM metadata organization. The Data Retriever provides an interface to download directly from the curated data basket. We use the Linux-based command-line interface of the tool.

2.2. Data Preprocessing

Preprocessing of the VS-SEG dataset involve a multi-step pipeline designed to standardize file structures and convert imaging data into formats suitable for downstream analysis. The original DICOM files and associated metadata—including tumour contours in JSON format and registration matrices in .tfm format—were first downloaded from TCIA and restructured using a custom Python script. This script reorganized the heterogeneous directory layout into a consistent format, aligning with the structure of the supplementary contour and registration files. Following this, files from the three modalities (images, contours, matrices) were manually merged to create unified subject-specific folders. Conversion to NIfTI format was performed using 3D Slicer with the SlicerRT extension. The command-line utility executed a preprocessing script to convert DICOM volumes and 2D contour annotations into registered 3D NIfTI segmentations. Optional flags allowed for rigid registration across modalities (e.g., aligning T2 to T1-weighted images) and export of all annotated structures beyond the tumour. This pipeline produced spatially aligned volumetric data with consistent naming conventions.

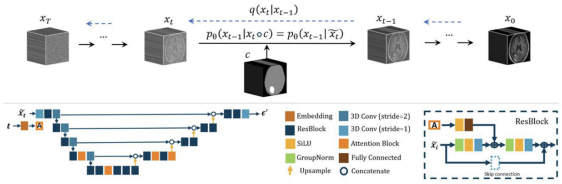

Figure 2. Med-DDPM, a conditional diffusion model. The top part illustrates the conditioning mechanism, encompassing the forward diffusion process $q(\mathbf{x}_t \mid \mathbf{x}_{t-1})$ and the denoising process $p_\theta(\mathbf{x}_{t-1} \mid \tilde{\mathbf{x}}_t)$. The conditioning mask $\mathbf{c}$ is concatenated with the input image $\mathbf{x}_t$, forming the concatenated image $\tilde{\mathbf{x}}_t$ used by p_θ for denoising. The bottom row presents a detailed view of the noise predictor U-Net model θ [7].

2.3. Semantic 3D brain MRI synthesis

We leverage Med-DDPM [7], a conditional denoising diffusion probabilistic model, as our data augmentation backbone. Med-DDPM is a state-of-the-art approach for semantic 3D brain MRI generation: it takes an input segmentation mask and synthesizes a corresponding MRI volume, optionally producing multiple modalities (T1, T1ce, T2, FLAIR) in a single pass. This enables controlled augmentation that addresses data scarcity and privacy constraints by generating realistic, anatomically coherent images from mask inputs. The overall architecture (Figure 2) is a 3D U-Net based diffusion model conditioned on the mask via channel-wise concatenation.

2.3.1 Mask generation

To enrich the training set with rare and challenging examples, we programmatically generate additional tumor masks that simulate missing or underrepresented cases. Starting from the available segmentation maps, we create variant masks that shrink the tumor size and alter its location while preserving anatomical plausibility.

Scaled-down samples. We synthesize small tumor masks to model diminutive lesions that were absent in the original training data. Each original tumor mask is algorithmically scaled down (via downsampling) to create a new mask with substantially smaller tumor volume (from 0.90 to 0.41). For example, we generate masks with volumes on the order of 10^{-2} mL, mimicking the tiny vestibular schwannomas (approximately a few voxels in size) that a prior model failed to detect [28]. These scaled-down masks are then used as conditioning inputs to the generative model, yielding synthetic images of very small tumors. By augmenting the training set with such cases, we expose the segmentation network to the end of the tumor size spectrum. This is crucial because a segmentation model trained on a homogeneous set of larger tumors [32] struggled to recognize lesions below a certain size in external validation [28]. Incorporating synthetic images of tiny tumors addresses this weakness by teaching the model how small tumors appear in MRI, improving its sensitivity to minute pathology [12, 19]. Figure 3 shows an example of segmentation masks scaled down derived from an original tumor mask.

