Conditional Diffusion Model with Spatial Attention and Latent Embedding for Medical Image Segmentation

Behzad Hejrati[1], Soumyanil Banerjee[1], Carri Glide-Hurst[2], and Ming Dong[1(✉)]

[1] Department of Computer Science, Wayne State University, Detroit, MI, USA
{b.hejrati,s.banerjee,mdong}@wayne.edu
[2] Department of Human Oncology, University of Wisconsin-Madison, Madison, WI, USA
glidehurst@humonc.wisc.edu

Abstract. Diffusion models have been used extensively for high quality image and video generation tasks. In this paper, we propose a novel conditional diffusion model with spatial attention and latent embedding (cDAL) for medical image segmentation. In cDAL, a convolutional neural network (CNN) based discriminator is used at every time-step of the diffusion process to distinguish between the generated labels and the real ones. A spatial attention map is computed based on the features learned by the discriminator to help cDAL generate more accurate segmentation of discriminative regions in an input image. Additionally, we incorporated a random latent embedding into each layer of our model to significantly reduce the number of training and sampling time-steps, thereby making it much faster than other diffusion models for image segmentation. We applied cDAL on 3 publicly available medical image segmentation datasets (MoNuSeg, Chest X-ray and Hippocampus) and observed significant qualitative and quantitative improvements with higher Dice scores and mIoU over the state-of-the-art algorithms. The source code is publicly available at https://github.com/Hejrati/cDAL/.

Keywords: medical image segmentation · diffusion models · generator · discriminator · spatial attention · latent embedding

1 Introduction

Medical image segmentation is a crucial task in clinical practice with applications including disease diagnosis, radiotherapy and surgical treatment planning [1,2]. A major challenge in the segmentation process are the manual annotations performed by a trained clinician, which is a time-consuming process that

B. Hejrati and S. Banerjee—Equal contribution.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-72114-4_20.

M. G. Linguraru et al. (Eds.): MICCAI 2024, LNCS 15009, pp. 202–212, 2024.
https://doi.org/10.1007/978-3-031-72114-4_20

is not scalable. Hence, automation of the segmentation with deep learning algorithms have been a key area of research for the last several years. The algorithms which produced state-of-the-art results for end-to-end 2D and 3D medical image segmentation task include the U-Net [3] and the 3D U-Net [4], respectively.

Diffusion models are a class of generative models where a neural network is trained to remove the noise from an image which was produced during the forward process with a pre-defined noise schedule. This trained neural network is then used in the sampling process to iteratively remove the Gaussian noise from an image and eventually generate high quality samples by starting from pure Gaussian noise [5–7]. Diffusion models generate more diverse images than Generative Adversarial Networks (GANs) and have recently outperformed GANs for the generation of high resolution images [8,9].

Image segmentation with diffusion models is a challenging task due to the deterministic nature of image segmentation as opposed to the stochastic nature of diffusion models. Hence, diffusion models have been used for supervised medical image segmentation tasks to model the distribution of labels resulting from independent annotators of the same image [12,13]. When the segmentation labels are scarce, the semantic representation from intermediate layers of a pretrained diffusion model is used to train a simple pixel-level classifier with the small set of available labels [14]. In [10,11], the image is used as a condition to a diffusion model during the label generation process, which is repeated a few times due to the stochastic nature of diffusion models. The mean of all such label generations is considered as the final segmentation map.

Diffusion models have a common drawback that the sampling procedure to generate the images from pure Gaussian noise is a time-consuming process. This problem was addressed with several interesting ideas such as non-markovian diffusion process [15] and distillation in diffusion models [16,17]. But, faster sampling typically results in degradation of the generated image quality. Hence, there was a need to tackle the generative learning trilemma of achieving fast sampling, higher quality and diversified image samples. This trilemma was addressed with the denoising diffusion GANs [18] by modeling the denoising distribution with a complex multimodal distribution.

In this work, we propose a novel method of using a conditional diffusion model with spatial attention and latent embedding (cDAL) for medical image segmentation. During training, cDAL uses a diffusion model to predict the unperturbed segmentation labels from a noisy label. The image is encoded and passed as a condition to the input of the label diffusion model. During each diffusion time-step, we incorporate a separate discriminator to distinguish between the ground-truth labels and the generated ones. We use the spatial attention map learned by the discriminator [20] to compute attention-based labels as the input of the diffusion model so that it can focus on these discriminative regions during segmentation. We also incorporate a random latent embedding into each layer of the diffusion model to reduce the number of diffusion time-steps in both training and sampling.

