




BerDiff: Conditional Bernoulli Diffusion Model for Medical Image Segmentation

Tao Chen¹, Chenhui Wang¹, and Hongming Shan^{1,2,3} 

¹ Institute of Science and Technology for Brain-Inspired Intelligence and MOE Frontiers Center for Brain Science, Fudan University, Shanghai, China

hmsan@fudan.edu.cn

² Key Laboratory of Computational Neuroscience and Brain-Inspired Intelligence (Fudan University), Ministry of Education, Shanghai, China

³ Shanghai Center for Brain Science and Brain-Inspired Technology, Shanghai, China

Abstract. Medical image segmentation is a challenging task with inherent ambiguity and high uncertainty attributed to factors such as unclear tumor boundaries and multiple plausible annotations. The *accuracy* and *diversity* of segmentation masks are both crucial for providing valuable references to radiologists in clinical practice. While existing diffusion models have shown strong capacities in various visual generation tasks, it is still challenging to deal with discrete masks in segmentation. To achieve accurate and diverse medical image segmentation masks, we propose a novel conditional **Bernoulli Diffusion** model for medical image segmentation (**BerDiff**). Instead of using the Gaussian noise, we first propose to use the Bernoulli noise as the diffusion kernel to enhance the capacity of the diffusion model for binary segmentation tasks, resulting in more accurate segmentation masks. Second, by leveraging the stochastic nature of the diffusion model, our **BerDiff** randomly samples the initial Bernoulli noise and intermediate latent variables multiple times to produce a range of diverse segmentation masks, which can highlight salient regions of interest that can serve as a valuable reference for radiologists. In addition, our **BerDiff** can efficiently sample sub-sequences from the overall trajectory of the reverse diffusion, thereby speeding up the segmentation process. Extensive experimental results on two medical image segmentation datasets with different modalities demonstrate that our **BerDiff** outperforms other recently published state-of-the-art methods. Source code is made available at <https://github.com/takimailto/BerDiff>.

Keywords: Conditional diffusion · Bernoulli noise · Medical image segmentation

1 Introduction

Medical image segmentation plays a crucial role in enabling better diagnosis, surgical planning, and image-guided surgery [8]. The inherent ambiguity and high

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43901-8_47.

uncertainty of medical images pose significant challenges [5] for accurate segmentation, attributed to factors such as unclear tumor boundaries in brain Magnetic Resonance Imaging (MRI) images and multiple plausible annotations of lung nodule in Computed Tomography (CT) images. Existing medical image segmentation methods typically provide a single, deterministic, most likely hypothesis mask, which may lead to misdiagnosis or sub-optimal treatment. Therefore, providing *accurate* and *diverse* segmentation masks as valuable references [17] for radiologists is crucial in clinical practice.

Recently, diffusion models [10] have shown strong capacities in various visual generation tasks [21, 22]. However, how to better deal with discrete segmentation tasks needs further consideration. Although many researches [1, 26] have combined diffusion model with segmentation tasks, all these methods do not take full account of the discrete characteristic of segmentation task and still use Gaussian noise as their diffusion kernel.

To achieve accurate and diverse segmentation masks, we propose a novel Conditional **B**ernoulli **D**iffusion model for medical image segmentation (**BerDiff**). Instead of using the Gaussian noise, we first propose to use the Bernoulli noise as the diffusion kernel to enhance the capacity of the diffusion model for segmentation, resulting in more accurate segmentation masks. Moreover, by leveraging the stochastic nature of the diffusion model, our **BerDiff** randomly samples the initial Bernoulli noise and intermediate latent variables multiple times to produce a range of diverse segmentation masks, highlighting salient regions of interest (ROI) that can serve as a valuable reference for radiologists. In addition, our **BerDiff** can efficiently sample sub-sequences from the overall trajectory of the reverse diffusion based on the rationale behind the Denoising Diffusion Implicit Models (DDIM) [25], thereby speeding up the segmentation process.

The contributions of this work are summarized as follows. 1) Instead of using the Gaussian noise, we propose a novel conditional diffusion model based on the Bernoulli noise for discrete binary segmentation tasks, achieving accurate and diverse medical image segmentation masks. 2) Our **BerDiff** can efficiently sample sub-sequences from the overall trajectory of the reverse diffusion, thereby speeding up the segmentation process. 3) Experimental results on LIDC-IDRI and BRATS 2021 datasets demonstrate that our **BerDiff** outperforms other state-of-the-art methods.

