Re-DiffiNet: Modeling discrepancies loss in tumor segmentation using diffusion models

Tianyi Ren[*1] TR1@UW.EDU
Abhishek Sharma[*1] AS711@UW.EDU
Juampablo Heras Rivera[1] JEHR@UW.EDU
Harshitha Rebala [2] LHREBALA@UW.EDU
Ethan Honey[1] EHONEY22@UW.EDU
Agamdeep Chopra[1] ACHOPRA4@UW.EDU
Jacob Ruzevick[3] RUZEVICK@NEUROSURGERY.WASHINGTON.EDU
Mehmet Kurt[1] MKURT@UW.EDU

[1] *Department of Mechanical Engineering, University of Washington*
[2] *Paul G. Allen School of Computer Science, University of Washington*
[3] *Department of Neurological Surgery, University of Washington*

Abstract

Identification of tumor margins is essential for surgical decision-making for glioblastoma patients and provides reliable assistance for neurosurgeons. Despite improvements in deep learning architectures for tumor segmentation over the years, creating a fully autonomous system suitable for clinical floors remains a formidable challenge because the model predictions have not yet reached the desired level of accuracy and generalizability for clinical applications. Generative modeling techniques have seen significant improvements in recent times. Specifically, Generative Adversarial Networks (GANs) and Denoising-diffusion-based models (DDPMs) have been used to generate higher-quality images with fewer artifacts and finer attributes. In this work, we introduce a framework called Re-Diffinet for modeling the discrepancy between the outputs of a segmentation model like U-Net and the ground truth, using DDPMs. By explicitly modeling the discrepancy, the results show an average improvement of 0.55% in the Dice score and 16.28% in HD95 from cross-validation over 5-folds, compared to the state-of-the-art U-Net segmentation model. The code is available: https://github.com/KurtLabUW/Re-DiffiNet.git.

Keywords: Tumor segmentation, DDPMs, MRI, Deep learning

1. Introduction

Background: Brain Tumor Segmentation

Glioblastoma is the most frequent primary malignant brain tumor in adults, representing approximately 57% of all gliomas and 48% of all primary malignant central nervous system (CNS) tumors (Ostrom et al., 2018; Tan et al., 2020). This heterogeneous group of tumors is characterized by their resemblance to glia that perform a variety of important functions including support to neurons (Isensee et al., 2021; Ahuja et al., 2020).

[*] Contributed equally

The treatment for glioma patients generally consists of surgery, radiotherapy, and chemotherapy and the outcomes of patients with gliomas vary widely according to the glioma type and prognostic factors. Due to the superior soft tissue contrast, multimodal MRI images which allow the complexity and the heterogeneity of the tumor lesion to be better visualized than a CT scan have become the golden standard for surgical decision-making for glioma patients (Hanif et al., 2017; Keunen et al., 2014; van Dijken et al., 2019). However, visual identification of tumor margins in CT or MRI still remains a challenge for neurosurgeons and researchers (Wang et al., 2019). Clinically, brain tumor masks are often obtained through Magnetic Resonance Imaging (MRI) scans, which require experienced radiologists to manually segment tumor sub-regions (Baid et al., 2021b). This is a long process that is unscalable to the needs of all patients. Thus, the recent growth of machine learning technologies holds promise to provide a reliable and automated solution to segmentation to save time and help medical professionals with this process (Luu and Park, 2021).

Deep learning techniques have been widely used in brain tumor segmentation. U-Net is the state of art for tumor segmentation. U-Net and its variants have been used in brain tumor segmentation. such as U-Net++ (Zhou et al., 2018), 3D U-Net (Çiçek et al., 2016), V-Net (Milletari et al., 2016), and Attention-U-Net (Oktay et al., 2018). Transformer architectures has also been applied in brain tumor segmentation. TransU-Net and Swin-U-Net show potential to predict accurate tumor margins. However, the state-of-the-art models in brain tumor segmentation are still based on the encoder-decoder architectures such as U-Net (Isensee et al., 2021) and its variations. For instance, Luu et. al (Luu and Park, 2021) modified the nnU-Net model by adding an axial attention in the decoder. Futrega et. al (Futrega et al., 2021) optimized the U-Net model by adding foreground voxels to the input data, increasing the encoder depth and convolutional filters. Siddiquee et. al (Siddiquee and Myronenko, 2021) applied adaptive ensembling to minimize redundancy under perturbations.

