Tumor Segmentation Using Diffusion & GAN + Transformer

April 2023

Yaxuan Hou(yaxuan@umich.edu), Saurav Telge(sauravt@umich.edu)
Bolin Wu(bolinw@umich.edu), Huiyao Yang(huiyao@umich.edu)
Jiaming Yao(yaojm@umich.edu)

Abstract

This paper proposes a comparison study between two generative models, GAN + Transformer and Diffusion Model, for bladder tumor whole slide image segmentation. Tumor segmentation on whole slide images is challenging and time-consuming for pathologists, and obtaining labelled data for medical image segmentation tasks is difficult. The proposed models aim to overcome these limitations and provide accurate segmentation results. The dataset used in this study consists of 397 patient-exclusive slides for non-invasive low-grade papillary urothelial carcinoma and non-invasive or invasive high-grade papillary urothelial carcinoma. The GAN + Transformer model uses a generator that generates tumor segmentation using Transformer, while the discriminator is trained using CNN based on maximizing L1 Loss. The diffusion model, on the other hand, is trained to estimate the probability distribution of the segmentation masks through the reverse diffusion process using a modified UNet conditioned on the raw image. The models' performances are evaluated using the DICE score, and the results are compared with state-of-the-art methods. The proposed models have the potential to provide a more efficient and accurate method for bladder tumor whole slide image segmentation.

1 Introduction

Pathologists use WSIs for cancer detection and produce one of the gold standards in the area. Whole slide images (WSIs), also known as virtual slides or digital slides, are high-resolution digital images of entire microscope slides. These images are created by scanning the entire glass slide using a high-resolution scanner, and then storing the resulting image as a large digital file. However, tumor segmentation by human hand on WSIs is extremely exhausting and posted a high requirement of professional knowledge on the executor. It is also a common challenge with ML models on medical image segmentation tasks that the labelled data is hard to obtain. Recent generative models have proven their excellence in overcoming this limitation. Hence, we would like to evaluate two generative models on this bladder tumour WSI image segmentation task and assess their robustness when the availability of training data is further limited.

2 Revelant Background

2.1 Baseline

Z. Zhang et al., 2019 presented a pathology whole-slide diagnosis method, powered by neural networks, that detects tumor regions given a whole slide image and makes diagnostic suggestions. The method framework can be divided into 3 parts, in which the first part, detecting tumor regions using the scanner network (S-Net), is related to our topic. We will use the S-Net model as our baseline model, we were able to obtain a portion of their dataset as our training dataset. They first construct their I-Slide dataset, composed of 913 haematoxylins and eosin (H&E) stained whole slides (with an average slide height $\times$ width of 80,386 $\times$ 59,143 pixels) from patients with bladder cancer, obtained from multiple medical sources. Each of these 913 slides is annotated through a strict diagnosis label verification process. One board-certified pathologist and one pathology-trained doctor are asked to annotate at most eight tumor regions, which they believed to contain diagnostically useful information and eight non-tumor regions. Note: the slides are partially annotated due to the great workload. They then construct their II-Image dataset by randomly sampling a set of images with 1,024 $\times$ 1,024 resolution (pathologists for cellular feature analysis suggest this resolution) around the annotated tumor and non-tumor regions. Each tissue image had a tumor region binary mask. S-net resembles the commonly used U-net. The s-net conducts tumor detection by classifying each pixel as tumor or non-tumor, represented by a probability. To bypass the partial annotation problem, they measure their model's performance by only computing the loss on annotated pixels, which we will adopt when evaluating our models.

2.2 GAN + Transformer

Transformer-based frameworks have reached state-of-the-art performance on various computer vision tasks. Wang et al., 2021 proposed a method by combining CNN and transformer to perform tumor segmentation.

The main structure of the method is an encoder and a decoder. The encoder first utilizes 3D CNN to extract the spatial feature maps. The decoder leverages the features embedded by Transformer and performs progressive upsampling to predict the detailed segmentation map. Our paper uses a similar transformer structure in the generators' transformer part.