2 Methodology

In this section, we first describe the problem definitions, and then demonstrate the Bernoulli forward and diverse reverse processes of our **BerDiff**, as shown in Fig. 1. Finally, we provide an overview of the training and sampling procedures.

2.1 Problem Definition

Let us assume that $\mathbf{x} \in \mathbb{R}^{H \times W \times C}$ denotes the input medical image with a spatial resolution of $H \times W$ and C channels. The ground-truth mask is represented

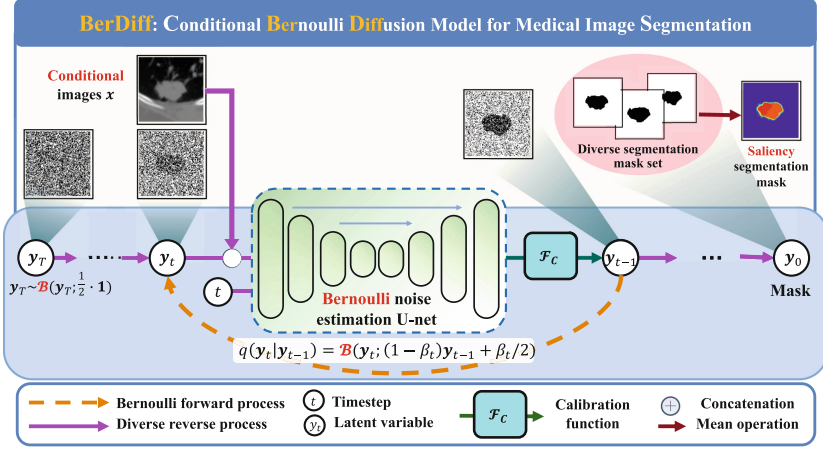


Fig. 1. Illustration of Bernoulli forward and diverse reverse processes of our **BerDiff**.

as $y_0 \in \{0, 1\}^{H \times W}$, where 0 represents background while 1 ROI. Inspired by diffusion-based models such as denoising diffusion probabilistic model (DDPM) and DDIM, we propose a novel conditional Bernoulli diffusion model, which can be represented as $p_\theta(y_0 | x) := \int p_\theta(y_{0:T} | x) dy_{1:T}$, where y_1, \dots, y_T are latent variables of the same size as the mask y_0 . For medical binary segmentation tasks, the diverse reverse process of our **BerDiff** starts from the initial Bernoulli noise $y_T \sim \mathcal{B}(y_T; \frac{1}{2} \cdot \mathbf{1})$ and progresses through intermediate latent variables constrained by the input medical image x to produce segmentation masks, where $\mathbf{1}$ denotes an all-ones matrix of the size $H \times W$.

2.2 Bernoulli Forward Process

In previous generation-related diffusion models, Gaussian noise is progressively added with increasing timestep t . However, for segmentation tasks, the ground-truth masks are represented by discrete values. To address this, our **BerDiff** gradually adds more Bernoulli noise using a noise schedule β_1, \dots, β_T , as shown in Fig. 1. The Bernoulli forward process $q(y_{1:T} | y_0)$ of our **BerDiff** is a Markov chain, which can be represented as:

$$q(y_{1:T} | y_0) := \prod_{t=1}^T q(y_t | y_{t-1}), \quad (1)$$

$$q(y_t | y_{t-1}) := \mathcal{B}(y_t; (1 - \beta_t)y_{t-1} + \beta_t/2), \quad (2)$$

where \mathcal{B} denotes the Bernoulli distribution with the probability parameters $(1 - \beta_t)y_{t-1} + \beta_t/2$. Using the notation $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{\tau=1}^t \alpha_\tau$, we can efficiently sample y_t at an arbitrary timestep t in closed form:

$$q(y_t | y_0) = \mathcal{B}(y_t; \bar{\alpha}_t y_0 + (1 - \bar{\alpha}_t)/2). \quad (3)$$