While U-Net-based architecture have led to significant improvements in region-based metrics for tumor segmentation e.g. Dice scores, it is also important to improve boundary-distance metrics like HD scores (Karimi and Salcudean, 2019; Yeghiazaryan and Voiculescu, 2018). Being able to locate boundaries of tumors is crucial for surgical planning. Thus, modeling techniques that are able to capture finer details and high frequency information at the boundaries, are desirable. One of the critical factors that makes predicting tumor boundaries difficult, is the inherent variability in tumor attributes at the boundaries. Thus, the modeling techniques also need to be able to capture the variability in tumor shapes.

Generative modeling techniques have seen great improvements in recent times. Specifically, Generative Adversarial Networks and Denoising-Diffusion based models have been used to generate desired images of greater quality. While GANs are able to generate images of high fidelity, they are also prone to mode collapse. Thus, they often fail to capture the variability of the data they seek to model. On the other hand, Denoising-Diffusion based models have been shown to be good at both mode coverage i.e. capturing the variability in the data (Kingma et al., 2021), as well as at generating high quality images (Dhariwal and Nichol, 2021). There have been very few instances of Diffusion models being used for brain tumor segmentation, that have shown promising results. (Xing et al., 2023; Wolleb et al., 2022; Wu et al., 2022).

In this work, we introduce a framework called Re-Diffinet, for modeling discrepancy between the outputs of a segmentation model like U-Net and the ground truth, using the advantages of Denoising Diffusion Probabilistic Models and U-Net model. By explicitly modeling the discrepancy, we intend to build upon previous segmentation models, force diffusion models to focus explicitly on the regions that other models miss, and exploit diffusion models' ability to capture finer details and variability in the data.

2. Methods

2.1. Dataset

2.1.1. BraTS2023

The training dataset provided for the BraTS23 challenge (Baid et al., 2021a) consists of 1251 brain MRI scans along with segmentation annotations of tumorous regions. The 3D volumes were skull-stripped and resampled to 1 mm^3 isotropic resolution, with dimensions of (240, 240, 155) voxels. For each example, four modalities were given: native (T1), post-contrast T1-weighted (T1Gd), T2-weighted (T2), and T2 Fluid Attenuated Inversion Recovery (T2-FLAIR). Segmentation labels were annotated manually by one to four experts. Annotations consist of three disjoint classes: enhancing tumor (ET), peritumoral edematous tissue (ED), and necrotic tumor core (NCR). To get the ground truth labels for these datasets, all imaging volumes have then been segmented using the STAPLE (Warfield et al., 2004) fusion of previous top-ranked BraTS algorithms, such as nnU-Net (Isensee et al., 2021). These segmented labels were then refined manually by volunteer neuroradiology experts following a consistently communicated annotation protocol. The manually refined annotations were finally approved by experienced board-certified attending neuro-radiologists.

2.1.2. Data Reprocessing

For all the MRI contrasts in the BraTS2023 dataset, we rescale the voxel intensity after Z-Score normalization as the preprocessing protocol.

2.1.3. Evaluation Metrics

The model will be evaluated on two metrics: Dice similarity coefficient (DICE) measures the similarity between the model prediction and the ground truth; Hausdorff distance (95%) measures the boundary distance between the model prediction and the ground truth.

2.2. Model Architectures

2.2.1. Baseline U-Net

We adapted the optimized U-Net (Futrega et al., 2021) as our baseline model architecture (Ren et al., 2024) for comparison purposes. U-Net has a symmetric U-shape that characterizes architecture and can be divided into two parts, i.e., encoder and decoder. The encoder comprises 5 levels of same-resolution convolutional layers with strided convolution downsampling. The decoder follows the same structure with transpose convolution upsampling and convolution operating on concatenated skip features from the encoder branch at the same level. The training dataset is comprised of the pairs $\{(I, x_0)\}_{i=1}^{N}$, where $I \in \mathbb{R}^{4\times D\times W\times H}$