Our main contributions are: (i) We incorporated a separate discriminator for each diffusion time-step and guided the diffusion process with the spatial

attention map learned from the discriminator. (ii) We used a random latent embedding for each layer of the diffusion model which helped in reducing both the training and sampling time-steps by modeling the denoising distribution with a complex multimodal distribution. (iii) We performed extensive experiments on two 2D binary (MoNuSeg and chest X-ray) and one 3D multi-class (Hippocampus) public medical image segmentation datasets and observed significant quantitative and qualitative improvements over state-of-the-art methods.

2 Method

In the following sections, we provide a detailed description of each component of our proposed cDAL architecture as shown in Fig. 1.

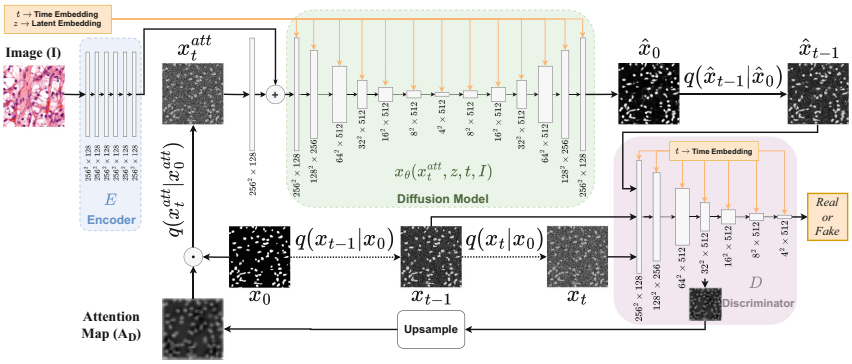

Fig. 1. Our proposed conditional diffusion model with spatial attention and latent embedding (cDAL) for medical image segmentation.

2.1 Conditional Diffusion Model for Image Segmentation

There is inherent ambiguity in medical image segmentation as the delineation of the same image differs among experts. In our proposed cDAL, we utilized the stochastic nature of DDPM to approximate this process and generate multiple predictions during inference. Subsequently, we take the mean of the predictions and threshold them to obtain more accurate segmentation masks compared to deterministic models such as U-Net.

DDPM [5] consists of a markov-chain forward process where Gaussian noise is gradually added to perturb the data distribution in T time-steps. The forward process q is given by the joint distribution: $q(x_{1:T}|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1})$, where for each step t, the forward process is: $q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I)$. Here, x_0 is sampled from the data distribution, T is the number of time-steps, β_t is the predefined noise schedule, $\mathcal{N}$ denotes the Gaussian distribution and I is a $n \times n$ shaped identity matrix of the same shape as the data x_0. The cumulative process from x_0 to x_t is represented as: $x_t = \sqrt{\bar{\alpha}_t}x_0 + (1-\bar{\alpha}_t)\epsilon, \epsilon \sim \mathcal{N}(0, I_{n\times n})$.

Here, $\bar{\alpha}_t = \prod_{s=1}^{t}(1-\beta_s)$ is the cumulative scaling factor used during the forward process $q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)I)$ to obtain sample x_t at arbitrary time-step t.

The reverse process of DDPM to iteratively denoise the latent variables $(x_1, ..., x_T)$ is parameterized by the joint distribution $p_\theta(x_{0:T})$ and given by:

$$p_\theta(x_{0:T}) = p(x_T)\prod_{t=1}^{T} p_\theta(x_{t-1}|x_t) = p(x_T)\prod_{t=1}^{T}\mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \sigma_t^2 I) \tag{1}$$

where, $\mu_\theta(x_t, t)$, σ_t^2 and θ denote the mean, variance and parameters of the denoising model $p_\theta(x_{t-1}|x_t)$ respectively.

By maximizing the evidence lower bound [5], we have the training loss function: $\arg\min_\theta \mathbb{E}_{x_0,\epsilon,t}[||\epsilon - \epsilon_\theta(x_t, t)||^2]$, where $\epsilon \sim \mathcal{N}(0, I_{n\times n})$ denotes pure Gaussian noise and ϵ_θ denotes the predicted noise by the denoising network. During the sampling stage, the trained model ϵ_θ is used to iteratively denoise the data, i.e. generate x_{t-1} from x_t for $t = T, T-1, ..., 1$ and eventually generate the data x_0 by starting from pure Gaussian noise $x_T \sim \mathcal{N}(0, I_{n\times n})$.