Algorithm 1. Training

```

repeat
   $(\mathbf{x}, \mathbf{y}_0) \sim q(\mathbf{x}, \mathbf{y}_0)$ 
   $t \sim \text{Uniform}(\{1, \dots, T\})$ 
   $\epsilon \sim \mathcal{B}(\epsilon; (1 - \bar{\alpha}_t)/2)$ 
   $\mathbf{y}_t = \mathbf{y}_0 \oplus \epsilon$ 
  Calculate Eq. (4)
  Estimate  $\hat{\epsilon}(\mathbf{y}_t, t, \mathbf{x})$ 
  Calculate Eq. (6)
  Take gradient descent on  $\nabla_{\theta}(\mathcal{L}_{\text{Total}})$ 
until converged

```

Algorithm 2. Sampling

```

 $\mathbf{y}_T \sim \mathcal{B}(\mathbf{y}_T; \frac{1}{2} \cdot \mathbf{1})$ 
for  $t = T$  to 1 do
   $\hat{\mu}(\mathbf{y}_t, t, \mathbf{x}) = \mathcal{F}_C(\mathbf{y}_t, \hat{\epsilon}(\mathbf{y}_t, t, \mathbf{x}))$ 
  For DDPM:
     $\mathbf{y}_{t-1} \sim \mathcal{B}(\mathbf{y}_{t-1}; \hat{\mu}(\mathbf{y}_t, t, \mathbf{x}))$ 
  For DDIM:
     $\mathbf{y}_{t-1} \sim \mathcal{B}(\mathbf{y}_{t-1}; \sigma_t \mathbf{y}_t + (\bar{\alpha}_{t-1} - \sigma_t \bar{\alpha}_t) |\mathbf{y}_t - \hat{\epsilon}(\mathbf{y}_t, t, \mathbf{x})| + ((1 - \bar{\alpha}_{t-1}) - (1 - \bar{\alpha}_t) \sigma_t) / 2)$ 
end for
return  $\mathbf{y}_0$ 

```

To ensure that the objective function described in Sect. 2.4 is tractable and easy to compute, we use the sampled Bernoulli noise $\epsilon \sim \mathcal{B}(\epsilon; \frac{1 - \bar{\alpha}_t}{2} \cdot \mathbf{1})$ to reparameterize \mathbf{y}_t of Eq. (3) as $\mathbf{y}_0 \oplus \epsilon$, where \oplus denotes the logical operation of “exclusive or (XOR)”. Additionally, let \odot denote elementwise product, and $\text{Norm}(\cdot)$ denote normalizing the input data along the channel dimension and then returning the second channel. The concrete Bernoulli posterior can be represented as:

$$q(\mathbf{y}_{t-1} \mid \mathbf{y}_t, \mathbf{y}_0) = \mathcal{B}(\mathbf{y}_{t-1}; \theta_{\text{post}}(\mathbf{y}_t, \mathbf{y}_0)), \quad (4)$$

where $\theta_{\text{post}}(\mathbf{y}_t, \mathbf{y}_0) = \text{Norm}([\alpha_t[1 - \mathbf{y}_t, \mathbf{y}_t] + \frac{1 - \bar{\alpha}_t}{2}] \odot [\bar{\alpha}_{t-1}[1 - \mathbf{y}_0, \mathbf{y}_0] + \frac{1 - \bar{\alpha}_{t-1}}{2}])$.

2.3 Diverse Reverse Process

The diverse reverse process $p_{\theta}(\mathbf{y}_{0:T})$ can also be viewed as a Markov chain that starts from the Bernoulli noise $\mathbf{y}_T \sim \mathcal{B}(\mathbf{y}_T; \frac{1}{2} \cdot \mathbf{1})$ and progresses through intermediate latent variables constrained by the input medical image \mathbf{x} to produce diverse segmentation masks, as shown in Fig. 1. The concrete diverse reverse process of our BerDiff can be represented as:

$$p_{\theta}(\mathbf{y}_{0:T} \mid \mathbf{x}) := p(\mathbf{y}_T) \prod_{t=1}^T p_{\theta}(\mathbf{y}_{t-1} \mid \mathbf{y}_t, \mathbf{x}), \quad (5)$$

$$p_{\theta}(\mathbf{y}_{t-1} \mid \mathbf{y}_t, \mathbf{x}) := \mathcal{B}(\mathbf{y}_{t-1}; \hat{\mu}(\mathbf{y}_t, t, \mathbf{x})). \quad (6)$$