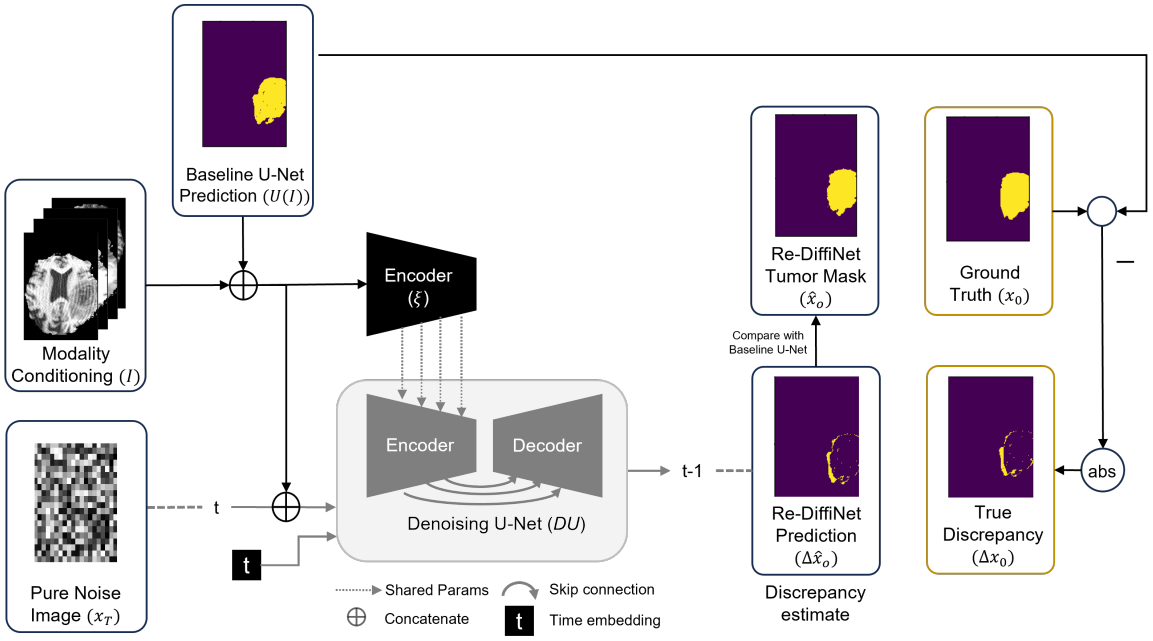

Figure 1: Re-DiffiNet uses MRI and predictions from baseline U-Net as inputs to generate predictions about incorrect voxels in U-Net predictions and corrects those voxels to generate redefined tumor masks.

represents the four 3D-MRI contrast as multi-channel input, $x_0 \in \mathbb{R}^{3\times D\times W\times H}$ represents the associated one-hot encoded segmentation mask, with 3 tumor labels: 1) Whole Tumor, 2) Enhancing Tumor, and 3) Necrotic Tumor Core. The baseline U-Net predicts the tumor labels $\hat{x}_0$ given the input I:

$$\hat{x}_0 = U(I) \tag{1}$$

2.2.2. U-Net augmented Diffusion (UA-Diffusion)

This model builds upon the Diff-U-Net model proposed by (Xing et al., 2023). Diff-U-Net conditions DDPMs with MRIs at each denoising step to generate the corresponding tumor masks. In comparison, we condition our diffusion model with predictions ($U(I)$) from baseline U-Net along with MRIs (I) as shown in figure 1. We tested 3 variants of this approach (3 different inputs) as explained in section 3. Among the 3 variants, we chose the best performing UA-Diffusion approach (Table 2.2.3) and used it for the remaining experiments. The expression for the best performing variant of UA-Diffusion is shown in equation 2:

$$\hat{x}_0 = DU(\text{cat}(U(I), I, x_t), t, \hat{I}_f) \tag{2}$$

where t is the time embedding, x_t is the corresponding noise masks, $\hat{I}_f = \xi(\text{cat}(U(I), I))$ are the multi-scale features extracted using a trainable copy (ξ) of the encoder of the denoising-

Table 1: Comparison of the proposed model architecture in section 2.2.

Model	Dice				HD95(mm)			
	WT	ET	TC	Avg	WT	ET	TC	Avg
Baseline U-Net	92.63%	**86.87%**	93.28%	90.93%	1.06	1.62	1.57	1.42
Diff-U-Net	87.98%	83.92%	86.25%	86.05%	2.46	3.56	3.32	3.11
UA-Diffusion (Input: U(I))	90.72%	83.76%	86.57%	87.02%	1.12	1.90	2.56	1.86
UA-Diffusion (Input: concat(I, U(I))	92.86%	85.08%	91.43%	89.79%	1.39	1.93	1.59	1.63
UA-Diffusion (Input: mask($I, U(I)$)) see eq.4	91.32%	84.46%	91.18%	88.99%	1.46	1.74	1.89	1.70
Re-DiffiNet	**93.23%**	86.79%	**93.98%**	**91.33%**	**0.87**	**1.27**	**1.34**	**1.16**

U-Net (DU). These multi-scale features are added to the outputs of the corresponding layers in the denoising-U-Net, DU (see Figure 1).

