MedSegDiff: Medical Image Segmentation with Diffusion Probabilistic Model

Author name(s) withheld EMAIL(S) WITHHELD
Address withheld

Editors: Under Review for MIDL 2023

Abstract

Diffusion probabilistic model (DPM) recently becomes one of the hottest topic in computer vision. Its image generation application such as Imagen, Latent Diffusion Models and Stable Diffusion have shown impressive generation capabilities, which aroused extensive discussion in the community. Many recent studies also found it is useful in many other vision tasks, like image deblurring, super-resolution and anomaly detection. Inspired by the success of DPM, we propose the first DPM based model toward general medical image segmentation tasks, which we named MedSegDiff. In order to enhance the step-wise regional attention in DPM for the medical image segmentation, we propose dynamic conditional encoding, which establishes the state-adaptive conditions for each sampling step. We further propose Feature Frequency Parser (FF-Parser), to eliminate the negative effect of high-frequency noise component in this process. We verify MedSegDiff on three medical segmentation tasks with different image modalities, which are optic cup segmentation over fundus images, brain tumor segmentation over MRI images and thyroid nodule segmentation over ultrasound images. The experimental results show that MedSegDiff outperforms state-of-the-art (SOTA) methods with considerable performance gap, indicating the generalization and effectiveness of the proposed model.

Keywords: diffusion probabilistic model, medical image segmentation, brain tumor, optic cup, thyroid nodule

1. Introduction

Medical image segmentation is the process of partitioning a medical image into meaningful regions. Segmentation is a fundamental step in many medical image analysis applications such as diagnosis, surgical planning, and image-guided surgery. This is important because it allows doctors and other medical professionals to better understand what they're looking at. It also makes it easier to compare images and track changes over time. In recent years, there has been a growing interest in automatic medical image segmentation methods. These methods have the potential to reduce the time and effort required for manual segmentation, and to improve the consistency and accuracy of results. With the development of the deep learning techniques, more and more studies successfully applied the neural network (NN) based models to the medical image segmentation tasks, from the popular convolution neural networks (CNN) (Ji et al., 2021) to the recent vision transformers (ViT) (Chen et al., 2021; Wang et al., 2021; Liu et al., 2022; Zhao et al., 2021).

Very recently, diffusion probabilistic model (DPM)(Ho et al., 2020) gained popularity as a powerful class of generative models(Zhao and Shi, 2021; Goodfellow et al., 2020), that is able to generate images with high diversity and synthesis quality. Recent large diffusion

models, such as DALL-E2(Ramesh et al., 2022), Imagen(Saharia et al., 2022a), and Stable Diffusion(Rombach et al., 2022) have shown incredible generation capability. Diffusion models are originally applied in fields in which there is no absolute ground-truth. However, recent studies show that it is also effective for the problems in which the ground-truth is unique, like super-resolution(Saharia et al., 2022b) and deblurring(Whang et al., 2022).

Inspired by the recent success of DPM, we design a unique DPM-based segmentation model for the medical image segmentation tasks. To our knowledge, we are the first to propose the DPM-based model under the background of general medical image segmentation with different image modalities. We note that in tasks of medical image segmentation, the lesions/organs are often ambiguous and hard to discriminate from the background. In that case, an adaptive calibration process is the key to obtain a delicate result. Following this mindset, we propose dynamic conditional encoding over vanilla DPM to design the proposed model, named MedSegDiff. Note that in the iterative sampling process, MedSegDiff conditions each of the step with image prior, in order to learn the segmentation map from it. Toward the adaptive regional attention, we integrate the segmentation map of current step into the image prior encoding at each step. The specific implementation is to fuse the current-step segmentation mask with the image prior on the feature level with a multi-scale manner. In this way, the corrupted current-step mask helps to dynamically enhance the condition features, thus improves the reconstruction accuracy. In order to eliminate the high-frequency noises in the corrupted given mask in this process, we further propose the feature frequency parser (FF-Parser) to filter the features in the Fourier space. FF-Parsers are adopted on each skip connection path for the multi-scale integration. We verify MedSegDiff on three different medical segmentation tasks, the optic-cup segmentation, the brain tumor segmentation, and the thyroid nodule segmentation. The images of these tasks have different modalities, which are the fundus images, brain CT images, the ultrasound images respectively. MedSegDiff outperforms the previous SOTA on all three tasks with different modalities, which shows the generalization and effectiveness the proposed method. In brief, the contributions of the paper are:

- The fist to propose DPM-based model toward general medical image segmentation.
- Dynamic conditional encoding strategy is proposed for step-wise attention.
- FF-Parser is proposed to eliminate the negative effects of high-frequency components.
- SOTA performance on three different medical segmentation tasks with different image modalities.

2. Method

We design our model based on diffusion model mentioned in (Ho et al., 2020). Diffusion models are generative models composed of two stages, a forward diffusion stage and a reverse diffusion stage. In the forward process, the segmentation label x_0 is gradually added Gaussian noise through a series of steps T. In the reverse process, a neural network is trained to recover the original data by reversing the noising process, which can be represented as:

$$p_\theta(x_{0:T-1}|x_T) = \Pi_{t=1}^{T} p_\theta(x_{t-1}|x_t), \tag{1}$$

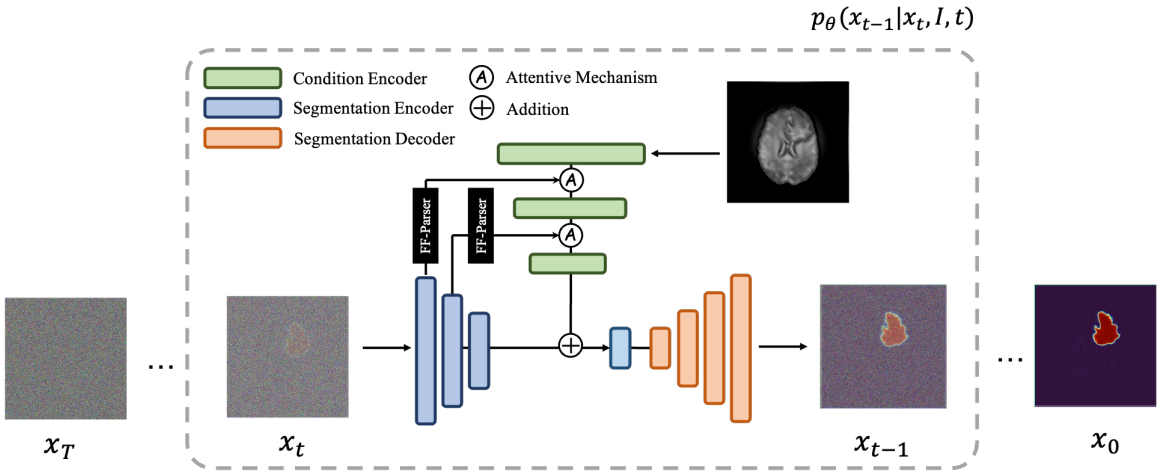

Figure 1: An illustration of MedSegDiff. For the clarity, the time step encoding is omitted in the figure.

where θ is reverse process parameters. Starting from a Gaussian noise, $p_\theta(x_T) = \mathcal{N}(x_T; 0, I_{n\times n})$, where I is the raw image, the reverse process transforms the latent variable distribution $p_\theta(x_T)$ to the data distribution $p_\theta(x_0)$. To be symmetrical to the forward process, the reverse process recovers the noise image step by step to obtain the final clear segmentation.

Following the standard implementation of DPM, we adopt a UNet as the network for the learning. An illustration is shown in Figure 1. In order to achieve the segmentation, we condition the step estimation function ϵ by raw image prior, which can be represented as:

$$\epsilon_\theta(x_t, I, t) = D((E_t^I + E_t^x, t), t), \tag{2}$$

where E_t^I is the conditional feature embedding, in our case, the raw image embedding, E_t^x is the segmentation map feature embedding of the current step. The two components are added and sent to a UNet decoder D for the reconstruction. The step index t is integrated with the added embedding and decoder features. In each of these, it is embedded using a shared learned look-up table, following (Ho et al., 2020).

