MedSegDiff: Medical Image Segmentation with Diffusion Probabilistic Model

Junde Wu
Baidu Research

Huihui Fang
Baidu Research

Yu Zhang
Mind Vogue

Yehui Yang
Baidu Research

Haoyi Xiong
Baidu Research

Yanwu Xu
Baidu Research

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.

Index Terms—diffusion probabilistic model, medical image segmentation, brain tumor, optic cup, thyroid nodule

I. 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) [1] to the recent vision transformers (ViT) [2], [3].

Very recently, diffusion probabilistic model (DPM) [4] gained popularity as a powerful class of generative models, that is able to generate images with high diversity and synthesis quality. Recent large diffusion models, such as DALL-E2 [5], Imagen [6], and Stable Diffusion [7] 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 [8] and deblurring [9].

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.

II. METHOD

We design our model based on diffusion model mentioned in [4]. 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:

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

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 [4].

A. 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.

B. 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:

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

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 [10]. 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.

C. Training and Architecture

MedSegDiff is trained following the standard process of DPM [4]. Specifically, the loss can be represented as:

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

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 [11], which we implement it with a ResNet encoder following a UNet decoder. The detailed network setting is following [12]. I and x_t are encoded with two individual
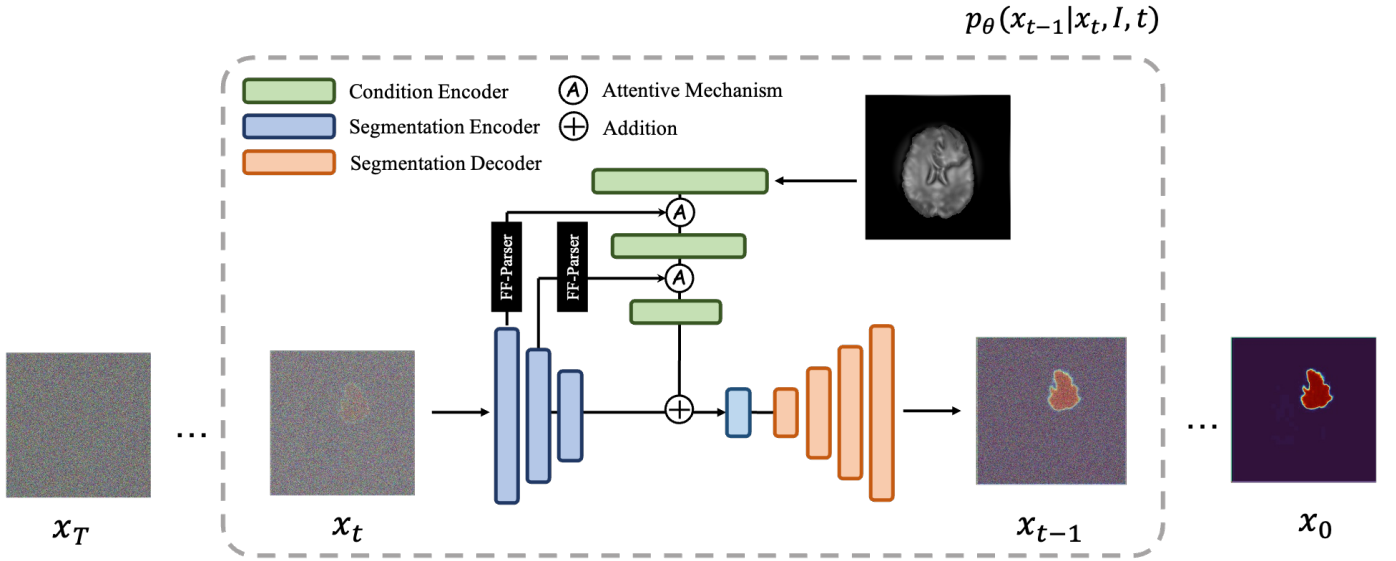
$$p_\theta(x_{t-1}|x_t, I, t)$$

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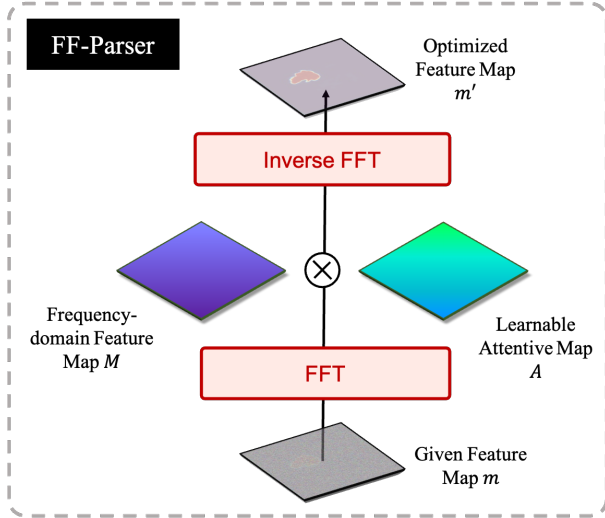

Fig. 2. An illustration of FF-Parser

encoders. The encoder is consisted of three convolution stages. Each stage contains several 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 [13] 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.

III. EXPERIMENTS

A. 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 [14], BraTs-2021 dataset [15] and DDTI dataset [16], 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.

B. 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 [4], [12]. 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 [17] 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}.