Shifted tumor masks. We also augment the spatial distribution of tumors by generating shifted tumor masks that reposition the lesion within the brain. Given a tumor mask, we apply random small translations (on the order of a few millimeters) to the tumor region within the mask volume. This process shifts the tumor to a new anatomical location while keeping the mask's shape and size unchanged. We constrain the shifts to remain in plausible intracranial locations. For example, a vestibular schwannoma mask might be shifted along the cerebellopontine angle region or towards the petrous temporal bone, but not into impossible locations outside the cranial volume. These spatial augmentations counteract anatomical biases in the training data. If all training tumors occupy similar locations (e.g., centered in the internal acoustic canal in a homogeneous dataset [32]), the model may implicitly learn location-specific features and fail to segment tumors in uncommon positions. By training on images where the tumor appears in varied locations, we encourage the model to rely on semantic cues of the tumor itself rather than its expected position. Moreover, some shifted masks deliberately place tumors adjacent to structures that have a similar intensity to pathology. This approach helps the model distinguish tumor tissue from look-alike anatomy. For example, one failure mode in the external study was a hyperintense petrous bone being mistakenly segmented as tumor [28], likely because the training lacked examples of tumors near bright bone. Our shifted-mask augmentation produces cases with tumors abutting different tissue interfaces (including bone), thereby introducing the needed variability for the model to learn proper discrimination. In summary, the mask generation strategy (scaled-down and shifted tumors) yields diverse semantic layouts for conditioning the synthesis model, targeting the specific rare scenarios (tiny lesions, atypical locations) that were problematic in previous segmentation efforts. Figure 6 in the Appendix shows an example of shifted segmentation masks.

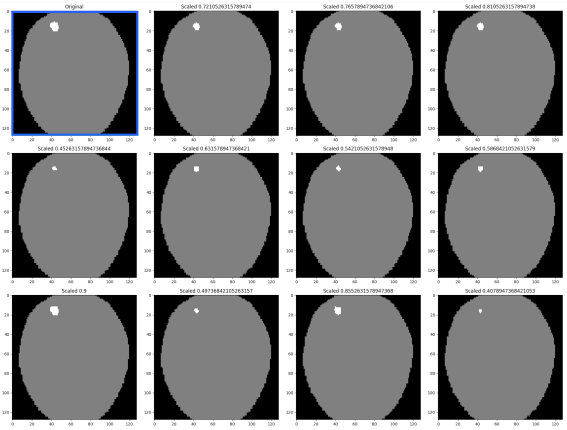

Figure 3. Illustration of scaled-down segmentation masks derived from an original tumor mask. The scaling factors range from 0.41 to 0.90, systematically simulating smaller tumor sizes to address model sensitivity limitations previously reported in segmentation validation studies.

2.3.2 Sample generation

Given an arbitrary mask (original or augmented as above), Med-DDPM generates a realistic 3D MRI volume that adheres to the mask's anatomy. We follow the diffusion model sampling procedure conditioned on the mask [7]. Let m denote the segmentation mask (with one-hot channels for structures) and let x denote the MRI image volume. The generative model is trained to learn the distribution $p(x|m)$ by gradually denoising random noise into an image consistent with m. Formally, during training, we add Gaussian noise to a real image x_0 in T steps: $x_t = \sqrt{\bar{\alpha}_t}\, x_0 + \sqrt{1-\bar{\alpha}_t}\, \epsilon$, where t indexes the diffusion timestep, $\epsilon \sim \mathcal{N}(0, I)$, and $\{\alpha_t\}$ is a variance schedule (we use a cosine schedule as in [7]). The model is a 3D U-Net ϵ_θ that takes as input the noisy image x_t concatenated with the mask m and the timestep t, and is trained to predict the added noise ϵ. The loss follows the standard DDPM formulation $L(\theta) = \mathbb{E}_{x_0,m,\epsilon,t}\left[\|\epsilon - \epsilon_\theta(x_t, m, t)\|^2\right]$.

Method	VS-SEG T1	VS-SEG T2	VS-SEG T1/T2	VS-SEG T1+	VS-SEG T2+	VS-SEG T1+/T2+
Baseline (Wang *et al.*, 2019) [32]	0.610	NA	NA	NA	NA	NA
Baseline (Shapey *et al.*, 2021) [26]	0.945	0.909	NA	**0.950**	**0.911**	0.580
+ Transformer bottleneck (Ours)	0.920	0.892	NA	0.935	0.907	0.538
+ Transformer bottleneck w/ residual (Ours)	0.926	0.906	NA	0.929	0.908	0.560
+ Fusion Decoder (Ours)	NA	NA	0.750	NA	NA	NA

Table 1. Mean Dice score over 46 random volumetric samples for various VS-SEG configurations. "+" indicates inclusion of diffusion-generated data; "NA" denotes unavailable results.