2.1. Dynamic Conditional Encoding

In most conditional DPM, the conditional prior will be a unique given information. However, medical image segmentation is notorious for its ambiguous objects. The lesions or tissues are often hard to discriminate from its background. The low-contrast image modalities, such as MRI or ultrasound images, make it even worse. Given only a static image I as the condition for each step will be hard to learn. To address this problem, we propose a dynamic conditional encoding for each step. We note that on the one hand, the raw image contains the accurate segmentation target information but hard to discriminate from the background, on the other hand, the current-step segmentation map contains the enhanced target regions but not accurate. This motivated us to integrate the current-step segmentation information x_t into the conditional raw image encoding for the mutual complement. To be specific, we

implement the integration on the feature level. In the raw image encoder, we enhance its intermediate feature with the current-step encoding features. Each scale of the conditional feature map m_I^k is fused with the x_t encoding features m_x^k with the same shape, k is the index of layer. The fusion is implemented by an attentive-like mechanism $\mathcal{A}$. In particular, two feature maps are first applied layer normalization and multiply together to get an affinity map. Then we multiply the affinity map with the condition encoding features to enhance the attentive region, which is:

$$\mathcal{A}(m_I^k, m_x^k) = (LN(m_I^k) \otimes LN(m_x^k)) \otimes m_I^k, \tag{3}$$

where $\otimes$ implies element-wise multiplication, LN denotes layer normalization. The operation is applied on the middle two stages, where each is the convolutional stage implemented following ResNet34. Such a strategy helps MedSegDiff dynamically localize and calibrate the segmentation. Although effective the strategy it is, another specific problem is that integrating x_t embedding will induce extra high-frequency noise. To address this problem, we propose FF-Parser to constrain the high-frequency components in the features.

2.2. FF-Parser

We connect FF-parser in the path ways of the feature integration. The function of it is to constrain the noise-related components in the x_t features. Our main idea is to learn a parameterized attentive (weight) map applying on the Fourier-space features. Given a decoder feature map $m \in \mathbb{R}^{H \times W \times C}$, we first perform 2D FFT(fast fourier transform) along the spatial dimensions, which we can represented as:

$$M = \mathcal{F}[m] \in \mathbb{C}^{H \times W \times C}, \tag{4}$$

where $\mathcal{F}[\cdot]$ denotes the 2D FFT. We then modulate the spectrum of m by multiplying a parameterized attentive map $A \in \mathbb{C}^{H \times W \times C}$ to M:

$$M' = A \otimes M, \tag{5}$$

where $\otimes$ denotes the element-wise product. Finally, we reverse M' back to the spatial domain by adopting inverse FFT:

$$m' = \mathcal{F}^{-1}[M']. \tag{6}$$

FF-Parser can be regarded as a learnable version of frequency filters which are wildly applied in the digital image processing (Pitas, 2000). Different from the spacial attention, it globally adjusts the components of the specific frequencies. Thus it can be learn to constrain the high-frequency component for the adaptive integration.

2.3. Training and Architecture

MedSegDiff is trained following the standard process of DPM (Ho et al., 2020). Specifically, the loss can be represented as:

$$\mathcal{L} = E_{x_0,\epsilon,t}[||\epsilon - \epsilon_\theta(\sqrt{\hat{a}_t}x_0 + \sqrt{1-\hat{a}_t}\epsilon, I_i, t)||^2]. \tag{7}$$