The GAN model is also popular in tumor segmentation. A novel end-to-end adversarial neural network, called SegAN, was proposed by Xue et al., 2017 . They use a fully convolutional neural network as the generator to generate segmentation label maps and propose a novel critique with a multi-scale L1 loss function to force the discriminator and generator to learn both global and local features that capture long- and short-range spatial relationships between pixels. Our paper adopts the main structure of the generator's strided convolution and deconvolution parts and discriminator.

Huang et al., 2022 demonstrated how to use GAN + Transformer model to do tumor segmentation in their paper. The main idea of their method is to combine a generator and a discriminator, which are trained in a min-max game. The generator consists of a 3D CNN-based down-sampling encoder, encoding transformer, and up-sampling decoder. The discriminator is based on CNN, and they used L1-norm distance as a loss function to measure the difference between prediction and the ground truth. The difference between our model and this model is the difference in the loss function of the discriminator and hyperparameters in the whole structure.

2.3 Diffusion Probabilistic Model

The paper by Wu et al., 2022 proposes the first diffusion probabilistic model (DPM)-based model for medical segmentation tasks, which the authors name MedSegDiff, and verify that it outperforms the state-of-the-art (SOTA) methods in three tasks with different image modalities.

The implementation follows standard conditional DPM architecture, as well as the learnable parameters, where a UNet is used for segmentation. In each step of the reverse diffusion step, the current segmentation map is taken as signal input, and the raw image is taken as the conditional prior. Then they are summed and sent to the UNet decoder for reconstruction. One of the main contributions of this work is Dynamic Conditional Encoding. To improve segmentation accuracy, MedSegDiff integrates the raw image and the current-step segmentation map in each step of the encoders by an attentive-like mechanism and uses the FF-Parser, a learnable frequency filter, to constrain the high-frequency component.

The authors conduct the experiments on three tasks, which are optic-cup segmentation from fundus images, brain tumor segmentation from MRI images, and thyroid nodule segmentation from ultrasound images. The results show that MedSegDiff achieves higher accuracy in all tasks than SOTA segmentation methods.

The same authors later published the paper "MedSegDiff-V2: Diffusion-based Medical Image Segmentation with Transformer" Wu et al., 2023. In this work, the authors integrate a transformer into the DPM-based model for medical image segmentation, which outperforms MedSegDiff version one.Additionally, the authors propose a new transformer architecture, SS-Former, used to learn the interaction between the segmentation noise and semantic feature. The anchor condition is also introduced to reduce the large input variance caused by transformer blocks. Their research shows that the model outperforms other state-of-the-art (SOTA) segmentation methods on multi-organ segmentation tasks over the AMOS dataset.

Meanwhile, in the field of denoising diffusion models, conventional approaches involve gradually adding Gaussian noise until it is possible to start the reverse diffusion process at isotropic normal distribution. However, A new approach Zheng et al., 2022 uses truncated diffusion probabilistic models (Truncated DPM) where the diffusion chain is cut down at time T_{trunc} in order to accelerate the diffusion process. As a result of truncation, there is a large gap between the Gaussian prior $p(x_T)$ and the truncated point $q(x_T|x_0)$, which can be bridged with an implicit generative distribution $p_\psi(x_T) = \int p_\psi(x_T|z)p(z)dz$, as Figure 3 illustrates.

In this research project, we explored implementing the diffusion-based model for segmentation tasks. It's important to note that our task is different from the previous works we referenced, which focused on organ image segmentation, while we focused on WSI segmentation. Given the complexity of cells and tissues in WSI images, we may need to seek different methods for conditioning and discrimination if the segmentation results are not ideal. Overall, our work provides valuable insights into implementing diffusion-based models for WSI segmentation tasks.

3 Proposed method

Dataset We used a subset of the dataset collected by Z. Zhang et al., 2019. It consists of whole slide images (WSI), which are digital representations of entire tissue samples scanned using high-resolution digital scanners. In medical imaging, whole slide images are commonly used in pathology to analyze tissue samples for diagnosis and research purposes. This WSI dataset is the largest annotated dataset we could find. We were able to obtain 397 patient-exclusive slides. Along with

the whole slide images, there is also the binary mask, which serves as the ground truth of the tumor segmentation. We followed their proposed pipeline to randomly sample images with 1,024 × 1,024 resolution around the annotated tumor and non-tumor regions of the whole slide images and their corresponding binary mask with the python library OpenSlide, and then we downsample the resolution of the images to 256 × 256 due to limitation of our computing resource. These sampled images will serve as our training dataset. Table 1 shows the composition of our dataset. Figure 7 is sample of training data.