At inference (image synthesis) time, we sample $x_T \sim \mathcal{N}(0, I)$ and iteratively apply the denoising process conditioned on m. At each step t, the model predicts $\hat{\epsilon} = \epsilon_\theta(x_t, m, t)$. We then compute the denoised image estimate $\mu_\theta(x_t, m, t) = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\,\hat{\epsilon}\right)$, which is the DDPM approximation of x_0 (the underlying clean image) at step t. The sample is then updated as:

$$x_{t-1} = \mu_\theta(x_t, m, t) + \sigma_t z,$$

where $z \sim \mathcal{N}(0, I)$ is fresh Gaussian noise and σ_t is the prescribed noise scale for step t [7]. This procedure is repeated for $t = T, T-1, \ldots, 1$ to obtain x_0, which is a synthetic MRI volume conditioned on mask m. The conditioning ensures that the output image exhibits the structures indicated in m (tumor at a given location, etc.) with high fidelity. In practice, the conditioned diffusion sampling produces anatomically plausible brain images where the tumor is precisely shaped and located as specified by the mask.

The stochastic nature of diffusion means that multiple diverse images can be generated from the same mask. This is a key advantage for data augmentation: for each mask (including each scaled or shifted variant), we sample several different images. All such images share the same ground-truth segmentation (the mask m) but have varied appearance characteristics. The model introduces realistic variation in background anatomy and in tumor texture and intensity. For instance, across different samples from one mask, the synthetic tumor may exhibit variable contrast enhancement (from barely enhancing to strongly enhancing) and differing surrounding edema extents, mirroring the heterogeneity seen in real patients. In our experiments, we observed that some samples displayed tumors with clear, well-defined borders, while others generated slight blur or infiltrative edges consistent with edema in T2-weighted contrasts. Intensity profiles of both normal tissues and the lesion also differ between samples. For instance, one image may have a uniformly isointense tumor, whereas another tumor might contain internal heterogeneity or necrotic darker regions. Med-DDPM intentionally captures this diversity in enhancement, edema, and intensity profiles. By training the segmentation model on this broadened distribution of appearances, we aim to reduce overfitting to any single imaging phenotype of the tumor. In effect, the synthetic data augments the number of training samples and their qualitative diversity, in line with our hypothesis in helping the downstream model generalize better. We directly counteract limitations observed in earlier studies by producing tiny tumors that were missing in training, as well as tumors against varied anatomical backgrounds, thereby addressing the failure to detect small tumors and the misclassifications of unusual intensities reported in [28]. From the 5,000 synthetic samples generated, we randomly select 500 T1 images and 500 T2 images for our experiments.

2.4. Model architecture

The baseline adopted from prior work is a conventional encoder–decoder **U–Net**. We extend this backbone with two lightweight variants that retain the same convolutional encoder and decoder but differ in how global context is injected at the bottleneck:

- **Variant T** inserts a single self-attention block—the *transformer bottleneck*—at the coarsest resolution to capture long-range slice-to-slice dependencies.
- **Variant TR** augments this bottleneck with a learnable residual gate α that adaptively blends the transformer output with the original convolutional features.
- **Fusion decoder** extracts a tumour-centred crop from co-registered T1-w and T2-w volumes, embeds the paired patches as tokens, and feeds their concatenation through a pretrained Swin-Transformer encoder; the fused representation is then forwarded to the standard nnU-Net decoder, enabling early cross-modality attention within the region of interest.

Their implementation details are provided in the subsections that follow.

2.4.1 Transformer bottleneck

A transformer bottleneck is inserted between the U–net encoder and the decoder. Self-attention operates in the coarsest resolution, where feature maps are smallest, and thus the computational overhead is limited. The bottleneck comprises one multi-head self-attention (MHSA) layer followed by a two-layer position-wise feed-forward network (FFN)

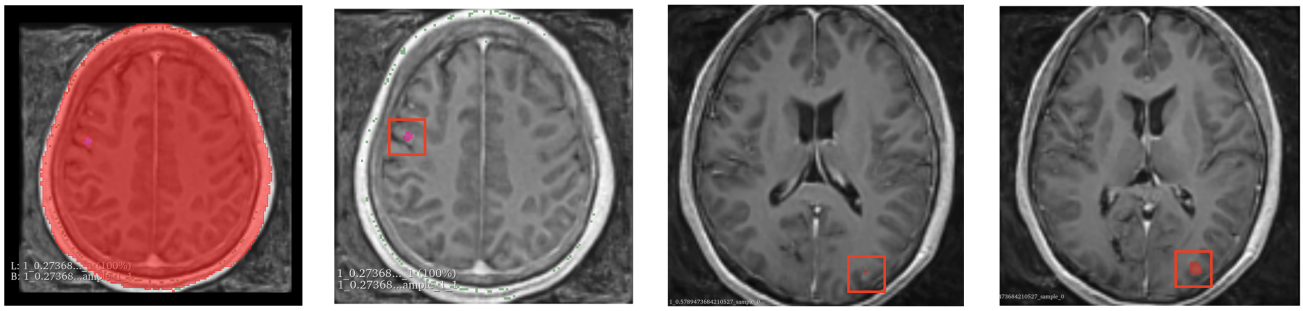

Figure 4. Example of synthetic MRI generation using shifted and scaled tumor masks. From left to right: whole brain segmentation mask (highlighted in red), synthetic sample with scaled tumor (red box), and synthetic samples with scaled + shifted tumor positions (red boxes). This augmentation technique diversifies the spatial distribution and size of tumors in the training data, addressing previously reported limitations related to small tumors and anatomical variability.