2.2.3. Re-DiffiNet

Our proposed Re-DiffiNet model architecture uses the same architecture as the U-Net augmented Diffusion (UA-Diffusion). However, instead of trying to generate ground truth segmentation masks x_0, we generate the absolute discrepancy between the ground truth segmentation masks and baseline U-Net's predictions i.e. $\Delta x_0 = abs(U(I) - x_0)$ (Figure 1). These discrepancy masks (Δx_0) will have a value of 1 for each voxel that is predicted incorrectly by the baseline U-Net (U), and 0 for voxels where the predictions are correct. Once, we have generated the estimated discrepancies $\Delta\hat{x}_0$, the tumor mask predictions can be obtained as shown in equation 3 :

$$\Delta\hat{x}_0 = DU(\text{cat}(U(I), I, x_t), t, \hat{I}_f) \Rightarrow \hat{x}_0 = abs(U(I) - \Delta\hat{x}_0) \tag{3}$$

Subtracting the estimated discrepancy $\Delta\hat{x}_0$ from $U(I)$ and taking the absolute, we flip every incorrect voxel (as per our estimate) in $U(I)$ i.e. $1 \rightarrow 0$ and $0 \rightarrow 1$. While, the correct voxels (as per our estimation) in the baseline U-Net, remain the same.

2.3. Training details

Our models were implemented in Pytorch and MONAI, and trained on 2 NVIDIA A40 GPUs. The model was trained on overlapping regions, whole tumor (WT), tumor core (TC), and enhancing tumor(ET). TC entails the ET, as well as the necrotic (NCR) parts of the tumor, and WT describes the complete extent of the disease. The diffusion models was trained using a compound loss function including DICE loss, Binary cross entropy (BCE) loss, and Mean square error(MSE) loss. The model was trained using the AdamW optimizer with a learning rate of 0.0001 and a weight decay equal to 0.0001. The network's performance was evaluated using 5-fold cross-validation. The data were randomly shuffled and equally split into 5 groups for cross-validation.

Table 2: 5 fold cross-validation results for 3 models: Baseline U-Net, U-Net augmented Diffusion (UA-Diffusion) with concatenation of MRI and Baseline U-Net predictions as input, and Re-DiffiNet

Fold #	Model	Dice				HD95(mm)			
		WT	ET	TC	Avg	WT	ET	TC	Avg
fold1	Baseline U-Net	92.63%	**86.87%**	93.28%	90.93%	1.06	1.62	1.57	1.42
	UA-Diffusion	92.72%	86.76%	93.57%	91.02%	1.12	1.40	1.56	1.36
	Re-DiffiNet	**93.23%**	86.79%	**93.98%**	**91.33%**	**0.87**	**1.27**	**1.34**	**1.16**
fold2	Baseline U-Net	92.60%	**88.30%**	93.79%	91.56%	1.18	1.77	1.24	1.40
	UA-Diffusion	92.62%	87.86%	94.09%	91.52%	1.19	1.73	1.18	1.37
	Re-DiffiNet	**93.04%**	87.34%	**94.48%**	**91.62%**	**0.97**	**1.67**	**0.85**	**1.16**
fold3	Baseline U-Net	92.40%	87.04%	92.47%	90.64%	1.41	1.78	1.46	1.55
	UA-Diffusion	**92.93%**	87.22%	**93.21%**	**91.12%**	**1.05**	1.62	**1.14**	**1.27**
	Re-DiffiNet	92.86%	**87.23%**	93.11%	91.07%	1.07	**1.60**	1.21	1.29
fold4	Baseline U-Net	91.21%	86.90%	92.66%	89.06%	1.62	1.74	1.30	1.55
	UA-Diffusion	91.32%	86.25%	92.99%	88.79%	1.61	1.73	1.26	1.53
	Re-DiffiNet	**91.73%**	**87.18%**	**92.91%**	**89.46%**	**1.58**	**1.64**	**1.21**	**1.48**
fold5	Baseline U-Net	91.30%	86.61%	93.25%	90.39%	1.34	1.72	1.16	1.41
	UA-Diffusion	91.43%	87.01%	93.56%	90.67%	1.30	1.68	1.18	1.39
	Re-DiffiNet	**92.72%**	**87.81%**	**94.30%**	**91.61%**	**1.15**	**1.37**	**1.06**	**1.04**