This unconditional generation process of DDPM is suitable for image generation tasks where the goal is to model a data distribution. For image segmentation tasks, there exists an image and label pair (I, x), where I denotes the image and x denotes the corresponding ground-truth label. Hence, for image segmentation tasks, diffusion models are helpful to generate a distribution of labels but it needs to have the image as a condition to generate relevant labels.

In cDAL, we use the diffusion model as a generator x_θ with the image I as a condition to guide the diffusion model to generate the label (x) corresponding to the image I as shown in Fig. 1. In our approach, instead of predicting the noise with the diffusion model ϵ_θ, we use the formulation provided by [19] and directly predict the clean label $\hat{x}_0$ using our diffusion model conditioned on the image, i.e. $\hat{x}_0 = x_\theta(x_t, t, I)$, where t is the time embedding.

2.2 cDAL: Spatial Attention Maps

In the cDAL architecture, we incorporate a distinct CNN-based discriminator D as shown in Fig. 1. This discriminator is trained to differentiate between the ground-truth segmentation labels and the labels generated using our diffusion model $x_\theta(x_t, t, I)$.

More specifically, first the conditional diffusion model is frozen. The perturbed label x_{t-1} is generated using the forward process and ground-truth labels, i.e. $x_{t-1} := q(x_{t-1}|x_0)$. Then, the discriminator D uses $x_t := q(x_t|x_{t-1})$, x_{t-1} and time-step t as inputs to predict the label as real. The cross-entropy loss is used to update D. Subsequently, with the diffusion model still frozen, the output of the diffusion model is: $\hat{x}_0 = x_\theta(x_t, t, I)$. With $\hat{x}_0$, $\hat{x}_{t-1}$ is sampled using the posterior distribution $q(\hat{x}_{t-1}|x_t, \hat{x}_0)$. Then, the discriminator uses $\hat{x}_{t-1}$, x_t and t as inputs to predict the label as fake (class 0), and the cross-entropy loss is used to update D again.

Clearly, the discriminator D learns the most discriminative features to differentiate between the real x_{t-1} and predicted $\hat{x}_{t-1}$. Here, we use the feature maps of D to generate the spatial attention map $A_D = \frac{1}{C}\sum_{i=1}^{C} F_i$, where, F_i denotes the i^{th} feature map of D with C channels.

The attention map A_D highlights the spatial regions in the labels which are essential for our model to generate labels $\hat{x}_0$ that are close to the ground-truth x_0. We upsample the attention map A_D to match the shape of ground-truth label x_0 and then perform element-wise multiplication with x_0 to get $x_0^{att} = x_0 \odot A_D$, where $\odot$ represents the Hadamard product. Subsequently, the forward process is used to transform x_0^{att} to x_t^{att} using $q(x_t^{att}|x_0^{att})$. The perturbed x_t^{att} is fed to the conditional diffusion model to predict $\hat{x}_0$ as depicted in Fig. 1. With discriminator D fixed, the diffusion model loss is $||x_0 - x_\theta(x_t^{att}, t, I)||^2$, where, x_0 is the ground-truth label, x_θ is the denoising model dependent on the attention incorporated x_t^{att}. This loss is used to update the parameters θ of the conditional diffusion model.

2.3 cDAL: Latent Embedding

DDPM [5] typically uses a large number of time-steps for both training and sampling since they use small step-sizes. Hence, the true denoising distribution is closer to a Gaussian distribution. When the denoising step size becomes larger, the denoising distribution deviates from a Gaussian and becomes a complex multi-modal distribution [18]. In cDAL, we use larger step sizes to perturb the label data $x_0 \sim q(x_0)$ in T time-steps ($T \leq 4$) using the forward process $q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I)$ with large variance β_t in each time step. For the reverse process, the denoising model is given by:

$$p_\theta(\hat{x}_{t-1}|x_t) := q(\hat{x}_{t-1}|x_t, \hat{x}_0 = x_\theta(x_t^{att}, t, I)) \tag{2}$$

where, $\hat{x}_0$ is first predicted using $x_\theta(x_t^{att}, t, I)$ and then from the posterior distribution $q(\hat{x}_{t-1}|x_t, \hat{x}_0)$, $\hat{x}_{t-1}$ is sampled as shown in Fig. 1.