For the sake of better debugging experience and visualization of the models' performance and also due to our restricting time and knowledge, we decided to simplify the learning objective of our models by merging the non-tumor region and unlabeled region by assigning them both to the negative class whereas the tumor region will remain as the positive class. We fully acknowledge the loss of information under current scheme, and we will seek for better encoding methods and loss functions in future iterations.

3.1 Models

3.1.1 Baseline

The S-Net shares similar structure with the well U-Net model, but with additional layers. One thing to note about the S-Net model is that when calculating the Binary Cross Entropy Loss (as mentioned earlier, we merged the unlabeled region with the non-tumor regions and hence we are using binary loss), we assign pixels different weights so that the unlabeled pixels will have a contribution of $\frac{1}{2}(1 - y_{ij})log(1 - \hat{y_{ij}})$, where y_{ij} is the groudtruth which is 0, and $\hat{y_{ij}}$ is the predicted value, to the aggregated loss over the image, i.e. half of a negative pixel's contribution to the aggregated loss. The intuition is that we do not want to fully discount the information contained in the unlabeled region during the training phase.

3.1.2 GAN + Transformer

The GAN + Transformer model combines Generative Adversarial Network (GAN) with the Transformer model. The overview of the proposed method is in Figure 1, which consists of a generator and a discriminator for competing training.

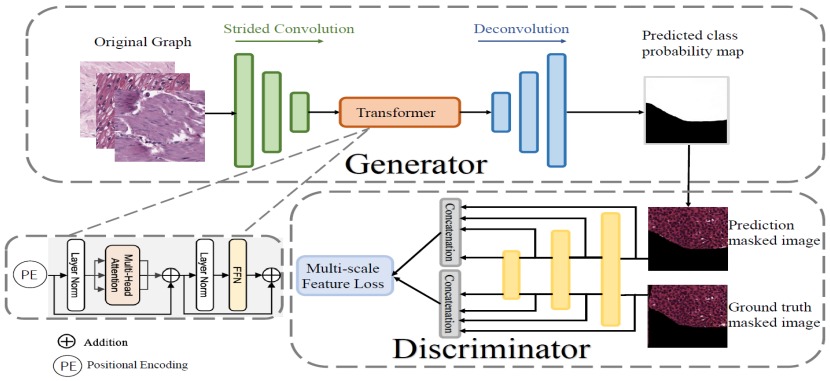

Figure 1: GAN + Transformer Model Structure

The generator part consists of strided convolution, transformer, and deconvolution, and finally generates a predicted class probability map. For the discriminator part, it maximizes the distance between the prediction-masked image and ground truth-masked image to distinguish the difference between them. These two are trained in an alternating fashion to improve the performance of the other.

For the ground truth, the white area values are changed to 1 since the value scales fit with the probability map from 0 to 1.

Generator The size of Image $3 \times 256 \times 256$ transforms to a sequence by utilizing six down-sampling layers with filter size 5×5 convolution (stride = 2), one down-sampling layer with filter size 2×2 convolution (stride = 1), and one down-sampling layer with filter size 3×3 convolution (stride = 2). Each convolution operation is followed by a batch Norm layer and a LeakyReLU activation layer. For the first six layers, there is a residual block between each layer.

Before the transformer structure, there is positional encoding. We used a standard method, which is the combination of $\sin(\frac{k}{n^{\frac{2i}{d}}})$ and $\cos(\frac{k}{n^{\frac{2i}{d}}})$.

The Transformer is composed of L Transformer layers, each of which has a standard architecture, which consists of a Multi-Head Attention (MHA) block and a Feed Forward Network (FFN).