3. Experiment and Results

We trained 3 models 1) Baseline U-Net (section 2.2.1), 2) U-Net augmented diffusion or UA-Diffusion (section 2.2.2), and 3) *Re-DiffiNet* (section 2.2.3). We first trained the baseline U-Net model. Then, the predictions of the baseline U-Net model were used as inputs in U-Net augmented diffusion (UA-Diffusion), and Re-DiffiNet. In a preliminary study, we trained 3 variants of the U-Net augmented diffusion (UA-Diffusion) on a random train-test split and compared them with the baseline-U-Net: 1) Conditioning the diffusion model with only the U-Net output U(I)U(I), 2) Conditioning the diffusion model with a concatenation of MRI contrasts and baseline U-Net predictions U(I)U(I), 3) conditioning the diffusion model with MRIs II masked by U-Net predictions U(I)U(I). A mask which has a value 1 for each tumor voxel and 0.2 for non-tumor voxel is applied to each of the 4 MRI contrasts, which are concatenated and used as inputs. The resulting masked input is represented as shown in equation 4 :

$$M'(x,y,z) = \begin{cases} 1 & \text{if } U(I)[x,y,z] > 0, \\ 0.2 & \text{if } U(I)[x,y,z] = 0, \text{where x,y,x are voxel indices} \end{cases} \tag{4}$$

$$mask(I, U(I)) = concat(I_i \circ M' | i = 1,2,3,4), \text{where i denotes an MRI contrast}$$

We found that using diffusion directly to predict tumor masks doesn't lead to any significant performance gains over the baseline U-Net, as shown in Table 2.2.3. On the other hand, using diffusion model to predict discrepancies and using them to correct U-Net's outputs leads to significant performance gains specially in terms of HD95 score. Among the 3 UA-Diffusion approaches concatenating the U-Net prediction and MRI yielded the

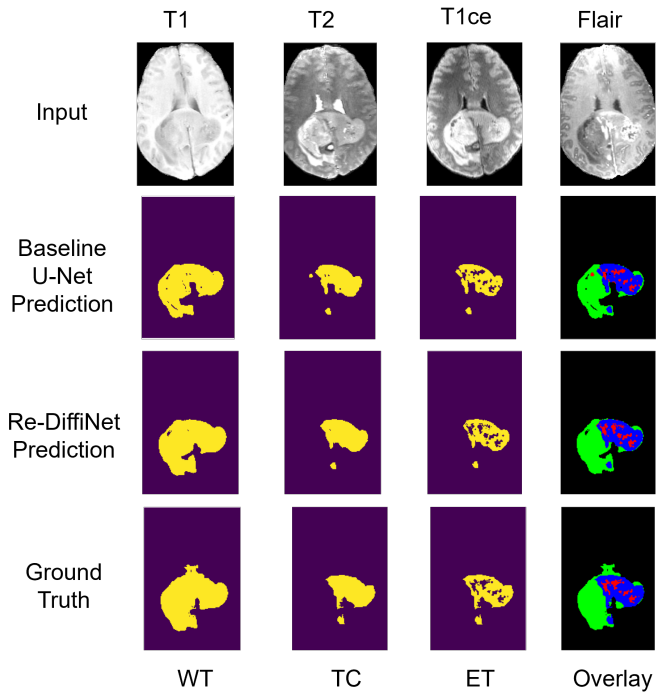

Figure 2: A comparison between the segmentations generated by baseline U-Net, and Re-DiffiNet. In this example, Re-DiffiNet can predict the false positive lesion on Tumor core masks that was predicted by baseline U-Net. Meanwhile, Re-DiffiNet predicts a smoother boundary

best performance. Thus, we use the UA-Diffusion with concatenation of MRI and U-Net prediction, for 5-fold cross-validation in Table 3.