C. Main Results

We compare with SOTA segmentation methods proposed for the three specific tasks and general medical image seg-
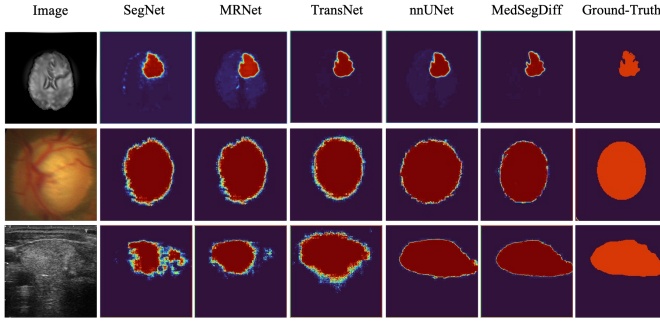
Fig. 3. The visual comparison of Top-4 general medical image segmentation methods in Table I. From top to down are brain-tumor segmentation, optic-cup segmentation and thyroid nodule segmentation, respectively.

mentation methods. The main results are shown in Table I. In the table, ResUnet [11] and BEAL [18] are proposed for optic disc/cup segmentation, TransBTS [3] and EnsemDiff [19] are proposed for the brain tumor segmentation, MTSeg [20] and UltraUNet [21] are proposed for the Thyroid Nodule segmentation, CENet [22], MRNet [1], SegNet [23], nnUNet [24] and TransUNet [2] are proposed for the general medical image segmentation. We evaluate the segmentation performance by Dice score and IoU. The compared methods are all implemented with their default setting.

In Table I, 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 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.

TABLE I
THE COMPARISON OF MEDSEGDIFF WITH SOTA SEGMENTATION METHODS. BEST RESULTS ARE DENOTED AS **BOLD**.

	Optic-Cup		Brain-Tumor		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**

D. 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 II, 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.

TABLE II
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**

IV. 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 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

[1] W. Ji, S. Yu, J. Wu, K. Ma, C. Bian, Q. Bi, J. Li, H. Liu, L. Cheng, and Y. Zheng, "Learning calibrated medical image segmentation via multi-rater agreement modeling," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12 341–12 351.

[2] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "Transunet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021.

[3] W. Wang, C. Chen, M. Ding, H. Yu, S. Zha, and J. Li, "Transbts: Multimodal brain tumor segmentation using transformer," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2021, pp. 109–119.

[4] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020.

[5] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," *arXiv preprint arXiv:2204.06125*, 2022.

[6] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, S. K. S. Ghasemipour, B. K. Ayan, S. S. Mahdavi, R. G. Lopes *et al.*, "Photorealistic text-to-image diffusion models with deep language understanding," *arXiv preprint arXiv:2205.11487*, 2022.

[7] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 684–10 695.

[8] C. Saharia, J. Ho, W. Chan, T. Salimans, D. J. Fleet, and M. Norouzi, "Image super-resolution via iterative refinement," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[9] J. Whang, M. Delbracio, H. Talebi, C. Saharia, A. G. Dimakis, and P. Milanfar, "Deblurring via stochastic refinement," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 293–16 303.

[10] I. Pitas, *Digital image processing algorithms and applications*. John Wiley & Sons, 2000.

[11] S. Yu, D. Xiao, S. Frost, and Y. Kanagasingam, "Robust optic disc and cup segmentation with deep learning for glaucoma detection," *Computerized Medical Imaging and Graphics*, vol. 74, pp. 61–71, 2019.

[12] A. Q. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," in *International Conference on Machine Learning*. PMLR, 2021, pp. 8162–8171.

[13] S. Elfwing, E. Uchibe, and K. Doya, "Sigmoid-weighted linear units for neural network function approximation in reinforcement learning," *Neural Networks*, vol. 107, pp. 3–11, 2018.

[14] H. Fang, F. Li, H. Fu, X. Sun, X. Cao, J. Son, S. Yu, M. Zhang, C. Yuan, C. Bian *et al.*, "Refuge2 challenge: Treasure for multi-domain learning in glaucoma assessment," *arXiv preprint arXiv:2202.08994*, 2022.

[15] U. Baid, S. Ghodasara, S. Mohan, M. Bilello, E. Calabrese, E. Colak, K. Farahani, J. Kalpathy-Cramer, F. C. Kitamura, S. Pati *et al.*, "The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification," *arXiv preprint arXiv:2107.02314*, 2021.

[16] L. Pedraza, C. Vargas, F. Narváez, O. Durán, E. Muñoz, and E. Romero, "An open access thyroid ultrasound image database," in *10th International symposium on medical information processing and analysis*, vol. 9287. SPIE, 2015, pp. 188–193.

[17] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[18] S. Wang, L. Yu, K. Li, X. Yang, C.-W. Fu, and P.-A. Heng, "Boundary and entropy-driven adversarial learning for fundus image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 102–110.

[19] J. Wolleb, R. Sandkühler, F. Bieder, P. Valmaggia, and P. C. Cattin, "Diffusion models for implicit image segmentation ensembles," *arXiv preprint arXiv:2112.03145*, 2021.

[20] H. Gong, G. Chen, R. Wang, X. Xie, M. Mao, Y. Yu, F. Chen, and G. Li, "Multi-task learning for thyroid nodule segmentation with thyroid region prior," in *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2021, pp. 257–261.

[21] C. Chu, J. Zheng, and Y. Zhou, "Ultrasonic thyroid nodule detection method based on u-net network," *Computer Methods and Programs in Biomedicine*, vol. 199, p. 105906, 2021.

[22] Z. Gu, J. Cheng, H. Fu, K. Zhou, H. Hao, Y. Zhao, T. Zhang, S. Gao, and J. Liu, "Ce-net: Context encoder network for 2d medical image segmentation," *IEEE transactions on medical imaging*, vol. 38, no. 10, pp. 2281–2292, 2019.

[23] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[24] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnu-net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature methods*, vol. 18, no. 2, pp. 203–211, 2021.