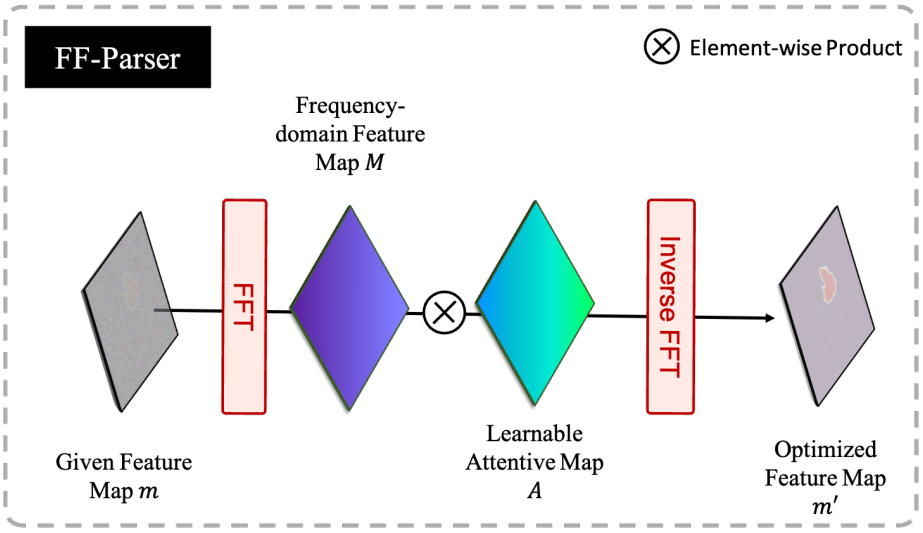

Figure 2: An illustration of FF-Parser. FFT denotes Fast Fourier Transform.

In each of the iteration, a random couple of raw image I_i and segmentation label S_i will be sampled for the training. The iteration number is sampled from a uniform distribution and ϵ from a Gaussian distribution.

The main architecture of MedSegDiff is a modified ResUNet(Yu et al., 2019), which we implement it with a ResNet encoder following a UNet decoder. The detailed network setting is following (Nichol and Dhariwal, 2021). I and x_t are encoded with two individual encoders. The encoder is consisted of the convolution stages containing multiple residual blocks. The number of residual blocks in each stage is following that of ResNet34. Each residual block is composed of two convolutional blocks, each one consists of group-norm and SiLU(Elfwing et al., 2018) active layer and a convolutional layer. The residual block receives the time embedding through a linear layer, SiLU activation, and another linear layer. The result is then added to the output of the first convolutional block. The obtained E^I and E^{x_t} are added together and sent to the last encoding stage. A standard convolutional decoder is connected to predict the final result.

3. Experiments

3.1. Dataset

We conduct the experiments on three different medical tasks with different image modalities, which are optic-cup segmentation from fundus images, brain tumor segmentation from MRI images, and thyroid nodule segmentation from ultrasound images. The experiments of glaucoma, thyroid cancer and melanoma diagnosis are conducted on REFUGE-2 dataset (Fang et al., 2022), BraTs-2021 dataset (Baid et al., 2021) and DDTI dataset (Pedraza et al., 2015), which contain 1200, 2000, 8046 samples, respectively. The datasets are publicly available with both segmentation and diagnosis labels. Train/validation/test sets are split following the default settings of the dataset.

3.2. Implementation Details

We experiment with huge, large, basic, and small variants of our model, *MedSegDiff++*, *MedSegDiff-L*, *MedSegDiff-B*, and *MedSegDiff-S*, respectively. In *MedSegDiff-S*, *MedSegDiff-B MedSegDiff-L*, *MedSegDiff++*, we use UNet with 4x, 5x, 6x, 6x downsamples respectively. In the experiments, we employ 100 diffusion steps for the inference, which is much smaller than most of the previous studies(Ho et al., 2020; Nichol and Dhariwal, 2021). All the experiments are implemented with the PyTorch platform and trained/tested on 4 Tesla P40 GPU with 24GB of memory except *MedSegDiff++* and *MedSegDiff-L*. All images are uniformly resized to the dimension of 256×256 pixels. The networks are trained in an end-to-end manner using AdamW(Loshchilov and Hutter, 2017) optimizer. *MedSegDiff-B* and *MedSegDiff-S* are trained with 32 batch size, *MedSegDiff-L* and *MedSegDiff++* are trained with 64 batch size. The learning rate is initially set to 1×10^{-4}. All models are set 25 times of ensemble in the inference. We use STAPLE(Warfield et al., 2004) algorithm to fuse the different samples. The diffusion based competitor EnsemDiff(Wolleb et al., 2021) is reproduced with the same setting for the fair comparison.