Specifically, we utilize the estimated Bernoulli noise $\hat{\epsilon}(\mathbf{y}_t, t, \mathbf{x})$ of \mathbf{y}_t to parameterize $\hat{\mu}(\mathbf{y}_t, t, \mathbf{x})$ via a calibration function \mathcal{F}_C , as follows:

$$\hat{\mu}(\mathbf{y}_t, t, \mathbf{x}) = \mathcal{F}_C(\mathbf{y}_t, \hat{\epsilon}(\mathbf{y}_t, t, \mathbf{x})) = \theta_{\text{post}}(\mathbf{y}_t, |\mathbf{y}_t - \hat{\epsilon}(\mathbf{y}_t, t, \mathbf{x})|), \quad (7)$$

where $|\cdot|$ denotes the absolute value operation. The calibration function aims to calibrate the latent variable \mathbf{y}_t to a less noisy latent variable \mathbf{y}_{t-1} in two steps: 1) estimating the segmentation mask \mathbf{y}_0 by computing the absolute deviation between \mathbf{y}_t and the estimated noise $\hat{\epsilon}$; and 2) estimating the distribution of \mathbf{y}_{t-1} by calculating the Bernoulli posterior, $p(\mathbf{y}_{t-1} \mid \mathbf{y}_t, \mathbf{y}_0)$, using Eq. (4).

2.4 Detailed Procedure

Here, we provide an overview of the training and sampling procedure in Algorithms 1 and 2. During the training phase, given an image and mask data pair $\{\mathbf{x}, \mathbf{y}_0\}$, we sample a random timestep t from a uniform distribution $\{1, \dots, T\}$, which is used to sample the Bernoulli noise ϵ .

We then use ϵ to sample \mathbf{y}_t from $q(\mathbf{y}_t | \mathbf{y}_0)$, which allows us to obtain the Bernoulli posterior $q(\mathbf{y}_{t-1} | \mathbf{y}_t, \mathbf{y}_0)$. We pass the estimated Bernoulli noise $\hat{\epsilon}(\mathbf{y}_t, t, \mathbf{x})$ through the calibration function \mathcal{F}_C to parameterize $p_\theta(\mathbf{y}_{t-1} | \mathbf{y}_t, \mathbf{x})$. Based on the variational upper bound on the negative log-likelihood in previous diffusion models [3], we adopt Kullback-Leibler (KL) divergence and binary cross-entropy (BCE) loss to optimize our **BerDiff** as follows:

$$\mathcal{L}_{\text{KL}} = \mathbb{E}_{q(\mathbf{x}, \mathbf{y}_0)} \mathbb{E}_{q(\mathbf{y}_t | \mathbf{y}_0)} [D_{\text{KL}}[q(\mathbf{y}_{t-1} | \mathbf{y}_t, \mathbf{y}_0) \| p_\theta(\mathbf{y}_{t-1} | \mathbf{y}_t, \mathbf{x})]], \quad (8)$$

$$\mathcal{L}_{\text{BCE}} = -\mathbb{E}_{(\epsilon, \hat{\epsilon})} \sum_{i,j}^{H,W} [\epsilon_{i,j} \log \hat{\epsilon}_{i,j} + (1 - \epsilon_{i,j}) \log (1 - \hat{\epsilon}_{i,j})]. \quad (9)$$

Finally, the overall objective function is presented as: $\mathcal{L}_{\text{Total}} = \mathcal{L}_{\text{KL}} + \lambda_{\text{BCE}} \mathcal{L}_{\text{BCE}}$, where λ_{BCE} is set to 1 in our experiments.

During the sampling phase, our **BerDiff** first samples the initial latent variable \mathbf{y}_T , followed by iterative calculation of the probability parameters of \mathbf{y}_{t-1} for different t . In Algorithm 2, we present two different sampling strategies from DDPM and DDIM for the latent variable \mathbf{y}_{t-1} . Finally, our **BerDiff** is capable of producing diverse segmentation masks. By taking the mean of these masks, we can further obtain a saliency segmentation mask to highlight salient ROI that can serve as a valuable reference for radiologists. Note that our **BerDiff** has a novel parameterization technique, *i.e.* calibration function, to estimate the Bernoulli noise of \mathbf{y}_t , which is different from previous works [3, 11, 24].