with *ReLU*, consistent with TransUNet [3] and the general ViT paradigm [8].

Let the encoder output be

$$\mathbf{E} \in \mathbb{R}^{C_e \times D \times H \times W}. \tag{1}$$

A $1\times1\times1$ convolution projects it to an embedding of width C_p:

$$\mathbf{Z} = W_p * \mathbf{E}, \quad \mathbf{Z} \in \mathbb{R}^{C_p \times D \times H \times W}. \tag{2}$$

Flattening the spatial axes ($N = DHW$) yields the token sequence

$$\mathbf{z} = \text{reshape}(\mathbf{Z}) \in \mathbb{R}^{N \times C_p}. \tag{3}$$

Sinusoidal positional encodings $\mathbf{P} \in \mathbb{R}^{N \times C_p}$ [30] give $\mathbf{x} = \mathbf{z} + \mathbf{P}$.

Multi-head self-attention. For each head h:

$$\mathbf{Q}_h = \mathbf{x}W_h^Q,\ \mathbf{K}_h = \mathbf{x}W_h^K,\ \mathbf{V}_h = \mathbf{x}W_h^V, \tag{4}$$

$$\text{head}_h = \text{softmax}\Big(\frac{\mathbf{Q}_h\mathbf{K}_h^\top}{\sqrt{d_k}}\Big)\mathbf{V}_h. \tag{5}$$

Heads are concatenated and projected:

$$\text{MHA}(\mathbf{x}) = \text{Concat}(\text{head}_1, \ldots, \text{head}_H)W^O. \tag{6}$$

Residual connections and an FFN complete the block:

$$\mathbf{h}_1 = \mathbf{x} + \text{MHA}(\mathbf{x}), \tag{7}$$
$$\mathbf{h}_2 = \mathbf{h}_1 + \text{FFN}(\mathbf{h}_1), \tag{8}$$
$$\text{FFN}(\mathbf{u}) = \sigma(\mathbf{u}W_1 + b_1)W_2 + b_2,\ \sigma = \text{ReLU}. \tag{9}$$

The sequence is reshaped to 3-D and a final $1\times1\times1$ convolution prepares the decoder input:

$$\mathbf{B} = W_o * \text{reshape}(\mathbf{h}_2). \tag{10}$$

2.4.2 Transformer bottleneck with residual

A learnable scalar gate α modulates the transformer contribution:

$$\mathbf{D} = \mathbf{E} + \alpha\,\mathbf{B}, \qquad \alpha \sim \mathcal{U}(0, 10^{-3}) \text{ at init.} \tag{11}$$

The idea follows DeepNet's residual-scaling principle [33] and retains an identity path [11]. Similar gating has proven effective in fully volumetric hybrids such as TransBTSv2 [31]. Only $\mathbf{D}$ is forwarded to the decoder, enabling the network to adaptively balance local convolutions and global self-attention.

Hybrid CNN–Transformer designs (e.g. TransUNet [3], UNETR [10], TransBTSv2 [31]) report state-of-the-art results across 3-D segmentation tasks. By restricting self-attention to the bottleneck, our variant preserves these benefits with minimal compute and readily complements anisotropic 2.5-D encoders [26].

2.4.3 Fusion Decoder

In an effort to increase tumor resolution, we sought means of combining data from both T1 and T2-weighted MRI channels. Applying pre-trained SWIN transformers to crops equidistant from the point of maximal tumor diameter in each dimension, we aimed to test how applying attention across modalities at the level of raw data would compare against nnUnet features. At the very least, we intend this alternate preference to increase robustness during the translation to clinical practice. Early-fusion of T1/T2 features with hierarchical 7x7x7 cropped tumor edges using pre-trained SWIN transformers yielded unstable training and a threshold at 50-75 In all likelihood, the poor performance of fusing these modalities early is a result of them depicting fundamentally distinct spatial entities. Generally speaking, The actual border between tumor and surrounding CSF is much higher in post-contrast T1, while brain anatomy, and

the displacement of the nerve show up best on T2. Being two separate views of the surrounding layer, their relative feature spaces span distinct portions of the image: concatenating the features from spatially contiguous T1 and T2 features is likely hampered by the distinct boundary between where these modalities draw tumor boundaries. Depending on the degree of cystic components in the tumor, the border will appear more or less "fuzzy," between post-contrast T1 and T2. In future work, we aims to test the performance of a fusion decoder with separate Vision Transformer encoder heads. The difference in contrast and gap between edges depicted in T1 and T2, we believe, constitute distinct feature spaces that should be learned separately before a joint decision is reached.