3.3. Main Results

We compare with SOTA segmentation methods proposed for the three specific tasks and general medical image segmentation methods. The main results are shown in Table 1. In the table, ResUnet(Yu et al., 2019) and BEAL(Wang et al., 2019) are proposed for optic disc/cup segmentation, TransBTS(Wang et al., 2021) and EnsemDiff(Wolleb et al., 2021) are proposed for the brain tumor segmentation, MTSeg(Gong et al., 2021) and UltraUNet(Chu et al., 2021) are proposed for the Thyroid Nodule segmentation, CENet(Gu et al., 2019), MRNet(Ji et al., 2021), SegNet(Badrinarayanan et al., 2017), nnUNet(Isensee et al., 2021) and TransUNet(Chen et al., 2021) are proposed for the general medical image segmentation. We evaluate the segmentation performance by Dice score and IoU.

In Table 1, we compare with the methods implemented with various network architectures, including CNN (ResUNet, BEAL, nnUNet, SegNet), vision transformer (TransBTS, TransUNet) and DPM (EnsemDiff). We can see the advanced network architectures commonly gain better results. For example, in optic-cup segmentation, ViT-based general segmentation method: TransUNet is even better than the CNN-based task toward method: BEAL. On brain tumor segmentation, recently proposed DPM-based segmentation method EnsemDiff outperforms all those previous ViT-based competitors, i.e., TransBTS and TransUNet. MedSegDiff not only adopts the recent successful DPM, but also designs an appropriate strategy over it specifically towards the general medical image segmentation task. We can see MedSegDiff outperforms all the other methods on three different tasks, which shows the generalization toward different medical segmentation tasks and different image modalities. Comparing against DPM-based model proposed specifically for the brain tumor segmentation, i.e., EnsemDiff, it improves 2.3% on Dice and 2.4% on IoU, which indicates the effectiveness of our unique techniques, i.e, dynamic conditioning and FF-Parser.

Figure 3 shows several typical examples generated by our MedSegDiff and other SOTA methods. It can be seen the target lesions/tissues are all ambiguous on the images so that they are hard to be recognized by human eyes. Comparing with these computer-aided methods, it is obvious that the segmentation maps generated by the proposed method are