3 Experiment

3.1 Experimental Setup

Dataset and Preprocessing. The data used in this experiment are obtained from LIDC-IDRI [2, 7] and BRATS 2021 [4] datasets. LIDC-IDRI contains 1,018 lung CT scans with plausible segmentation masks annotated by four radiologists. We adopt a standard preprocessing pipeline for lung CT scans and the train-validation-test partition as in previous work [5, 15, 23]. BRATS 2021 consists of four different sequence (T1, T2, FLAIR, T1CE) MRI images for each patient. All 3D scans are sliced into axial slices and discarded the bottom 80 and top 26 slices. Note that we treat the original four types of brain tumors as one type following previous work [25], converting the multi-target segmentation problem into binary. Our training set includes 55,174 2D images scanned from 1,126 patients, and the test set comprises 3,991 2D images scanned from 125 patients. Finally, the sizes of images from LIDC-IDRI and BRATS 2021 are resized to a resolution of 128×128 and 224×224 , respectively.

Table 1. Ablation results of hyperparameters on LIDC-IDRI.

Loss	Estimation Target	GED				HM-IoU
		16	8	4	1	
\mathcal{L}_{KL}	Bernoulli noise	0.332	0.365	0.430	0.825	0.517
\mathcal{L}_{BCE}	Bernoulli noise	<u>0.251</u>	<u>0.287</u>	<u>0.359</u>	<u>0.785</u>	<u>0.566</u>
$\mathcal{L}_{BCE} + \mathcal{L}_{KL}$	Bernoulli noise	0.249	0.287	0.358	0.775	0.575
$\mathcal{L}_{BCE} + \mathcal{L}_{KL}$	Ground-truth mask	0.277	0.317	0.396	0.866	0.509

‡ The best and second best results are highlighted in **bold** and underlined, respectively.

Table 2. Ablation results of diffusion kernel on LIDC-IDRI.

Training Iteration	Diffusion Kernel	GED				HM-IoU
		16	8	4	1	
21,000	Gaussian	0.671	0.732	0.852	1.573	0.020
	Bernoulli	0.252	0.287	0.358	0.775	0.575
86,500	Gaussian	0.251	0.282	0.345	0.719	0.587
	Bernoulli	0.238	0.271	0.340	0.748	0.596

Implementation Details. We implement all the methods with the PyTorch library and train the models on NVIDIA V100 GPUs. All the networks are trained using the AdamW [19] optimizer with a mini-batch size of 32. The initial learning rate is set to 1×10^{-4} for BRATS 2021 and 5×10^{-5} for LIDC-IDRI. The Bernoulli noise estimation U-net network in Fig. 1 of our **BerDiff** is the same as previous diffusion-based models [20]. We employ a linear noise schedule for $T = 1000$ timesteps for all the diffusion models. And we use the sub-sequence sampling strategy of DDIM to accelerate the segmentation process. During mini-batch training of LIDC-IDRI, our **BerDiff** learns diverse expertise by randomly sampling one from four annotated segmentation masks for each image. Four metrics are used for performance evaluation, including Generalized Energy Distance (GED), Hungarian-Matched Intersection over Union (HM-IoU), Soft-Dice and Dice coefficient. We compute GED using varying numbers of segmentation samples (1, 4, 8, and 16), HM-IoU and Soft-Dice using 16 samples.

3.2 Ablation Study

We start by conducting ablation experiments to demonstrate the effectiveness of different losses and estimation targets, as shown in Table 1. All experiments are trained for 21,000 training iterations on LIDC-IDRI. We first explore the selection of losses in the top three rows. We find that the combination of KL divergence and BCE loss can achieve the best performance. Then, we explore the selection of estimation targets in the bottom two rows. We observe that estimating Bernoulli noise, instead of directly estimating the ground-truth mask, is

Table 3. Results on LIDC-IDRI.

Methods	GED 16	HM-IoU 16	Soft-Dice 16
Prob.U-net [15]	0.320 ± 0.03	0.500 ± 0.03	-
Hprob.U-net [16]	0.270 ± 0.01	0.530 ± 0.01	0.624 ± 0.01
CAR [14]	0.264 ± 0.00	0.592 ± 0.01	0.633 ± 0.00
JPro.U-net [29]	0.260 ± 0.00	0.585 ± 0.00	-
PixelSeg [28]	0.260 ± 0.00	0.587 ± 0.01	-
SegDiff [1]	0.248 ± 0.01	0.585 ± 0.00	0.637 ± 0.01
MedSegDiff [26]	0.420 ± 0.03	0.413 ± 0.03	0.453 ± 0.02
Zhang <i>et al.</i> [27]	0.400	0.534	0.599
Ji <i>et al.</i> [13]	0.658	0.447	0.616
Liao <i>et al.</i> [18]	0.593	0.453	0.587
BerDiff (Ours)	0.238 ± 0.01	0.596 ± 0.00	0.644 ± 0.00

Table 4. Results on BRATS 2021.

