CSCI 1430 Final Project Report: MedSegDiff: A Diffusion Model for Medical Image Segmentation

All attention is on U: Yiyang Zhang, Zhuoyang Lyu, Xiangxi Mu, Xilin Wang.
TA name: Julia Fu. Brown University

Abstract

Recent advancements in diffusion probabilistic methods have enhanced their capability to generate high-quality images from simple prompts. Leveraging this power, we specifically adapted diffusion models for medical image segmentation, implementing an end-to-end approach. Our model demonstrates slightly improved performance across various metrics compared to traditional segmentation methods such as Unet and transformer architectures, suggesting its robustness and effectiveness in generating segmentation masks conditioned on the original image inputs.

Our code can be found at https://github.com/Rice-wxl/CSCI1430-Final-Project-MedImage-Segmentation, it is mainly based on https://github.com/tomeramit/SegDiff.

1. Introduction

Medical image segmentation is a computer vision task that divides a medical image into multiple meaningful segments. The goal of medical image segmentation is to provide a precise and accurate representation of objects for diagnosis, treatment planning, and quantitative analysis. However, medical images present several challenges for deep learning due to their unique characteristics.

First, the medical images have a large variance due to different effects including patient anatomy and imaging modalities, which could make the generalization process across different datasets difficult. In addition to that, in medical imaging tasks, the regions of interest usually occupy a small scatter fraction of the original images. Such an imbalance between the target and the background could make the model struggle to detect the small but crucial features.

Diffusion models have recently gained attention in image processing, including medical image segmentation. Diffusion models iteratively refine their outputs, starting from random noise and progressively improving the segmentation map. This iterative refinement can lead to more stable and robust segmentation, particularly in noisy or low-quality images. They are also adept at capturing complex structures and fine details in images. This is particularly beneficial for medical image segmentation, where accurately delineating complex anatomical structures is crucial. Due to their generative nature, diffusion models can produce high-quality, detailed segmentations. They model the data distribution more comprehensively, leading to segmentations that are closer to the ground truth, even in challenging conditions. By learning to generate data that follows the distribution of the training data, diffusion models often generalize better to unseen data.

We have implemented a diffusion probabilistic-based segmentation model for medical image segmentation tasks. In the iterative sampling process, the model conditions each step with image priors to learn the segmentation map. To achieve adaptive regional attention, the model integrate the segmentation map of the current step into the image prior encoding at each step. This approach allows the corrupted current-step mask to dynamically enhance the condition features, thus improving reconstruction accuracy.

We verify our SegDiff model on skin lesion medical segmentation tasks, and the model outperforms the baseline we set with U-Net, achieving reasonable segmentation masks.

2. Related Work

2.1. U-Net U-Net is a convolutional neural network that was developed for biomedical image segmentation. The network is based on a fully convolutional neural network whose architecture was modified and extended to work with fewer training images and yield more precise segmentation.

It consists of a contracting path and an expansive path. The contracting path follows the typical architecture of a convolutional network, consist the repeated application of two unpadded convolution layers, each followed by a rectified linear unit and a max pooling operation for downsampling. At each downsampling step, the model doubles the number of feature channels. Every step in the expansive path consists of an upsampling of the feature map followed by an up-convolution that halves the number of feature channels, a concatenation with the correspondingly cropped feature map from the contracting path, and two convolution layers, each followed by a rectified linear unit.

However, such an excessive down-sampling process leads to more loss in spatial information.

2.2. Medical Transformer Recent breakthroughs in large language models (LLMs) utilizing transformer architecture have impacted every field. The Medical Transformer is an architecture specifically designed for medical image segmentation tasks. While traditional transformers use self-attention, this model employs a gated axial-attention mechanism, which adds an additional control feature to the attention modules. The model was trained with a Local-Global training strategy (LoGo) that further improved performance by analyzing the entire input image globally and in patches separately. The Medical Transformer has demonstrated state-of-the-art performance, representing another category of architectures for segmentation tasks based on the attention mechanism.