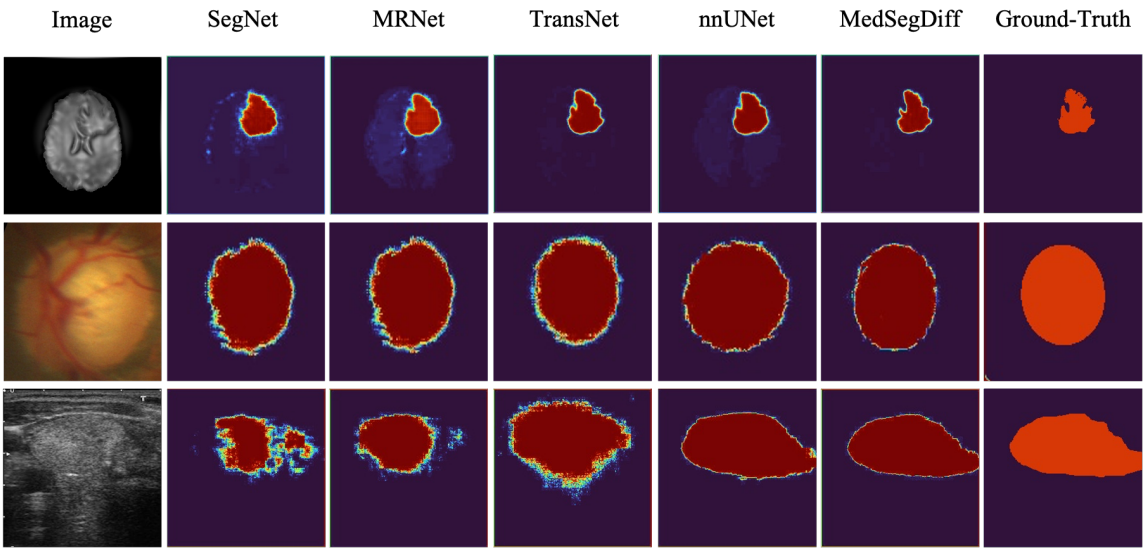

Figure 3: The visual comparison of Top-4 general medical image segmentation methods in Table 1. From top to down are brain-tumor segmentation, optic-cup segmentation and thyroid nodule segmentation, respectively.

more accurate than the other methods, especially for the ambiguous regions. To be benefited from DPM together with the proposed dynamic conditioning and FF-Parser, it can better localize and calibrate the segmentation on the low-contrast or ambiguous images.

3.4. Ablation Study

We do comprehensive ablation study to verify the effectiveness of the proposed dynamic conditioning and FF-Parser. The results are shown in Table 2, where Dy-Cond denotes dynamic conditioning. We evaluate the performance by Dice score(%) on all three tasks. From the table, we can see Dy-Cond gains considerable improvements over vanilla DPM. On the case which the region localization is important, i.e., optic-cup segmentation, it improves 2.1%. On the cases which the images are low-contrast, like brain tumor and thyroid nodule segmentation, it improves 1.6% and 1.8% respectively. It shows Dy-Cond is a generally effective strategy on DPM for both of the cases. FF-Parser which established over Dy-Cond mitigates the high-frequency noises thus further optimize the segmentation results. It helps MedSegDiff further improve near 1% performance and achieve the best on all three tasks.

4. Conclusion

In this paper, we provided a scheme for DPM-based general medical image segmentation, named MedSegDiff. We propose two novel techniques to promise the performance of it, i.e., the dynamic conditional encoding and FF-Parser. The comparison experiments are conducted on three medical image segmentation tasks with different image modalities, which

Table 1: The comparison of MedSegDiff with SOTA segmentation methods. Best results are denoted as **bold**. The grey background denotes the methods are proposed for that/these particular tasks.

	Optic-Cup		Brain-Turmor		Thyroid Nodule	
	Dice	IoU	Dice	IoU	Dice	IoU
ResUnet	80.1	72.3	-	-	-	-
BEAL	83.5	74.1	-	-	-	-
TransBTS	-	-	87.6	78.3	-	-
EnsemDiff	-	-	88.7	80.9	-	-
MTSeg	-	-	-	-	82.3	75.2
UltraUNet	-	-	-	-	84.5	76.2
CENet	78.6	69,4	76.2	68.9	78.9	71.2
MRNet	84.2	75.1	83.4	75.6	80.4	73.4
SegNet	80.4	70.7	80.2	72.9	81.7	74.5
nnUNet	84.9	75.1	88.2	80.4	84.2	76.2
TransUNet	85.6	75.9	86.6	79.0	83.5	75.1
MedSegDiff-S	81.2	71.7	82.3	73.6	80.8	73.7
MedSegDiff-B	85.9	76.2	88.9	81.2	84.8	76.4
MedSegDiff-L	86.9	78.5	89.9	82.3	86.1	79.6
MedSegDiff++	**87.5**	**79.1**	**90.5**	**82.8**	**86.6**	**80.2**

Table 2: An ablation study on dynamic condition encoding and FF-Parser. Dice score(%) is used as the metric.