2.5. Training and Evaluation

For baseline, transformer bottleneck, and transformer bottleneck with residual, we follow the training and evaluation pipeline used by Shapey, J. *et al.* [26, 27] and Wang, G. *et al.* [32]. We train and evaluate our models on NVIDIA A40 and L40S GPUs, with an average training time of 8 hours per experiment. For the experiments including diffusion-generated data, we use Google Cloud Vertex AI via a custom Training Job. Eight n1-standard-8 worker pools, each equipped with one NVIDIA Tesla V100 GPU, execute a containerized train.py script for 100 epochs using DistributedDataParallel. Each experiment takes 4 hours to execute with distributed training and 48 hours without.

For the fusion decoder approach, we proceed as follows. We randomly partitioned the T1- and T2-weighted MRI datasets into 178 training, 20 validation, and 47 testing cases. Each scan was manually cropped using a cubic bounding box of dimensions 100 mm $\times$ 50 mm $\times$ 50 mm to isolate the region of interest. Intensity normalization was performed on each image using its own mean and standard deviation. All convolutional neural networks (CNNs) were implemented using `TensorFlow` and `NiftyNet`, and trained on a Linux compute node equipped with an NVIDIA GTX 1080 Ti GPU. Training was conducted using the Adam optimizer with a weight decay of 1×10^{-7}, a batch size of 2, and an initial learning rate of 1×10^{-4}, which was reduced by half every 10,000 iterations. Early stopping was employed based on stagnation in validation performance.

For quantitative evaluation, we report Dice similarity coefficients (DSC) on the held-out test set to assess segmentation accuracy across modalities.

3. Results

We report the volumetric Dice coefficient

$$\mathrm{Dice}(P, G) = \frac{2\,|P \cap G|}{|P| + |G|}, \tag{12}$$

averaged over the same test volumes used in earlier studies. Tab. 1 summarizes the mean scores for all variants.

Baselines. The historical U Net of Wang [32] attains a mean Dice of 0.61 on post-contrast T1, underscoring the progress achieved by the Shapey pipeline, which reaches 0.945 (T1) and 0.909 (T2).

Transformer bottleneck. Introducing a single self-attention block (Variant T) yields $\mu_{\mathrm{Dice}} = 0.920$ on T1 and 0.892 on T2, [-2.5 pp and -1.7 pp relative to Shapey]. Adding the learnable gate (Variant TR) recovers part of this gap, improving T2 to 0.906 and T1 to 0.926 (+0.6 pp over T on T1).

Effect of synthetic data. Augmenting the training set with diffusion-generated crops ("+" columns) increases the Shapey baseline from $0.945 \rightarrow 0.950$ on T1 and $0.909 \rightarrow 0.911$ on T2. Variant T benefits more markedly ($+1.5$ pp), whereas Variant TR shows only a marginal gain (+0.3 pp), suggesting that the residual gate already captures part of the small-lesion variance introduced by synthesis.

Early multi-modal fusion. The Swin-based early-fusion decoder (Sec. 2.4.3) achieves a mean Dice of 0.750 on co-registered T1/T2 inputs, substantially higher than chance but $\sim$ 20 pp below single-modality Shapey. The training instability observed (Sec. 2.4.3) is therefore reflected in the final metric.

While the Shapey baseline remains the top performer in native inputs, the gated transformer bottleneck narrows the gap in T2 (0.906 vs 0.909) and leverages synthetic augmentation without catastrophic overfitting. In contrast, naive early fusion of T1 and T2 volumes is suboptimal, motivating the late-fusion strategy planned for future work.