Now, a random latent embedding $z \sim p(z) := \mathcal{N}(z; 0, I)$ is introduced in cDAL x_θ such that $\hat{x}_0 = x_\theta(x_t, t, z, I)$. Hence, the denoising model $p_\theta(\hat{x}_{t-1}|x_t)$ is given by:

$$p_\theta(\hat{x}_{t-1}|x_t) := \int p_\theta(\hat{x}_0|x_t) q(\hat{x}_{t-1}|x_t, \hat{x}_0) d\hat{x}_0 = \int p(z) q(\hat{x}_{t-1}|x_t, \hat{x}_0 = x_\theta(x_t^{att}, t, z, I)) dz \tag{3}$$

where, $p_\theta(\hat{x}_0|x_t)$ is the implicit distribution by our conditional diffusion model generator $x_\theta(x_t^{att}, t, z, I)$ that uses a L-dimensional latent variable z. Hence, the mapping of our conditional diffusion model label generator is $x_\theta(x_t^{att}, t, z, I)$.

The predicted label $\hat{x}_0$ is not a deterministic mapping of x_t as in DDPM but it is produced by the denoising model with a random latent variable z. This process makes the denoising distribution $p_\theta(\hat{x}_{t-1}|x_t)$ multimodal and hence larger step sizes could be used. The final loss to update x_θ (with D frozen) is given by:

$$||x_0 - x_\theta(x_t^{att}, t, z, I)||^2. \tag{4}$$

The training and sampling details of cDAL are given by Algorithms 1 and 2, respectively, which are described in the supplemental material.

3 Experiments and Results

We performed extensive experiments with our proposed cDAL algorithm on three public datasets and compared cDAL with several state-of-the-art (SOTA) segmentation methods, including SegDiff [11], the best diffusion-based image segmentation model.

3.1 Datasets

MoNuSeg Dataset (2D Binary) - This dataset [22,23] consists of H&E stained tissue images of patients with tumors of different organs. It contains 30 training and 14 held-out testing color images and corresponding binary labels.
CXR Dataset (2D Binary) - The National Library of Medicine in Maryland, USA created a standard digital chest X-ray dataset. This dataset comprises of 704 grayscale images and binary labels for the lungs, divided into 566 images for training and 138 images for testing, with a 3-fold cross-validation.
Hippocampus Dataset (3D Multi-class) - This dataset [21] is collection of 3D T1-weighted MRI images where each volume was annotated by using 2 labels for hippocampus and parts of the subiculum. The labels comprised of 3 classes: background, anterior and posterior and hence a slice by slice one-hot encoding was used to train and test the model. We divided the dataset into 130 and 65 for training and testing, with a 4-fold cross validation.

3.2 Experimental Setup and Implementation Details

We compared cDAL with several SOTA medical image segmentation models. These include U-Net [3], U-Net++ [26], MedT [27], Res-UNet [28], MSU-Net [24], Multi-SegCaps [30], EM-SegCaps [31], 3D-UCaps [29] and SegDiff [11]. We briefly describe SegDiff below since it is the SOTA diffusion model based algorithm for image segmentation.
SegDiff [11] - Segdiff is an integration of the advanced image generation approach of diffusion models for image segmentation tasks. For the diffusion model, it uses a U-Net architecture with the input image passed as a condition through an image encoder that consists of several Residual in Residual Dense Blocks (RRDB) [25]. SegDiff uses 100 diffusion time-steps in its experiments.
Implementation Details - For cDAL, the discriminator architecture resembles the encoder part of the diffusion network, comprising of Residual blocks. Similar to other diffusion models, we utilized sinusoidal positional embeddings for time-step t, for both the discriminator and the diffusion model. Since the diffusion model's inference is not deterministic, following SegDiff [11], we ran cDAL for 5 instance generations during the inference stage and calculated the mean segmentation map. We used PyTorch and MONAI framework for our experiments and trained our models on a NVIDIA Quadro RTX 6000 GPU.

Table 1. Ablation study with the MoNuSeg and chest X-ray (CXR) dataset.