The Deconvolution uses eight layers and three different filter sizes to transpose convolution for up-sampling, including 2×2, 3×3, and 4×4. The output is a $2 \times 256 \times 256$ figure. We utilized the SoftMax function, transferred the output to the probability map of two classes and chooe the first layer as the probability of the positive class. Besides, setting 0.5 as the benchmark, the probability with $P > 0.5$ transfer the value to 1, while the probability with $P < 0.5$ transfer the value to 0. Finally, the output is a binary mask. As shown in the Figure 1, the black area is considered negative, and the white area is considered negative. This model uses the Dice loss as the the generator loss.

Discriminator The discriminator inputs are prediction-masked image and ground truth-masked image. The discriminator is composed of six blocks. Each of these blocks consists of a different filter sizes convolution layer, including 4×4, 5×5, and 7×7, a batch normalization layer and a LeakyReLU activation layer. Instead of only using the final output of discriminator, this model extract the 2^{nd}, 3^{rd}, 4^{th} output from two images and cat as one feature matrix, and calculates the multi-scale L1 loss of this feature matrix, and uses this L1 loss as loss which is better for comparing two images more precisely.

3.1.3 Diffusion Probabilistic Model

The proposed diffusion model is based on Wu et al., 2022. Figure 2 illustrates the architecture of the reverse diffusion process, in which a modified UNet is adopted into every time step for segmentation tasks. In every time step t, the UNet takes the current segmentation map x_t and the raw image I as the signal input and conditional input, respectively. The output of the UNet decoder x_{t-1} is taken as the input in the next time step.

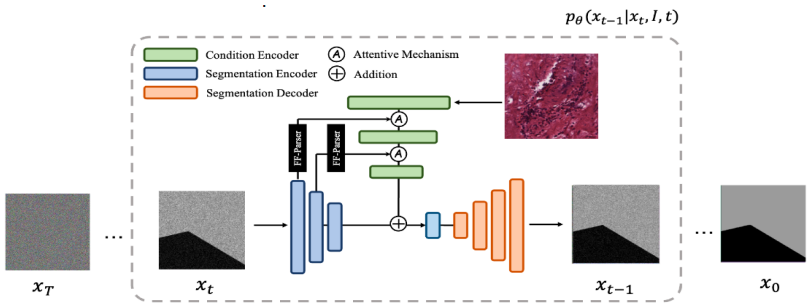

Figure 2: MedSegDiff Architecture

Gaussian Diffusion The diffusion process of the model follows standard DPM, which consists a forward process and a reverse process. In the forward process, the segmentation map x_0 is added with gaussian noise at every time step t until it becomes isotropic Gaussian distribution x_T. In the reverse process, on the other hand, the modified UNet learns to denoise x_T by adding it with an estimated distribution p_θ, where θ is the reverse process parameter. The time step t is integrated into the embedded feature in the UNet decoder. During UNet training, the output of the UNet decoder is used as the estimation for calculating the loss using Mean Square Error.

UNet The architecture of the modified UNet resembles ResUNet. Both the segmentation encoder and the condition enconder consist of 4 layers, where the first 3 layers contain 2 residual blocks and an additional downsampling convolutional layer with dimension 2 and stride 2, while the last layer contains only 2 residual blocks. The bidge block consists of 2 residual blocks. The decoder, similar to the encoders, consists of 4 layers, where the first 3 layers contain 3 residual blocks and an additional upsampling convolutional layer with dimension 2, while the last layer contains only 3 residual blocks. The residual blocks are implemented following ResNet34, with 2 sequential blocks that consist of a normalization layer, a SiLU activation layer and a convolutional layer with kernel size 3 and stride 2. In addition, time step t is integrated through an embedding layer with a SiLU layer and a linear layer.

Conditional Encoding The modified UNet differs from the one adopted by standard DPM mainly in the use of dynamic conditional encoding mechanism, an attentive-like mechanism that integrates the conditional and signal inputs on feature level. Specifically, the attention blocks are added between the residual blocks in the condition encoder and the bridge block, which first normalizes both the conditional encoding feature m_I and the x_t encoding feature m_x, multiplies the normalized feature maps together and multiplies the product by m_I again to enhance the conditional feature. This effective mechanism helps the model dynamically calibrate the segmentation target region.