Methods	Dice
nnU-net [12]	88.2
TransU-net [6]	88.6
Swin UNETR [9]	89.0
U-net [‡]	89.2
SegDiff [1]	<u>89.3</u>
BerDiff (Ours)	89.7

[‡] The U-net has the same architecture as the noise estimation network in our BerDiff and previous diffusion-based models.

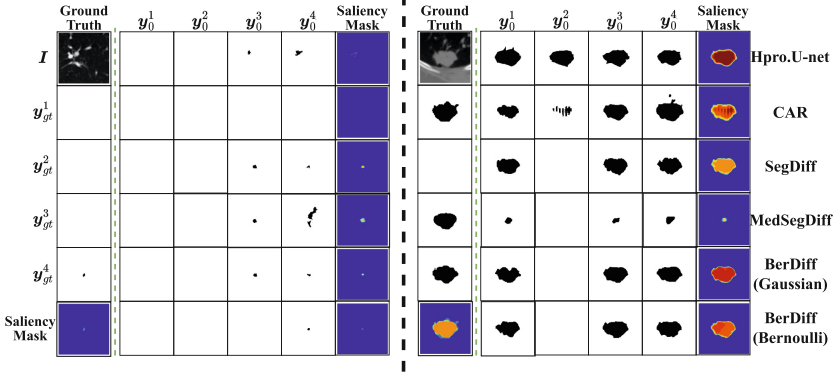


Fig. 2. Diverse segmentation masks and the corresponding saliency mask of two lung nodules randomly selected in LIDC-IDRI. y_0^i and y_{gt}^i refer to the i -th generated and ground-truth segmentation masks, respectively. Saliency Mask is the mean of diverse segmentation masks.

more suitable for our binary segmentation task. All of these findings are consistent with previous works [3, 10].

Here, we conduct ablation experiments on our BerDiff with Gaussian or Bernoulli noise, and the results are shown in Table 2. For discrete segmentation tasks, we find that using Bernoulli noise can produce favorable results when training iterations are limited (*e.g.* 21,000 iterations) and even outperform using Gaussian noise when training iterations are sufficient (*e.g.* 86,500 iterations). We also provide a more detailed performance comparison between Bernoulli- and Gaussian-based diffusion models over training iterations in Fig. S3.

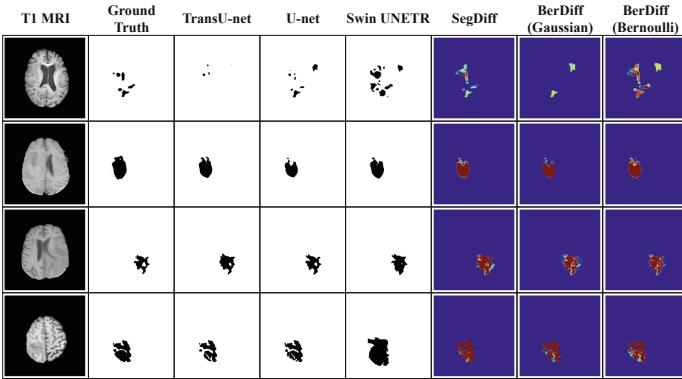


Fig. 3. Segmentation masks of four MRI images randomly selected in BRATS 2021. The segmentation masks of diffusion-based models (SegDiff and ours) presented here are saliency segmentation mask.