Model	MoNuSeg			CXR		
	mIoU (%)	Dice (%)	Attn. scale	mIoU (%)	Dice (%)	Attn. scale
cDAL w\o Latent	65.95	79.37	32	92.50	96.06	16
cDAL w\o Attention	70.11	82.36	-	93.27	96.48	-
cDAL	70.70	82.77	16	**94.00**	**96.87**	**16**
cDAL	**70.96**	**82.94**	**32**	93.87	96.80	32
cDAL	70.38	82.53	64	93.68	96.70	64

Evaluation Metrics - We employed three quantitative evaluation metrics. Following the literature, we used the Dice score and mIoU (mean Intersection over Union) for the CXR and MoNuSeg datasets [11,24] and used the Dice score, precision and recall for the Hippocampus dataset [29].

3.3 Ablation Study

To assess the impact of each component in our model, we conducted an ablation study as shown in Table 1.

The incorporation of attention map in cDAL increases the Dice score and mIoU by up to 0.49% and 0.79%, respectively, on average for both the datasets. Subsequently, we identified the optimal layer in the discriminator from which we could extract the attention map. For MoNuSeg, the best layer was 32×32, while for CXR it was 16×16. One reason for this difference is that the middle layer attention maps usually contain more information about boundaries and edges (smaller-sized labels as in MoNuSeg), whereas the attention maps of the later layers typically focus on entire objects (larger labels as in CXR). Additionally, we examined the importance of the random latent embedding in our model. Without the latent embedding, the mIoU and Dice score for cDAL drops significantly with similar diffusion time-steps as our proposed cDAL, and more number of time-steps would be necessary to match the performance of cDAL.

3.4 Segmentation Using MoNuSeg Dataset

Table 2 (left) presents the performance of cDAL and its comparison with several SOTA segmentation models on the MoNuSeg dataset. On a held-out test set, cDAL demonstrates a significant improvement over the other models, and an improvement of 1.96% in mIoU and 1.35% in Dice score over SegDiff, the current best diffusion-based segmentation model. It is worth mentioning that the notable enhancement in performance was achieved using a much lighter conditional image encoder, with 95% less parameters than the SegDiff image encoder. Additionally, the cDAL was much faster with inference time of 1 s as it used just 4 time-steps ($T = 4$) for training and sampling, compared to 100 time-steps ($T = 100$) in SegDiff which takes 60 s inference time. Hence, our method is not

Table 2. Segmentation results on the MoNuSeg and chest X-ray (CXR) dataset. For the CXR dataset, the standard deviation across the 3-folds is indicated inside the parenthesis.

Model	MoNuSeg		Model	CXR	
	mIoU (%)	Dice (%)		mIoU (%)	Dice (%)
U-Net [3]	65.99	79.43	U-Net [3]	91.91 (± 0.28)	95.75 (± 0.17)
U-Net++ [26]	66.04	79.49	U-Net++ [26]	92.03 (± 0.79)	95.80 (± 0.45)
MedT [27]	66.17	79.55	MSU-Net [24]	92.19 (± 0.61)	95.90 (± 0.35)
Res-UNet [28]	66.07	79.49	-	-	-
SegDiff [11]	69.00	81.59	SegDiff [11]	92.33 (± 0.69)	95.95 (± 0.40)
cDAL (ours)	**70.96**	**82.94**	cDAL (ours)	**93.04** (±0.97*)	**96.35** (±0.54*)

* indicates that the performance improvement by cDAL is statistically significant based on a t-test.

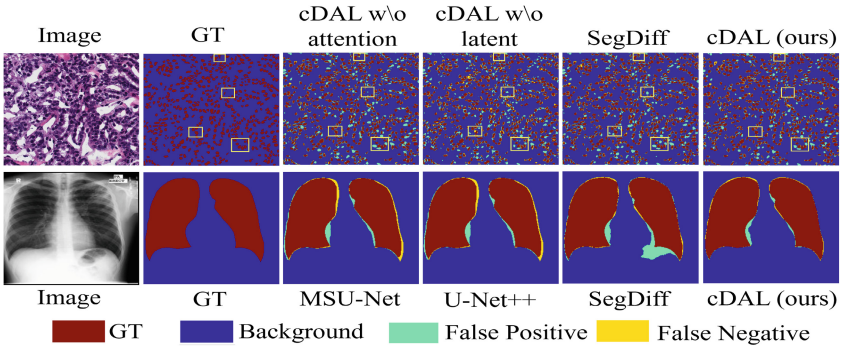

Fig. 2. Visualization of the Image, ground-truth (GT) and predictions with different models for MoNuSeg (top row) and CXR (bottom row) dataset. The detailed zoomed-in visual comparison is provided in the supplemental material.

computationally expensive as in inference we remove the discriminator and perform sampling with a much smaller number of steps. A qualitative comparison is provided in Fig. 2 (top row).