3.2 Evaluation metric

To quantify the performance and quality of the segmentation result, we consider using the (1 - DICE score) as the loss function of our models. For evaluating and comparing our model performances, we use both DICE and IoU as metrics. Below is the definition of DICE score and IoU:

$$DICE = \frac{2 * y_{pred} \cdot y_{true} + \epsilon}{sum(y_{pred} + y_{true}) + \epsilon}$$

$$IoU = \frac{y_{pred} \cdot y_{true} + \epsilon}{sum(y_{pred} + y_{true}) - y_{pred} \cdot y_{true} + \epsilon}$$

Where ϵ is a smoothing factor.

4 Experimental results

4.1 Predictions from the models

Baseline and GAN + Transformer The baseline model shows sign of overfitting starting in 6th epoch. The proposed GAN + Transformer model was trained for 50 epochs using two labels. For Baseline model and GAN + Transformer model, DICE scores are used to evaluate the performance of the model, and the unlabeled pixels are ignored. Table 2 compares the two models' performances on the test set.

The prediction generated by the Baseline Model shows as the Figure 4 . The black region represents non-tumor, white with tumor, grey with unlabeled tumor in ground truth.

As shown in the Figure 5, the prediction by the GAN + Transformer captures some features of the original image but is not precise enough, while it tends to select regions based on different colors.

Overall, as the number of epochs increases, the Dice score of the GAN + Transformer improves with some minor fluctuations. However, even after 50 epochs, there are some characteristics of the original images not fully captured, which may be attributed to the size and complexity of the GAN + Transformer model. It should be noted that GANs are difficult to train and can be unstable. Therefore, it may be helpful to explore pre-trained models for GAN as an alternative approach.

Diffusion Probabilistic Model Figure 6 shows some of the resulting prediction images we obtained after training the proposed diffusion model for 40k steps with two labels (255 for positive and 0 for negative/unlabeled), which took around 6 hours for every 5000 steps on four T4 GPUs. However, the segmentation masks for the entire dataset couldn't be generated due to a GPU memory overflow problem. While we attempted to incorporate ideas from truncated DPM, we faced implementation errors that prevented us from training the model. Here the results of some data points are presented individually. To evaluate the results using IoU and Dice scores, the probability maps are converted to binary based on a threshold value. Table 3 displays the evaluation scores of 4 images after conversion.

As we can see, the predicted segmentation masks are not that good compared to the GAN + Transformer results, and the evaluation scores are low for all 4 images. We think it is because of the following reasons -

a) Usually, diffusion models require a large number of iterations/steps to train properly. It helps the model learn the appropriate noise distribution and incorporate it into the signal generation process. Hence, if we have the computational resources, training for more number of iterations (around 100k steps) should produce better results.

b) Because of the GPU memory issue, we could not evaluate the results for the whole dataset. Hence, commenting on whether the model underfits or overfits is very vague.

5 Discussion

5.1 Overview

We have implemented two models for whole slide image segmentation - a GAN + Transformer model and a diffusion probabilistic model, using the WSI dataset by Z. Zhang et al., 2019. For the GAN + Transformer model, we used a discriminator based on CNN, with the architecture and algorithms similar to the one proposed by Xue et al., 2017, and a Transformer-based generator that generates pixels over a spatial region, which is based on work by Wang et al., 2021. For the diffusion model, we have a UNet integrated into the reverse diffusion stage for segmentation, similar to Wu et al., 2022. We tried to modify the architecture and training algorithms of DPM by shortening the diffusion stages using the idea of truncated DPM Zheng et al., 2022 but failed to train it due to implementation errors. We trained and tested our models on the WSI dataset and compared the result of these two models and the S-Net model by Z. Zhang et al., 2019 based on the Dice score.

5.2 Current limitations

The present study evaluates the performance of state-of-the-art models in segmenting Whole Slide Images (WSIs) and identifies challenges that arise due to the complex molecular structures present in WSIs. Despite exceptional performance on MRI segmentation tasks, these models struggle to distinguish between positive and negative regions in WSIs. Furthermore, the models should be trained and evaluated on a dataset with three labels (tumor, non-tumor, and unlabeled), but for the sake of simplicity, the proposed models merge non-tumor and unlabeled regions into a single label, missing the opportunity to learn the differences between these regions and the uncertainty on the unlabeled regions. The study acknowledges that due to limited time and restricted compute engines, the models were trained without data augmentation and fine-tuning, which may have contributed to overfitting issues observed in the results.