4. Conclusions

We presented an augmented U-Net architecture for vestibular schwannoma segmentation, combining a conditional diffusion model for synthetic training data with a lightweight transformer bottleneck. The aim was to mitigate two well-documented issues in automated segmentation systems: lack of small-tumor examples in training and insufficient global context in standard CNNs. Our experiments showed that training with diffusion-synthesized lesions provides limited improvements in sensitivity to small or shifted tumors, helping in certain—but not all—cases. Meanwhile, inserting a transformer bottleneck with residual gating offered only marginal gains over a strong baseline, particularly for T1 data; results for T2 were mixed

and, in some instances, remained below the existing benchmark. Although these findings suggest that diffusion-driven augmentation and transformer modules can bolster CNN-based architectures in specific scenarios, they do not universally resolve the variability observed in real-world MRI protocols. Future research could investigate more targeted multi-modal fusion strategies and explore larger, more diverse datasets, aiming for more robust and consistent generalization across different scanning conditions.

5. Contributions

Contributions: All members contributed equally. Tasks are divided into diffusion model, building on existing model, preprocessing, and other tasks.

References

[1] P Rajith Bhargav and Niladri B Puhan. Novel contraharmonic correlative attention loss for microaneurysm segmentation in fundus images. *IEEE Sensors Letters*, 7(7):1–4, 2023. 2

[2] Matthew L Carlson and Michael J Link. Vestibular schwannomas. *New England Journal of Medicine*, 384(14):1335–1348, 2021. 1

[3] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021. 6

[4] Jieneng Chen, Jieru Mei, Xianhang Li, Yongyi Lu, Qihang Yu, Qingyue Wei, Xiangde Luo, Yutong Xie, Ehsan Adeli, Yan Wang, et al. Transunet: Rethinking the u-net architecture design for medical image segmentation through the lens of transformers. *Medical Image Analysis*, 97:103280, 2024. 3

[5] Stephanie J Chiu, Simon J Hickman, Irene M Pepper, Jennifer HY Tan, John Yianni, and Joanna M Jefferis. Neuro-ophthalmic complications of vestibular schwannoma resection: Current perspectives. *Eye and Brain*, pages 241–253, 2021. 1

[6] Kenneth Clark, Bruce Vendt, Kirk Smith, John Freymann, Justin Kirby, Paul Koppel, Stephen Moore, Stanley Phillips, David Maffitt, Michael Pringle, et al. The cancer imaging archive (tcia): maintaining and operating a public information repository. *Journal of digital imaging*, 26:1045–1057, 2013. 3

[7] Zolnamar Dorjsembe, Hsing-Kuo Pao, Sodtavilan Odonchimed, and Furen Xiao. Conditional diffusion models for semantic 3d brain mri synthesis. *IEEE Journal of Biomedical and Health Informatics*, 2024. 3, 4, 5

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, and ... An image is worth 16×16 words: Transformers for image recognition at scale. *ICLR*, 2021. 6

[9] Evgin Goceri. Medical image data augmentation: techniques, comparisons and interpretations. *Artificial Intelligence Review*, 56(11):12561–12605, 2023. 1, 3

[10] A Hatamizadeh, D Yang, H Roth, and D Unetr Xu. Transformers for 3d medical image segmentation. arxiv 2021. *arXiv preprint arXiv:2103.10504*, 2021. 6

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 630–645. Springer, 2016. 6

[12] Qixin Hu, Junfei Xiao, Yixiong Chen, Shuwen Sun, Jie-Neng Chen, Alan Yuille, and Zongwei Zhou. Synthetic tumors make ai segment tumors better. *arXiv preprint arXiv:2210.14845*, 2022. 4

[13] Alex Ling Yu Hung, Haoxin Zheng, Kai Zhao, Xiaoxi Du, Kaifeng Pang, Qi Miao, Steven S Raman, Demetri Terzopoulos, and Kyunghyun Sung. Csam: A 2.5 d cross-slice attention module for anisotropic volumetric medical image segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5923–5932, 2024. 2

[14] Noemi Jester, Manwi Singh, Samantha Lorr, Steven M Tommasini, Daniel H Wiznia, and Frank D Buono. The development of an artificial intelligence auto-segmentation tool for 3d volumetric analysis of vestibular schwannomas. *Scientific Reports*, 15(1):5918, 2025. 1

[15] Chunhui Jiang, Yi Wang, Qingni Yuan, Pengju Qu, and Heng Li. A 3d medical image segmentation network based on gated attention blocks and dual-scale cross-attention mechanism. *Scientific Reports*, 15(1):6159, 2025. 3

[16] Lan Jiang, Ye Mao, Xiangfeng Wang, Xi Chen, and Chao Li. Cola-diff: Conditional latent diffusion model for multi-modal mri synthesis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 398–408. Springer, 2023. 2