3.5 Segmentation Using Chest X-Ray (CXR) Dataset

Table 2 (right) provides a comprehensive comparison between cDAL and other SOTA methods for segmentation tasks on the Chest X-ray (CXR) dataset. A 3-fold cross validation was performed to compare the performance of all models. On average, cDAL shows significant improvement in mIoU and Dice score compared to other models and an increase of 0.71% and 0.40% over SegDiff for mIoU and Dice, respectively. Again, cDAL is with less parameters and much faster as it achieved this performance with just 2 time-steps for training and

Table 3. Segmentation results for the Hippocampus dataset with the standard deviation across the 4-folds indicated inside the parenthesis.

Model	Precision (%)		Recall (%)		Dice (%)	
	Anterior	Posterior	Anterior	Posterior	Anterior	Posterior
Multi-SegCaps [30]	65.65	60.49	80.76	84.46	72.42	70.49
EM-SegCaps [31]	20.01	34.55	17.51	19.00	18.67	24.52
3D-UCaps [29]	87.79 (± 0.50)	85.79 (± 1.88)	84.10 (± 1.80)	82.17 (± 1.76)	85.73 (± 1.02)	83.77 (± 0.55)
SegDiff [11]	88.51 (± 0.94)	86.42 (± 0.67)	87.74 (± 1.53)	86.36 (± 1.18)	87.80 (± 0.24)	86.38 (± 0.22)
cDAL (ours)	**88.76** (± 1.15)	**87.43** (± 1.50)	**87.85** (± 0.79)	**86.72** (± 1.25)	**88.13** (±0.18*)	**86.90** (±0.13*)

* indicates that the performance improvement by cDAL is statistically significant based on a t-test.

sampling, compared to 100 in SegDiff. A qualitative comparison is provided in Fig. 2 (bottom row).

3.6 Segmentation Using Hippocampus Dataset

Table 3 compares cDAL with other SOTA methods for segmentation tasks on the Hippocampus dataset. A 4-fold cross validation was performed to compare the performance of all models. On average, cDAL shows significant improvement in precision, recall and Dice score compared to other models. Compared to SegDiff, an average improvement of 0.64%, 0.24% and 0.43% in precision, recall and Dice score was observed, with just 2 time-steps (T = 2) for training and sampling. The visualization of the predictions is provided in the supplemental material. One limitation of our model is that it can only be applied on a 2D slice and in the future we will have a 3D version of our model.

4 Conclusion

In this paper, we proposed cDAL, a novel conditional diffusion model for medical image segmentation. cDAL incorporates the spatial attention from the discriminator to guide the label generation process. It also includes the random latent embedding which helped significantly reduce the number of time-steps during training and sampling. cDAL demonstrated superior results on benchmarking medical image segmentation datasets.

Acknowledgement. Research reported in this publication was partly supported by the National Institute of Health (NIH R01HL153720).

Disclosure of Interests. The authors declare that they have no competing interests in the paper.