3.3 Comparison to Other State-of-the-Art Methods

Results on LIDC-IDRI. Here, we present the quantitative comparison results of LIDC-IDRI in Table 3, and find that our **BerDiff** performs well for discrete segmentation tasks. Probabilistic U-net (Prob.U-net), Hierarchical Prob.U-net (Hprob.U-net), and Joint Prob.U-net (JPro.U-net) use conditional variational autoencoder (cVAE) to accomplish segmentation tasks. Calibrated Adversarial Refinement (CAR) employs generative adversarial networks (GAN) to refine segmentation. PixelSeg is based on autoregressive models, while SegDiff and MedSegDiff are diffusion-based models. There are also methods that attempt to model multi-annotators explicitly [13, 18, 27]. We have the following three observations: 1) diffusion-based methods demonstrate significant superiority over traditional approaches based on VAE, GAN, and autoregression models for discrete segmentation tasks; 2) our **BerDiff** outperforms other diffusion-based models that use Gaussian noise as the diffusion kernel; and 3) our **BerDiff** also outperforms the methods that explicitly model the annotator, striking a good balance between diversity and accuracy. At the same time, we present comparison segmentation results in Fig. 2. Compared to other models, our **BerDiff** can effectively learn diverse expertise, resulting in more diverse and accurate segmentation masks. Especially for small nodules that can create ambiguity, such as the lung nodule on the left, our **BerDiff** approach produces segmentation masks that are more in line with the ground-truth masks.

Results on BRATS 2021. Here, we present the quantitative and qualitative results of BRATS 2021 in Table 4 and Fig. 3, respectively. We conducted a comparative analysis of our **BerDiff** with other models such as nnU-net, transformer-based models like TransU-net and Swin UNETR, as well as diffusion-based methods like SegDiff. First, we find that diffusion-based methods show superior performance compared to traditional U-net and transformer-based

approaches. Besides, the high performance achieved by U-net, which shares the same architecture as our noise estimation network, highlights the effectiveness of the backbone design in diffusion-based models. Moreover, our proposed **BerDiff** surpasses other diffusion-based models that use Gaussian noise as the diffusion kernel. Finally, from Fig. 3, we find that our **BerDiff** segments more accurately on parts that are difficult to recognize by the human eye, such as the tumor in the 3rd row. At the same time, we can also generate diverse plausible segmentation masks to produce a saliency segmentation mask. We note that some of these masks may be false positives, as shown in the 1st row, but they can be filtered out due to low saliency. Please refer to Figs. S1 and S2 for more examples of diverse segmentation masks generated by our **BerDiff**.

4 Conclusion

In this paper, we proposed to use the Bernoulli noise as the diffusion kernel to enhance the capacity of the diffusion model for binary segmentation tasks, achieving accurate and diverse medical image segmentation results. Our **BerDiff** only focuses on binary segmentation tasks and takes much time during the iterative sampling process as other diffusion-based models; *e.g.* our **BerDiff** takes 0.4s to segment one medical image, which is ten times of traditional U-net. In the future, we will extend our **BerDiff** to the multi-target segmentation problem and implement additional strategies for speeding up the segmentation process.

Acknowledgements. This work was supported in part by Natural Science Foundation of Shanghai (No. 21ZR1403600), National Natural Science Foundation of China (No. 62101136), Shanghai Municipal Science and Technology Major Project (No. 2018SHZDZX01) and ZJLab, Shanghai Municipal of Science and Technology Project (No. 20JC1419500), and Shanghai Center for Brain Science and Brain-inspired Technology.

References

1. Amit, T., Shaharbany, T., Nachmani, E., Wolf, L.: SegDiff: image segmentation with diffusion probabilistic models. arXiv preprint [arXiv:2112.00390](https://arxiv.org/abs/2112.00390) (2021)
2. Armato, S.G., III., et al.: The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans. *Med. Phys.* **38**(2), 915–931 (2011)
3. Austin, J., Johnson, D.D., Ho, J., Tarlow, D., van den Berg, R.: Structured denoising diffusion models in discrete state-spaces. *Adv. Neural. Inf. Process. Syst.* **34**, 17981–17993 (2021)
4. Baid, U., et al.: The RSNA-ASNR-MICCAI BraTS 2021 benchmark on brain tumor segmentation and radiogenomic classification. arXiv preprint [arXiv:2107.02314](https://arxiv.org/abs/2107.02314) (2021)
5. Baumgartner, C.F., et al.: PHISeg: capturing uncertainty in medical image segmentation. In: Shen, D., et al. (eds.) *MICCAI 2019. LNCS*, vol. 11765, pp. 119–127. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32245-8_14