3. Method

We leverage a diffusion probabilistic model originally designed for semantic image segmentation tasks, SegDiff, on medical images of skin lesion. A diffusion probabilistic model generates the output given a randomly samplied noisy iteratively, one step at a time. For the task of image segmentation, our diffusion model produces the output segmentation mask as a denoised version of a randomly Gaussian-sampled mask, conditioned on the original image of the skin lesion. This conditioning is analogous to a diffusion model that generates images based on an input text describing what the image should look like.

3.1. Architecture The SegDiff model specifically predicts the noise that is added to the segmentation mask, such that it is able to recover the denoised mask given an arbitrarily noisy one. Our model is an encoder-decoder based structure: there are two separate encoders for the noisy mask x_t and the conditioned image I, respectively; the result is then passed into another encoder to combine information, before finally passing into the decoder to predict the noise.

$$\epsilon_\theta(x_t, I, t) = D(E(F(x_t) + G(I), t), t). \tag{1}$$

The final encoder-decoder structure, E and D, is based on U-net, in which each level comprises of convolutional residual blocks along with attention layers at certain levels. The encoder for the conditioned image, I, contains multiple 2D convolutional layers and leverages Residual in Residual Dense Blocks (RRDBs) that combine multi-level residual connections. The segmentation mask encoder, F, is a simple 2D convolutional layer.

3.2. Training During training, a time t is uniformly sampled from 1 to T, where T is a predetermined total diffusion steps. Then, a ground truth noise is randomly generated from Gaussian distribution and applied to the ground truth segmentation mask (Equation 2).

$$x_t = \alpha \cdot x_0 + \beta \cdot \epsilon, \epsilon \sim N(0, I_{n\times n}) \tag{2}$$

The model then takes in the noisy mask, x_t, as well as the conditioned image to predict the noise following Equation 1. The training process minimizes the distance between the predicted noise ϵ_θ and the true noise ϵ across the dataset:

$$E_{x_0,\epsilon,t}\left[\|\epsilon - \epsilon_\theta(x_t, I, t)\|^2\right] \tag{3}$$

3.3. Inference At inference time, an arbitrarily noisy segmentation mask is first sampled as x_T. At each diffusion timestep t, our model is fed with the current estimate of the mask x_t along with the conditioned image, namely the original skin lesion image, to predict the noise imposed to the noisy mask. This noise is then used to predict the previous timestep's segmentation mask, x_{t-1}, as illustrated by Figure 1. The process is iteratively performed T times to eventually produce the segmentation mask, x_0.

4. Results

4.1. Data We used the ISIC (International Skin Imaging Collaboration) 2016 dataset got from ISIC Challenge website (https://challenge.isic-archive.com/data/). The dataset contains skin lesion images and masks for the lesion area. We used 900 images and masks for training and 379 images and masks for test.

4.2. Model Performance Our model was trained for 30,000 steps on the ISIC dataset. IoU (Intersection over Union) was used as the metrics for our diffusion model. Our model was trained successully, and its train loss decreases ideally (Figure 1). According to Table 1, We get an average IoU of 0.501 which is not very ideal in general, and our baseline Unet is also not performing well. The main reason is our dataset is not within itself. Some images have dark boundaries but some images are clean, and our model can't handle this situation very well.

Model	IoU
Unet	0.490
Ours	0.501

Table 1. Model Performance

4.3. Dark Corners The model incorrectly predicts a skin lesion in the top left corners, as seen in the example prediction. This error, likely due to the model's reliance on pixel

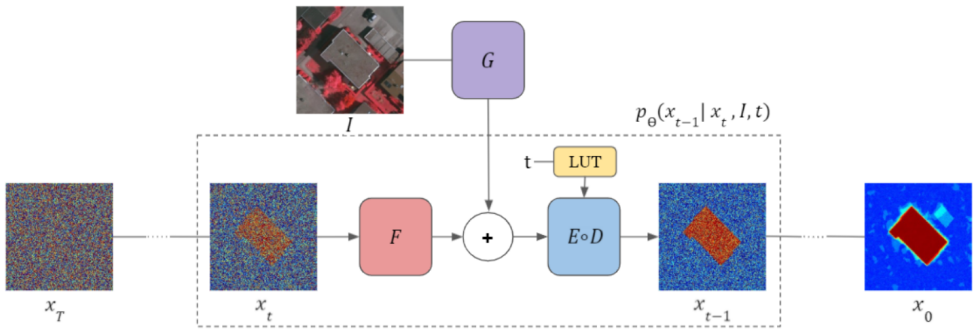

Figure 1. The SegDiff model structure for predicting medical segmentation mask from a conditioned image.