5.3 Future directions

Inspecting the output segmentation masks, we noticed our models are picking up color-related features. This is undesired as the colors in WSI's fluctuate due to artifacts. Hence one thing we would like to utilize stain augmentation techniques to make our models robust against artifacts. Data augmentation will also be implemented in the next iteration–it was skipped since the goal of current iteration is to get everything working. We would also investigate different encoding scheme of the unlabeled pixels to avoid information loss as mentioned in the previous section. For both GAN+Transformer and the diffusion model, we will implement different loss functions in the next iteration–for instance, KL divergence for the diffusion model, and multi scale L2 loss for the GAN+Transformer. To accelerate the diffusion process, we plan to further experiment with merging Truncated DPM into the current diffusion model. Both models will be further fine-tuned in the next iteration, and we will incorporate additional regularization techniques such as mixup and cutmix in deep long-tailed learning modelsY. Zhang et al., 2023 to enhance the model's segmentation performance on imbalanced datasets. Last but not the least, we will apply the models on other medical image datasets to further explore the potential and limitations of our proposed models.

6 Author Contributions

- **Jiaming Yao:** Proposed the initial idea of what the GAN model should look like; built the initial training data pipeline; conducted exploratory literature research on GAN and Transformer; responsible for the training and evaluation of the S-Net model.
- **Saurav Telge**: Reviewed and researched different diffusion models for segmentation. Implemented a diffusion probabilistic model (DPM) based on the Wu et al., 2022 paper. Trained the model on the whole slide image dataset described above and evaluated the results. Tried implementing the truncated DPM model Zheng et al., 2022 to improve performance.
- **Bolin Wu**: Explored the architecture and algorithms of diffusion models for image segmentation. Trained and Tested the diffusion model.
- **Yaxuan Hou**: Conducted the Exploratory literature research on GAN and transformer model. Trained and tested the GAN + Transformer model.
- **Huiyao Yang**: Performed extensive literature research about GAN and Transfomer model. Trained and tested GAN + Transformer model.
- All authors contributed equally to the project.

Contribution	Yaxuan Hou	Saurav Telge	Bolin Wu	Huiyao Yang	Jiaming Yao
Built the initial training data pipeline					✓
Train and test baseline model					✓
Conducted exploratory literature research on GAN and Transformer	✓			✓	✓
Train and test GAN + Transformer model	✓			✓	✓
Conducted exploratory literature research on diffusion model		✓	✓		
Train and test diffusion model		✓	✓		

7 Appendix

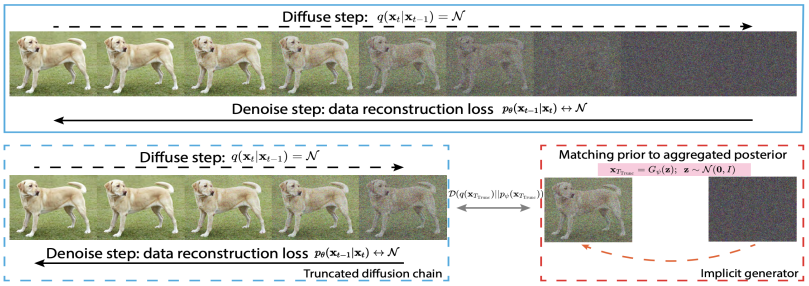

Figure 3: Illustration of DPM (above) vs. TDPM (below)