References

1. Wang, R., Lei, T., Cui, R., Zhang, B., Meng, H., Nandi, A.: Medical image segmentation using deep learning: a survey. IET Image Proc. **16**, 1243–1267 (2022)
2. Masood, S., Sharif, M., Masood, A., Yasmin, M., Raza, M.: A survey on medical image segmentation. Current Med. Imaging **11**, 3–14 (2015)
3. Ronneberger, O., Fischer, P., Brox, T.: U-net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
4. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3D U-net: learning dense volumetric segmentation from sparse annotation. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (eds.) MICCAI 2016. LNCS, vol. 9901, pp. 424–432. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46723-8_49
5. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Advances in Neural Information Processing Systems, vol. 33, pp. 6840–6851 (2020)
6. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: International Conference on Machine Learning, pp. 2256–2265 (2015)
7. Nichol, A., Dhariwal, P.: Improved denoising diffusion probabilistic models. In: International Conference on Machine Learning, pp. 8162–8171 (2021)
8. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10684–10695 (2022)
9. Dhariwal, P., Nichol, A.: Diffusion models beat GANs on image synthesis. In: Advances in Neural Information Processing Systems, vol. 34, pp. 8780–8794 (2021)
10. Wolleb, J., Sandkühler, R., Bieder, F., Valmaggia, P., Cattin, P.: Diffusion models for implicit image segmentation ensembles. In: International Conference on Medical Imaging with Deep Learning, pp. 1336–1348 (2022)
11. Amit, T., Shaharbany, T., Nachmani, E., Wolf, L.: SegDiff: image segmentation with diffusion probabilistic models. ArXiv Preprint ArXiv:2112.00390 (2021)
12. Rahman, A., Valanarasu, J., Hacihaliloglu, I., Patel, V.: Ambiguous medical image segmentation using diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11536–11546 (2023)
13. Amit, T., Shichrur, S., Shaharabany, T., Wolf, L.: Annotator consensus prediction for medical image segmentation with diffusion models. ArXiv Preprint ArXiv:2306.09004 (2023)
14. Baranchuk, D., Rubachev, I., Voynov, A., Khrulkov, V., Babenko, A.: Label-efficient semantic segmentation with diffusion models. ArXiv Preprint ArXiv:2112.03126 (2021)
15. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. ArXiv Preprint ArXiv:2010.02502 (2020)
16. Salimans, T., Ho, J.: Progressive distillation for fast sampling of diffusion models. ArXiv Preprint ArXiv:2202.00512 (2022)
17. Song, Y., Dhariwal, P., Chen, M., Sutskever, I.: Consistency models (2023)
18. Xiao, Z., Kreis, K., Vahdat, A.: Tackling the generative learning trilemma with denoising diffusion GANs. ArXiv Preprint ArXiv:2112.07804 (2021)
19. Benny, Y., Wolf, L.: Dynamic dual-output diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11482–11491 (2022)

20. Emami, H., Aliabadi, M., Dong, M., Chinnam, R.: SPA-GAN: spatial attention GAN for image-to-image translation. IEEE Trans. Multimedia **23**, 391–401 (2020)
21. Simpson, A., et al.: A large annotated medical image dataset for the development and evaluation of segmentation algorithms. ArXiv Preprint ArXiv:1902.09063 (2019)
22. Kumar, N., et al.: A multi-organ nucleus segmentation challenge. IEEE Trans. Med. Imaging **39**, 1380–1391 (2019)
23. Kumar, N., Verma, R., Sharma, S., Bhargava, S., Vahadane, A., Sethi, A.: A dataset and a technique for generalized nuclear segmentation for computational pathology. IEEE Trans. Med. Imaging **36**, 1550–1560 (2017)
24. Su, R., Zhang, D., Liu, J., Cheng, C.: MSU-Net: multi-scale U-Net for 2D medical image segmentation. Front. Genet. **12**, 639930 (2021)
25. Wang, X., et al.: ESRGAN: enhanced super-resolution generative adversarial networks. In: Leal-Taixé, L., Roth, S. (eds.) ECCV 2018. LNCS, vol. 11133, pp. 63–79. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-11021-5_5
26. Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J.: UNet++: a nested U-Net architecture for medical image segmentation. In: Stoyanov, D., et al. (eds.) DLMIA/ML-CDS -2018. LNCS, vol. 11045, pp. 3–11. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00889-5_1
27. Valanarasu, J.M.J., Oza, P., Hacihaliloglu, I., Patel, V.M.: Medical transformer: gated axial-attention for medical image segmentation. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12901, pp. 36–46. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87193-2_4
28. Xiao, X., Lian, S., Luo, Z., Li, S.: Weighted Res-UNet for high-quality retina vessel segmentation. In: 2018 9th International Conference on Information Technology in Medicine and Education (ITME), pp. 327–331 (2018)
29. Nguyen, T., Hua, B.-S., Le, N.: 3D-UCaps: 3D capsules UNet for volumetric image segmentation. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12901, pp. 548–558. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87193-2_52
30. LaLonde, R., Bagci, U.: Capsules for object segmentation. ArXiv Preprint ArXiv:1804.04241 (2018)
31. Survarachakan, S., Johansen, J., Pedersen, M., Amani, M., Lindseth, F.: Capsule nets for complex medical image segmentation tasks. In: CVCS (2020)