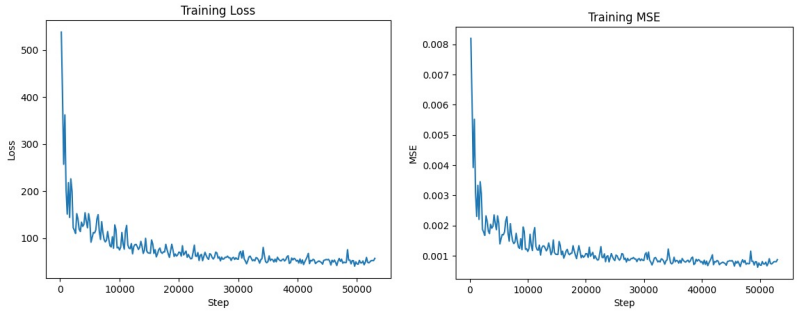

Figure 2. *Left:* Training Loss. *Right:* Training MSE.

darkness and color, is common. To address this, we plan to use data augmentation to create more images with dark corners, improving the model's accuracy by better training it to handle such scenarios.

4.4. Color Differences The model suffers from data where the skin lesion color is generally similar to the natural skin. Additionally, the gaps between the skin lesions affect the model's predictions. Moreover, several other scattered spots that appear far away from the lesion part can introduce noisy segmentation results.

5. Conclusion

For our project, we aimed to leverage deep learning models for medical segmentation tasks, specifically focusing on skin lesion analysis. Utilizing deep learning in this area can significantly enhance the efficiency of medical workflows, increase diagnostic accuracy, and ultimately save lives. We were particularly interested in the potential of diffusion models to generate accurate segmentation masks and wanted to compare their performance against traditional segmentation models like U-Net. However, our results have not met our expectations. Both our diffusion model and the U-Net model exhibited significant performance variations across different images. Moving forward, our primary objectives are to enhance the models' robustness in handling challenging image features, such as dark corners and color variations.

References

[1] Amit, T., Shaharbany, T., Nachmani, E., Wolf, L. (2021). SegDiff: Image Segmentation with Diffusion Probabilistic Models. arXiv. https://doi.org/10.48550/arXiv.2112.00390

[2] Ronneberger, O., Fischer, P., Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. arXiv. https://doi.org/10.48550/arXiv.1505.04597

[3] Valanarasu, J. M. J., Oza, P., Hacihaliloglu, I., Patel, V. M. (2021). Medical Transformer: Gated Axial-Attention for Medical Image Segmentation. Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI 2021). arXiv:2102.10662. https://doi.org/10.48550/arXiv.2102.10662

Appendix

Team contributions

Please describe in one paragraph per team member what each of you contributed to the project.

Xilin Wang I worked on implementing the SegDiff model on our intended dataset. This involves setting up the SegDiff model environment using an App container on Brown CCV OSCAR, understanding the codebase for the model implementation, and modifying the code as