[17] Yuxin Kang, Xuan Zhao, Yu Zhang, Hansheng Li, Guan Wang, Lei Cui, Yaqiong Xing, Jun Feng, and Lin Yang. Improving domain generalization performance for medical image segmentation via random feature augmentation. *Methods*, 218:149–157, 2023. 1

[18] Ron Kikinis and Steve Pieper. 3d slicer as a tool for interactive brain tumor segmentation. In *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 6982–6984. IEEE, 2011. 3

[19] Yuxiang Lai, Xiaoxi Chen, Angtian Wang, Alan Yuille, and Zongwei Zhou. From pixel to cancer: Cellular automata in computed tomography. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 36–46. Springer, 2024. 4

[20] Xiang Li, Chong Fu, Qun Wang, Wenchao Zhang, Chiu-Wing Sham, and Junxin Chen. Dmsa-unet: Dual multi-scale attention makes unet more strong for medical image segmentation. *Knowledge-Based Systems*, 299:112050, 2024. 3

[21] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1):654, 2024. 1

[22] Kerem Nernekli, Amit R Persad, Yusuke S Hori, Ulas Yener, Emrah Celtikci, Mustafa Caglar Sahin, Alperen Sozer, Batuhan Sozer, David J Park, and Steven D Chang.

Automatic segmentation of vestibular schwannomas: A systematic review. *World Neurosurgery*, 2024. 1

[23] Qiumei Pu, Zuoxin Xi, Shuai Yin, Zhe Zhao, and Lina Zhao. Advantages of transformer and its application for medical image segmentation: a survey. *BioMedical engineering online*, 23(1):14, 2024. 1

[24] Sina Radparvar. Vestibular schwannoma: evolution of diagnosis and treatment. *Egyptian Journal of Neurosurgery*, 40(1):5, 2025. 1

[25] Jo Schlemper, Ozan Oktay, Michiel Schaap, Mattias Heinrich, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. Attention gated networks: Learning to leverage salient regions in medical images. *Medical image analysis*, 53:197–207, 2019. 2

[26] Jonathan Shapey, Aaron Kujawa, Reuben Dorent, Guotai Wang, Alexis Dimitriadis, Diana Grishchuk, Ian Paddick, Neil Kitchen, Robert Bradford, Shakeel R Saeed, et al. Segmentation of vestibular schwannoma from mri, an open annotated dataset and baseline algorithm. *Scientific Data*, 8(1):286, 2021. 1, 2, 3, 5, 6, 7

[27] Jonathan Shapey, Guotai Wang, Reuben Dorent, Alexis Dimitriadis, Wenqi Li, Ian Paddick, Neil Kitchen, Sotirios Bisdas, Shakeel R Saeed, Sebastien Ourselin, et al. An artificial intelligence framework for automatic segmentation and volumetry of vestibular schwannomas from contrast-enhanced t1-weighted and high-resolution t2-weighted mri. *Journal of neurosurgery*, 134(1):171–179, 2019. 7

[28] Krish Suresh, Guibo Luo, Ryan A Bartholomew, Alyssa Brown, Amy F Juliano, Daniel J Lee, D Bradley Welling, Wenli Cai, and Matthew G Crowson. An external validation study for automated segmentation of vestibular schwannoma. *Otology & Neurotology*, 45(3):e193–e197, 2024. 1, 2, 4, 5

[29] Muhammad Usman Akbar, Måns Larsson, Ida Blystad, and Anders Eklund. Brain tumor segmentation using synthetic mr images-a comparison of gans and diffusion models. *Scientific Data*, 11(1):259, 2024. 2

[30] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 6

[31] Dongnan Wang, Geoffroy Ji, Lequan Yu, and ... Transbts v2: Towards better and more efficient volumetric medical image segmentation. *Medical Image Analysis*, 2022. 6

[32] Guotai Wang, Jonathan Shapey, Wenqi Li, Reuben Dorent, Alexis Dimitriadis, Sotirios Bisdas, Ian Paddick, Robert Bradford, Shaoting Zhang, Sébastien Ourselin, et al. Automatic segmentation of vestibular schwannoma from t2-weighted mri by deep spatial attention with hardness-weighted loss. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22*, pages 264–272. Springer, 2019. 3, 4, 5, 7

[33] Hongyu Wang, Shuming Ma, Li Dong, Shaohan Huang, Dongdong Zhang, and Furu Wei. Deepnet: Scaling transformers to 1,000 layers. arxiv e-prints, art. *arXiv preprint arXiv:2203.00555*, 2022. 6

[34] Hesheng Wang, Tanxia Qu, Kenneth Bernstein, David Barbee, and Douglas Kondziolka. Automatic segmentation of vestibular schwannomas from t1-weighted mri with a deep neural network. *Radiation Oncology*, 18(1):78, 2023. 1, 3