Dy-Cond	FF-Parser	OpticCup	BrainTumor	ThyroidNodule
		84.6	88.2	84.1
✓		86.7	89.8	85.9
✓	✓	**87.5**	**90.5**	**86.6**

shows our model outperforms previous SOTA. As the first DPM application in general medical image segmentation, we believe MedSegDiff will serve as an essential benchmark for future research.

References

Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017.

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*, 2021.

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.

Chen Chu, Jihui Zheng, and Yong Zhou. Ultrasonic thyroid nodule detection method based on u-net network. *Computer Methods and Programs in Biomedicine*, 199:105906, 2021.

Stefan Elfwing, Eiji Uchibe, and Kenji Doya. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural Networks*, 107:3–11, 2018.

Huihui Fang, Fei Li, Huazhu Fu, Xu Sun, Xingxing Cao, Jaemin Son, Shuang Yu, Menglu Zhang, Chenglang Yuan, Cheng Bian, et al. Refuge2 challenge: Treasure for multi-domain learning in glaucoma assessment. *arXiv preprint arXiv:2202.08994*, 2022.

Haifan Gong, Guanqi Chen, Ranran Wang, Xiang Xie, Mingzhi Mao, Yizhou Yu, Fei Chen, and Guanbin Li. Multi-task learning for thyroid nodule segmentation with thyroid region prior. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 257–261. IEEE, 2021.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

Zaiwang Gu, Jun Cheng, Huazhu Fu, Kang Zhou, Huaying Hao, Yitian Zhao, Tianyang Zhang, Shenghua Gao, and Jiang Liu. Ce-net: Context encoder network for 2d medical image segmentation. *IEEE transactions on medical imaging*, 38(10):2281–2292, 2019.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

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.

Wei Ji, Shuang Yu, Junde Wu, Kai Ma, Cheng Bian, Qi Bi, Jingjing Li, Hanruo Liu, Li Cheng, and Yefeng Zheng. Learning calibrated medical image segmentation via multi-rater agreement modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12341–12351, 2021.

Chenyang Liu, Rui Zhao, and Zhenwei Shi. Remote-sensing image captioning based on multilayer aggregated transformer. *IEEE Geoscience and Remote Sensing Letters*, 19: 1–5, 2022.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021.

Lina Pedraza, Carlos Vargas, Fabián Narváez, Oscar Durán, Emma Muñoz, and Eduardo Romero. An open access thyroid ultrasound image database. In *10th International symposium on medical information processing and analysis*, volume 9287, pages 188–193. SPIE, 2015.

Ioannis Pitas. *Digital image processing algorithms and applications*. John Wiley & Sons, 2000.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022a.

Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022b.

Shujun Wang, Lequan Yu, Kang Li, Xin Yang, Chi-Wing Fu, and Pheng-Ann Heng. Boundary and entropy-driven adversarial learning for fundus image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 102–110. Springer, 2019.

Wenxuan Wang, Chen Chen, Meng Ding, Hong Yu, Sen Zha, and Jiangyun Li. Transbts: Multimodal brain tumor segmentation using transformer. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 109–119. Springer, 2021.

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.

Jay Whang, Mauricio Delbracio, Hossein Talebi, Chitwan Saharia, Alexandros G Dimakis, and Peyman Milanfar. Deblurring via stochastic refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16293–16303, 2022.

Julia Wolleb, Robin Sandkühler, Florentin Bieder, Philippe Valmaggia, and Philippe C Cattin. Diffusion models for implicit image segmentation ensembles. *arXiv preprint arXiv:2112.03145*, 2021.

Shuang Yu, Di Xiao, Shaun Frost, and Yogesan Kanagasingam. Robust optic disc and cup segmentation with deep learning for glaucoma detection. *Computerized Medical Imaging and Graphics*, 74:61–71, 2019.

Rui Zhao and Zhenwei Shi. Text-to-remote-sensing-image generation with structured generative adversarial networks. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2021.

Rui Zhao, Zhenwei Shi, and Zhengxia Zou. High-resolution remote sensing image captioning based on structured attention. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14, 2021.