6. Chen, J., et al.: Transunet: transformers make strong encoders for medical image segmentation. arXiv preprint [arXiv:2102.04306](https://arxiv.org/abs/2102.04306) (2021)
7. Clark, K., et al.: The cancer imaging archive (TCIA): maintaining and operating a public information repository. *J. Digit. Imaging* **26**(6), 1045–1057 (2013)
8. Haque, I.R.I., Neubert, J.: Deep learning approaches to biomedical image segmentation. *Inform. Med. Unlocked* **18**, 100297 (2020)
9. Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H.R., Xu, D.: Swin UNETR: swin transformers for semantic segmentation of brain tumors in MRI images. In: Crimi, A., Bakas, S. (eds.) *BrainLes 2021*. LNCS, vol. 12962, pp. 272–284. Springer, Cham (2021). https://doi.org/10.1007/978-3-031-08999-2_22
10. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. *Adv. Neural. Inf. Process. Syst.* **33**, 6840–6851 (2020)
11. Hoogeboom, E., Nielsen, D., Jaini, P., Forré, P., Welling, M.: Argmax flows and multinomial diffusion: Learning categorical distributions. *Adv. Neural. Inf. Process. Syst.* **34**, 12454–12465 (2021)
12. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat. Methods* **18**(2), 203–211 (2021)
13. Ji, W., et al.: Learning calibrated medical image segmentation via multi-rater agreement modeling. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 12336–12346 (2021)
14. Kassapis, E., Dikov, G., Gupta, D.K., Nugteren, C.: Calibrated adversarial refinement for stochastic semantic segmentation. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 7037–7047 (2020)
15. Kohl, S., et al.: A probabilistic U-Net for segmentation of ambiguous images. In: *Advances in Neural Information Processing Systems*, vol. 31 (2018)
16. Kohl, S.A., et al.: A hierarchical probabilistic U-Net for modeling multi-scale ambiguities. arXiv preprint [arXiv:1905.13077](https://arxiv.org/abs/1905.13077) (2019)
17. Lenchik, L., et al.: Automated segmentation of tissues using CT and MRI: a systematic review. *Acad. Radiol.* **26**(12), 1695–1706 (2019)
18. Liao, Z., Hu, S., Xie, Y., Xia, Y.: Modeling annotator preference and stochastic annotation error for medical image segmentation. arXiv preprint [arXiv:2111.13410](https://arxiv.org/abs/2111.13410) (2021)
19. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint [arXiv:1711.05101](https://arxiv.org/abs/1711.05101) (2017)
20. Nichol, A.Q., Dhariwal, P.: Improved denoising diffusion probabilistic models. In: *International Conference on Machine Learning*, pp. 8162–8171. PMLR (2021)
21. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint [arXiv:2204.06125](https://arxiv.org/abs/2204.06125) (2022)
22. Saharia, C., et al.: Photorealistic text-to-image diffusion models with deep language understanding. *Adv. Neural. Inf. Process. Syst.* **35**, 36479–36494 (2022)
23. Selvan, R., Faye, F., Middleton, J., Pai, A.: Uncertainty quantification in medical image segmentation with normalizing flows. In: Liu, M., Yan, P., Lian, C., Cao, X. (eds.) *MLMI 2020*. LNCS, vol. 12436, pp. 80–90. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59861-7_9
24. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: *International Conference on Machine Learning*, pp. 2256–2265. PMLR (2015)

25. Wolleb, J., Sandkühler, R., Bieder, F., Valmaggia, P., Cattin, P.C.: Diffusion models for implicit image segmentation ensembles. In: International Conference on Medical Imaging with Deep Learning, pp. 1336–1348. PMLR (2022)
26. Wu, J., Fang, H., Zhang, Y., Yang, Y., Xu, Y.: MedSegDiff: medical image segmentation with diffusion probabilistic model. arXiv preprint [arXiv:2211.00611](https://arxiv.org/abs/2211.00611) (2022)
27. Zhang, L., et al.: Disentangling human error from ground truth in segmentation of medical images. *Adv. Neural. Inf. Process. Syst.* **33**, 15750–15762 (2020)
28. Zhang, W., Zhang, X., Huang, S., Lu, Y., Wang, K.: PixelSeg: pixel-by-pixel stochastic semantic segmentation for ambiguous medical images. In: Proceedings of the 30-th ACM International Conference on Multimedia, pp. 4742–4750 (2022)
29. Zhang, W., Zhang, X., Huang, S., Lu, Y., Wang, K.: A probabilistic model for controlling diversity and accuracy of ambiguous medical image segmentation. In: Proceedings of the 30-th ACM International Conference on Multimedia, pp. 4751–4759 (2022)