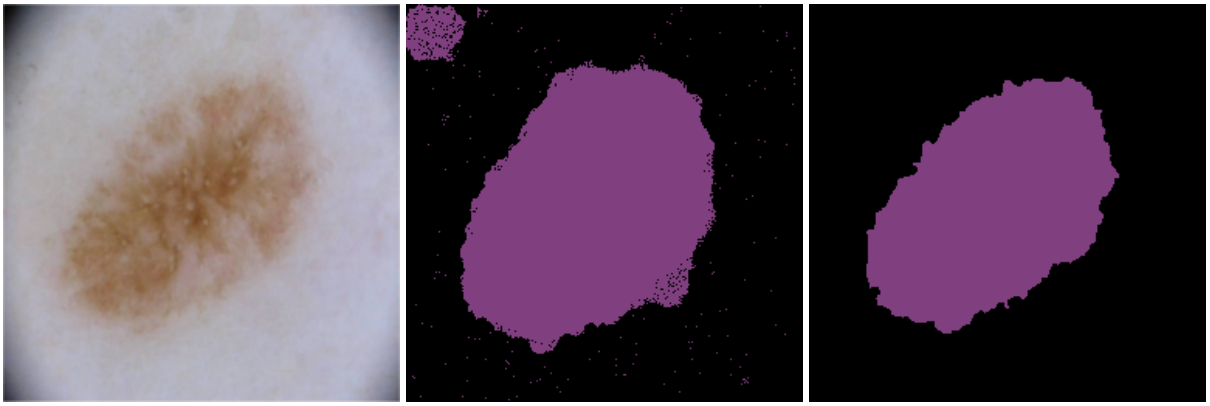

Figure 3. Dark Corners. *Left:* Input. *Mid:* Our Prediction. *Right:* Ground Truth.

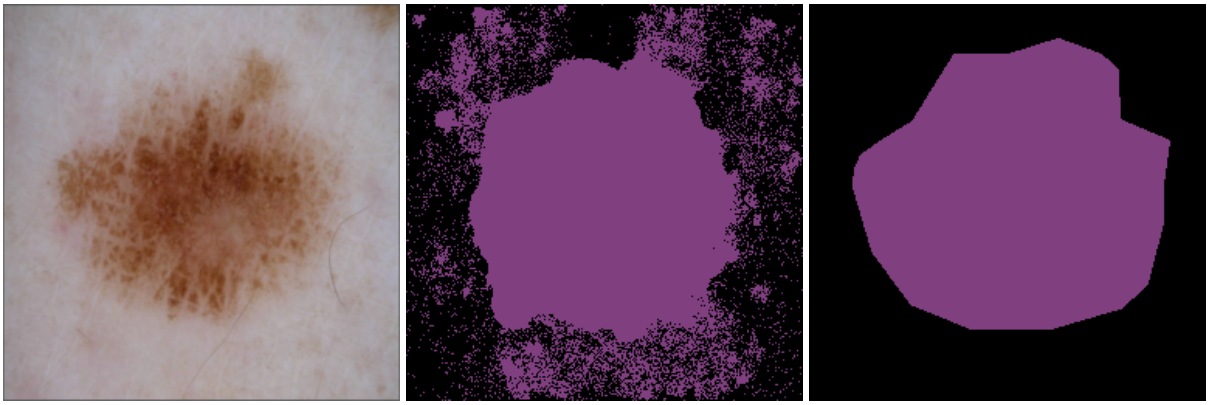

Figure 4. Color Difference. *Left:* Input. *Mid:* Our Prediction. *Right:* Ground Truth.

needed to be compatible with our dataset. Our codebase as well as the environment were eventually compatible and could be trained on OSCAR. For the poster and the final report, I focus on the methodology and model architecture, and help with the result and evaluation section.

Zhuoyang Lyu I was responsible for writing the code for the model performance experiment. To monitor and log the training process, as well as to tune and standardize various hyperparameters, we needed a standardized pipeline to train, test, and evaluate the models. Additionally, I did the motivation, problem definition, training process visualization, and analysis in the poster. I also help with the introduction, related work, and results sections of the report.

Yiyang Zhang I have taken part in adjusting the SegDiff model structures so that the model could run on our dataset. Given the differences between the datasets, the image size and format could lead to different input forms. Therefore, I have developed different dataloaders that can handle various datasets and adjusted the sampling method for our model. I helped analyze the output results. I have also contributed to writing the introduction, related work, and analysis sections of this report.

Xiangxi Mu I was responsible for writing preprocess and finding dataset. So that the data can be used for various models. I also participated in looking for and running other models. I helped with writing the poster. I contribute to the results and concolution part of the report.