[35] Wenjian Yao, Jiajun Bai, Wei Liao, Yuheng Chen, Mengjuan Liu, and Yao Xie. From cnn to transformer: A review of medical image segmentation models. *Journal of Imaging Informatics in Medicine*, 37(4):1529–1547, 2024. 1

Appendix

A. Mask Generation

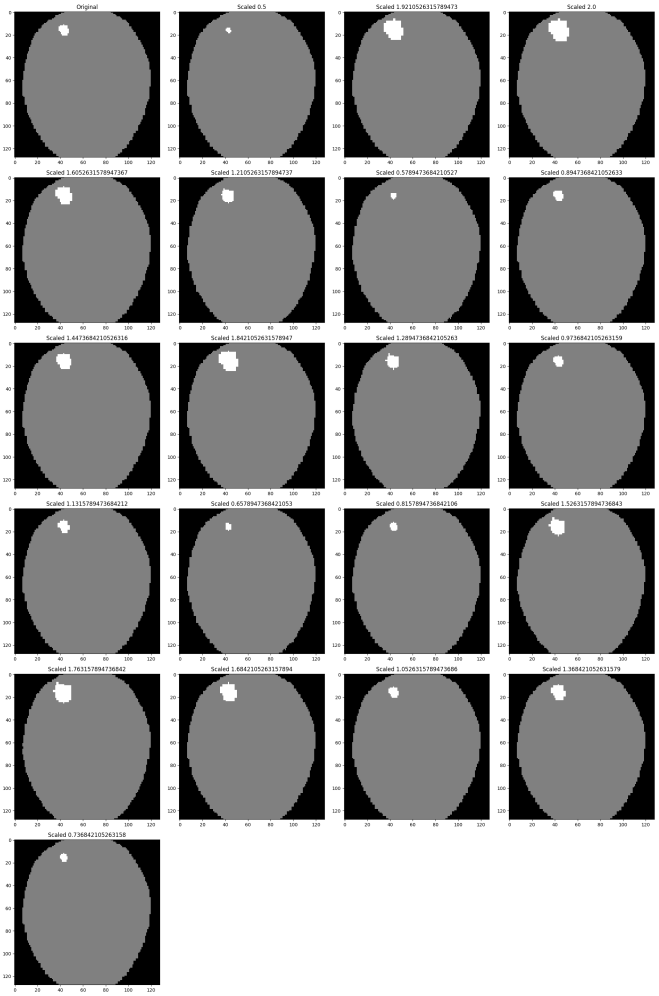

Figure 5. Largest tumor slice in original image for axial plane comparative plotting with all scaled images

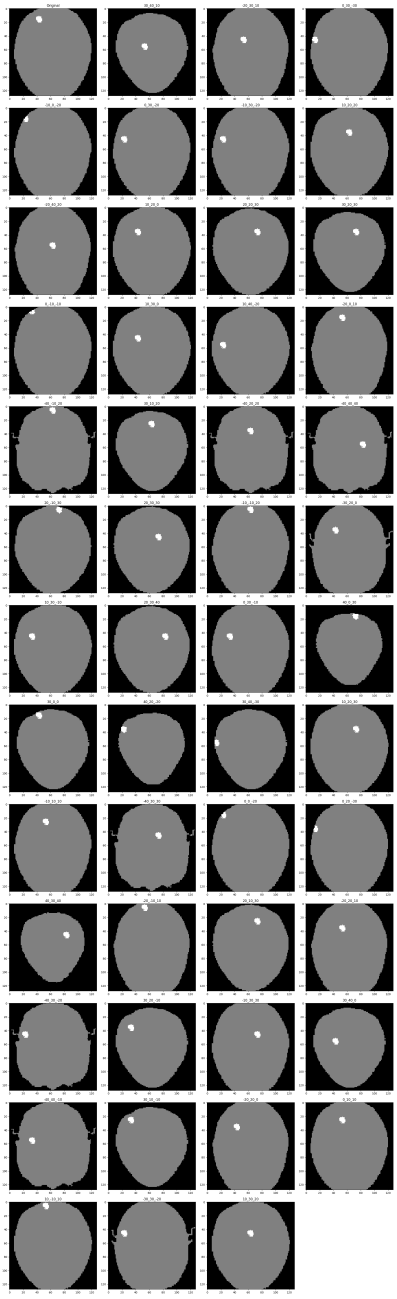

Figure 6. Largest tumor slice comparison between each shifted image and original image