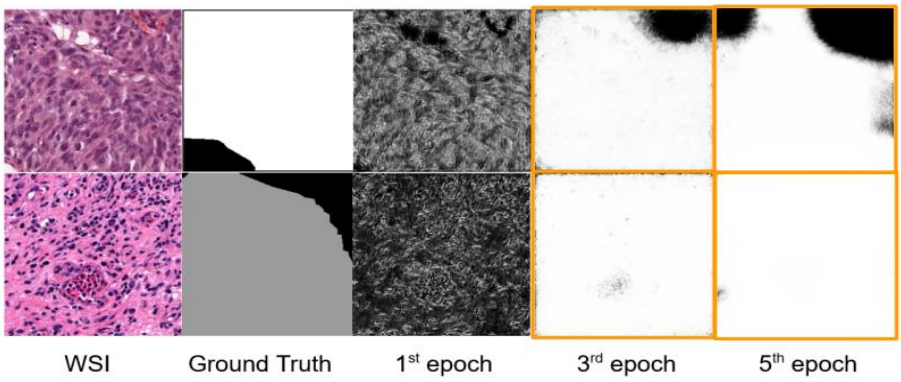

Figure 4: Baseline Model Results

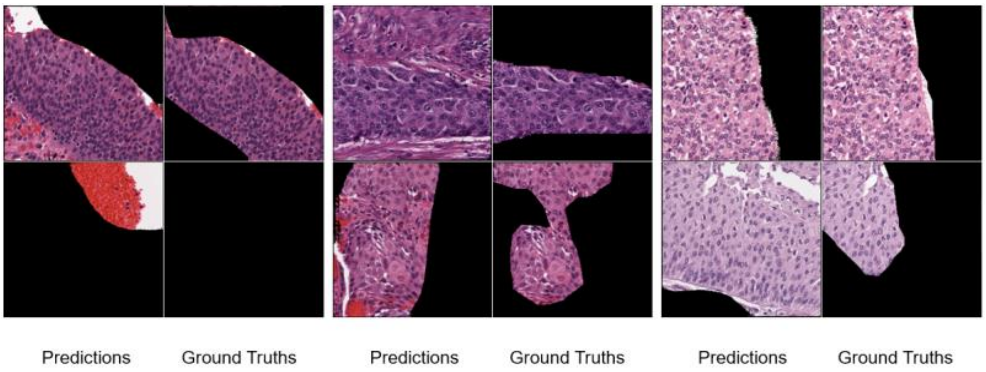

Figure 5: GAN + Transformer Model Results

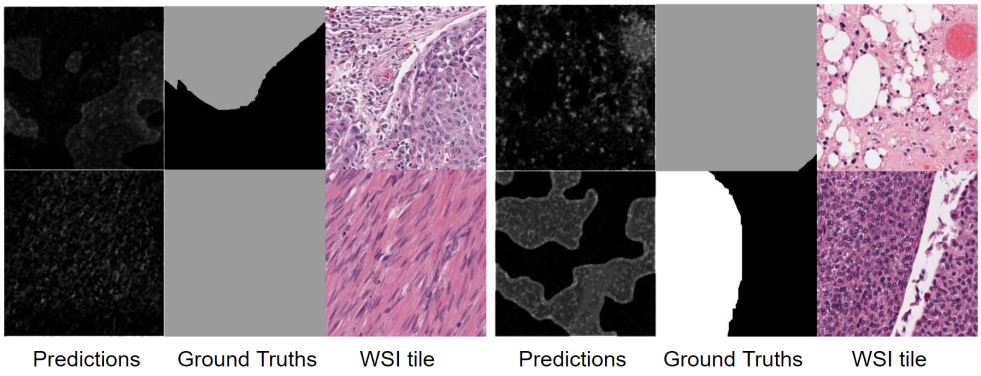

Figure 6: Diffusion Model Results

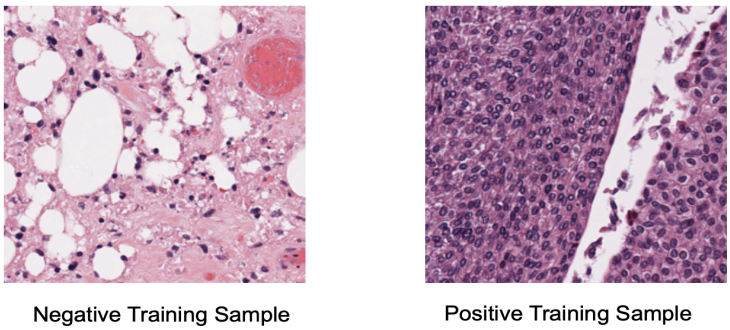

Figure 7: Sample Training Data Images

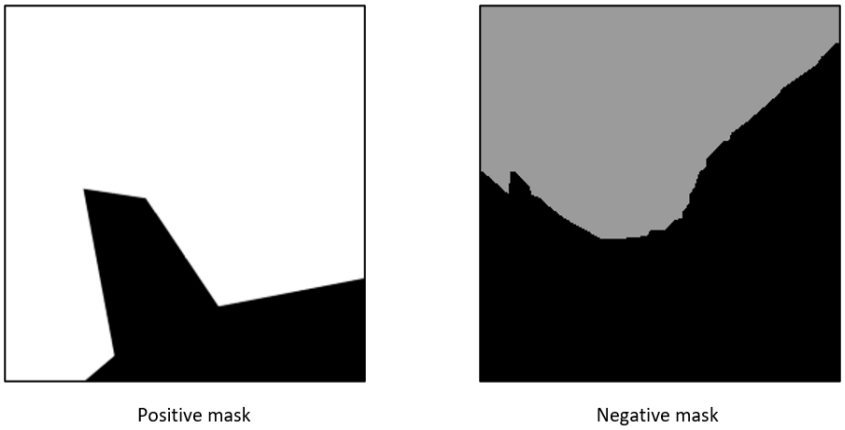

Figure 8: Sample Training Data Ground Truth

Dataset	Training split size	Testing split size
Whole Slide Image	293	104
Sampled Image	1778 (POS) + 3717 (NEG)	1210 (POS) + 1236 (NEG)

Table 1: Dataset Statistics

Approach	Epoch	DICE Score	IoU
Baseline	5	0.4066	0.2559
GAN + Transformer	50	0.3298	0.2034

Table 2: Model Performances of Baseline and GAN + Transformer

Image Name[1]	IoU Score	DICE Score
2_00119_sub0_0_region11_5_Neg	0.0162	0.0311
2_00175_sub0_0_region7_4_Neg	0.0381	0.0675
2_00316_sub0_0_region2_8_Neg	0.0602	0.2002
2_00377_sub0_0_region13_4_Pos	0.0883	0.1365
Overall	0.05	0.08

Table 3: Model Performances of Diffusion Model

[1]For naming the images, we followed a convention where the first part denotes the specific tile out of all the tiles extracted from the whole slide image, and then the second part states whether it is a positive or negative label.

References

Xue, Y., Xu, T., Zhang, H., Long, R., & Huang, X. (2017). Segan: Adversarial network with multi-scale l1 loss for medical image segmentation. https://doi.org/10.48550/arXiv.1706.01805