The results of 5-fold cross-validation are shown in Table 3, which reports the Dice Score (DICE) and 95 percentile Hausdorff distance (HD95) and the average scores of all methods on the three overlapping regions whole tumor (WT), tumor core (TC) and Enhancing tumor (ET) for the BraTS2023 dataset (Figure 2). We found that while using the diffusion model to directly output the tumor segmentation mask does lead to improvements over the U-Net model, the improvements are modest: 0.12% improvement in Dice, and 5.61% improvement in HD95 score. On the other hand, with Re-DiffiNet we found a 16.28% improvement in HD95 score, indicating the benefits of modeling discrepancy using diffusion models, while simultaneously the Dice score was comparable with the baseline U-Net (0.55% improvement). Figure 2 shows an example of the segmented masks of baseline U-Net and Re-DiffiNet.

4. Discussion and Conclusion

In this research, we proposed a tumor segmentation framework Re-DiffiNet, which uses diffusion models to refine and improve predictions of a tumor segmentation model (like optimized U-Net). Most tumor segmentation studies optimize for region-based metrics like Dice scores, and have been able to show high Dice score in the range of 90% or greater. However, boundary-distance metrics like HD scores are also critical, and being able to improve upon these score while not sacrificing performance on Dice score is highly desirable. In this work, we investigated the potential to refine predictions generated by state-of-the-art U-Net models using diffusion models.

We found that while using diffusion models to directly generate tumor masks did lead to improvements in performance over the baseline U-Net, it was the use of discrepancy modeling i.e. predicting the differences between ground truth masks and baseline U-Net's outputs, that led to most significant improvements. This was indicated by 16.28% improvement in HD-95 score, highlighting significant improvements on the boundaries of tumors. While, discrepancies can be modeled by any other modeling technique (even a U-Net), effectively acting as a boosting method, we chose diffusion primarily because of its ability to generate high-fidelity visual attributes, as well as capture variability in the data distribution, both of which are exhibited by brain tumors. Another benefit of using diffusion models instead of a U-Net to improve the baseline U-Net's predictions, would be the potential to learn more robust and diverse representations from the data, due to the inherently different mechanism using which diffusion models are trained.

Our work shows the potential of further improving tumor segmentation by combining diffusion models and discrepancy modeling. In this work, we investigated Re-DiffiNet for the segmentation of gliomas. In the future, we intend to test our approach to improve the segmentation of other kinds of tumors like meningioma, and pediatric brain tumors.

Acknowledgments

Juampablo Heras Rivera is supported by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research, Department of Energy Computational Science Graduate Fellowship under Award Number DE-SC0024386.

References

Sakshi Ahuja, B.K. Panigrahi, and Tapan Gandhi. Transfer learning based brain tumor detection and segmentation using superpixel technique. In *2020 International Conference on Contemporary Computing and Applications (IC3A)*, pages 244–249, 2020. doi: 10.1109/IC3A48958.2020.233306.

Ujjwal Baid, Satyam Ghodasara, Suyash Mohan, Michel Bilello, Evan Calabrese, Errol Colak, Keyvan Farahani, Jayashree Kalpathy-Cramer, Felipe C Kitamura, Sarthak Pati, et al. The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. *arXiv preprint arXiv:2107.02314*, 2021a.

Ujjwal Baid, Satyam Ghodasara, Suyash Mohan, Michel Bilello, Evan Calabrese, Errol Colak, Keyvan Farahani, Jayashree Kalpathy-Cramer, Felipe C Kitamura, Sarthak Pati,

et al. The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. *arXiv preprint arXiv:2107.02314*, 2021b.

Özgün Çiçek, Ahmed Abdulkadir, Soeren S. Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: Learning dense volumetric segmentation from sparse annotation. *CoRR*, abs/1606.06650, 2016. URL http://arxiv.org/abs/1606.06650.

Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis. *CoRR*, abs/2105.05233, 2021. URL https://arxiv.org/abs/2105.05233.

Michał Futrega, Alexandre Milesi, Michał Marcinkiewicz, and Pablo Ribalta. Optimized u-net for brain tumor segmentation. In *International MICCAI Brainlesion Workshop*, pages 15–29. Springer, 2021.

Farina Hanif, Kanza Muzaffar, Kahkashan Perveen, Saima M Malhi, and Shabana U Simjee. Glioblastoma multiforme: a review of its epidemiology and pathogenesis through clinical presentation and treatment. *Asian Pacific journal of cancer prevention: APJCP*, 18(1): 3, 2017.

Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021.

Davood Karimi and Septimiu E Salcudean. Reducing the hausdorff distance in medical image segmentation with convolutional neural networks. *IEEE Transactions on medical imaging*, 39(2):499–513, 2019.

Olivier Keunen, Torfinn Taxt, Renate Grüner, Morten Lund-Johansen, Joerg-Christian Tonn, Tina Pavlin, Rolf Bjerkvig, Simone P Niclou, and Frits Thorsen. Multimodal imaging of gliomas in the context of evolving cellular and molecular therapies. *Advanced drug delivery reviews*, 76:98–115, 2014.

Diederik P. Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. *CoRR*, abs/2107.00630, 2021. URL https://arxiv.org/abs/2107.00630.

Huan Minh Luu and Sung-Hong Park. Extending nn-unet for brain tumor segmentation. In *International MICCAI Brainlesion Workshop*, pages 173–186. Springer, 2021.

Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. *CoRR*, abs/1606.04797, 2016. URL http://arxiv.org/abs/1606.04797.

Ozan Oktay, Jo Schlemper, Loïc Le Folgoc, Matthew C. H. Lee, Mattias P. Heinrich, Kazunari Misawa, Kensaku Mori, Steven G. McDonagh, Nils Y. Hammerla, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. Attention u-net: Learning where to look for the pancreas. *CoRR*, abs/1804.03999, 2018. URL http://arxiv.org/abs/1804.03999.

Quinn T Ostrom, Haley Gittleman, Gabrielle Truitt, Alexander Boscia, Carol Kruchko, and Jill S Barnholtz-Sloan. Cbtrus statistical report: primary brain and other central

nervous system tumors diagnosed in the united states in 2011–2015. *Neuro-oncology*, 20 (suppl_4):iv1–iv86, 2018.

Tianyi Ren, Ethan Honey, Harshitha Rebala, Abhishek Sharma, Agamdeep Chopra, and Mehmet Kurt. An optimization framework for processing and transfer learning for the brain tumor segmentation. *arXiv preprint arXiv:2402.07008*, 2024.

Md Mahfuzur Rahman Siddiquee and Andriy Myronenko. Redundancy reduction in semantic segmentation of 3d brain tumor mris. *arXiv preprint arXiv:2111.00742*, 2021.

Aaron C Tan, David M Ashley, Giselle Y López, Michael Malinzak, Henry S Friedman, and Mustafa Khasraw. Management of glioblastoma: State of the art and future directions. *CA: a cancer journal for clinicians*, 70(4):299–312, 2020.

Bart RJ van Dijken, Peter Jan van Laar, Marion Smits, Jan Willem Dankbaar, Roelien H Enting, and Anouk van der Hoorn. Perfusion mri in treatment evaluation of glioblastomas: Clinical relevance of current and future techniques. *Journal of Magnetic Resonance Imaging*, 49(1):11–22, 2019.

Lei Wang, Buqing Liang, Yan Icy Li, Xiang Liu, Jason Huang, and Yan Michael Li. What is the advance of extent of resection in glioblastoma surgical treatment—a systematic review. *Chinese neurosurgical journal*, 5(1):1–6, 2019.

Simon K Warfield, Kelly H Zou, and William M Wells. Simultaneous truth and performance level estimation (staple): an algorithm for the validation of image segmentation. *IEEE transactions on medical imaging*, 23(7):903–921, 2004.

Julia Wolleb, Robin Sandkühler, Florentin Bieder, Philippe Valmaggia, and Philippe C Cattin. Diffusion models for implicit image segmentation ensembles. In *International Conference on Medical Imaging with Deep Learning*, pages 1336–1348. PMLR, 2022.

Junde Wu, Huihui Fang, Yu Zhang, Yehui Yang, and Yanwu Xu. Medsegdiff: Medical image segmentation with diffusion probabilistic model. *arXiv preprint arXiv:2211.00611*, 2022.

Zhaohu Xing, Liang Wan, Huazhu Fu, Guang Yang, and Lei Zhu. Diff-unet: A diffusion embedded network for volumetric segmentation. *arXiv preprint arXiv:2303.10326*, 2023.

Varduhi Yeghiazaryan and Irina Voiculescu. Family of boundary overlap metrics for the evaluation of medical image segmentation. *Journal of Medical Imaging*, 5(1):015006–015006, 2018.

Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. *CoRR*, abs/1807.10165, 2018. URL http://arxiv.org/abs/1807.10165.