Zhang, Z., Chen, P., McGough, M., Xing, F., Wang, C., Bui, M., Xie, Y., Sapkota, M., Cui, L., Dhillon, J., Ahmad, N., Khalil, F. K., Dickinson, S. I., Shi, X., Liu, F., Su, H., Cai, J., & Yang, L. (2019). Pathologist-level interpretable whole-slide cancer diagnosis with deep learning. *Nature Machine Intelligence*, *1*(5), 236–245. https://doi.org/10.1038/s42256-019-0052-1

Wang, W., Chen, C., Ding, M., Li, J., Yu, H., & Zha, S. (2021). Transbts: Multimodal brain tumor segmentation using transformer. https://doi.org/10.48550/arXiv.2103.04430

Huang, L., Chen, L., Zhan, B., & Chai, S. (2022). A transformer-based generative adversarial network for brain tumor segmentation. https://doi.org/10.48550/arXiv:2207.14134

Wu, J., Fu, R., Fang, H., Zhang, Y., Yang, Y., Xiong, H., Liu, H., & Xu, Y. (2022). Medsegdiff: Medical image segmentation with diffusion probabilistic model. https://doi.org/10.48550/ARXIV.2211.00611

Zheng, H., He, P., Chen, W., & Zhou, M. (2022). Truncated diffusion probabilistic models. *stat*, *1050*, 7.

Wu, J., Fu, R., Fang, H., Zhang, Y., & Xu, Y. (2023). Medsegdiff-v2: Diffusion based medical image segmentation with transformer. https://doi.org/10.48550/ARXIV.2301.11798

Zhang, Y., Kang, B., Hooi, B., Yan, S., & Feng, J. (2023). Deep long-tailed learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.