IEICE TRANS. INF. & SYST., VOL.E106–D, NO.11 NOVEMBER 2023

LETTER

Inverse Heat Dissipation Model for Medical Image Segmentation

Yu KASHIHARA[†a)], ***Nonmember and*** **Takashi MATSUBARA**[†b)], ***Member***

SUMMARY The diffusion model has achieved success in generating and editing high-quality images because of its ability to produce fine details. Its superior generation ability has the potential to facilitate more detailed segmentation. This study presents a novel approach to segmentation tasks using an inverse heat dissipation model, a kind of diffusion-based models. The proposed method involves generating a mask that gradually shrinks to fit the shape of the desired segmentation region. We comprehensively evaluated the proposed method using multiple datasets under varying conditions. The results show that the proposed method outperforms existing methods and provides a more detailed segmentation.

key words: *deep learning, diffusion-based models, inverse heat dissipation models, and medical segmentation*

1. Introduction

Segmentation is a critical task in computer vision that separates the regions of an image into subregions, each corresponding to a class. Segmentation has many applications in the medical field, such as disease detection and organ recognition. Automated segmentation plays an essential role for physicians by assisting in diagnosis, reducing undiagnosed diseases, and improving work efficiency. Medical segmentation requires the capability to capture the details of the disease. Inappropriate segmentation can lead to the misinterpretation of disease progression, resulting in medical errors. Several medical segmentation methods have been proposed [1]–[3]. UNet [1] is a segmentation model trained to minimize cross-entropy or squared error and generates masks for each pixel. The neighborhood of a definitively diagnosed area has a high probability of suffering from that disease, and it is inefficient to predict whether each independent pixel is a disease. Convolutional neural networks can capture features in patches and represent relationships between pixels. However, this is a network structure approach, and an approach from a theoretical perspective is desired. We propose a method for capturing the context of an image from a theoretical perspective to facilitate the development of segmentation tasks in medicine.

Diffusion-based models [4]–[6] have shown impressive performance in image generation and editing [7], [8]. Their contribution stems from their ability to produce high-quality images. In this paper, we suppose that the model's high generation capacity enables detailed segmentation with high accuracy. Diffusion-based models consist of two stochastic transitions: diffusion and reverse processes. During the diffusion process, the information in the image is lost little by little. The reverse process gradually generates an image by following the diffusion process in the reverse direction. The iterative reverse process is defined as a Markov chain in which the next image is generated from the previous step's image. The reverse process is obtained by a joint probability distribution defined by the Markov chain. Stochastic transitions allow the model to capture the context of the image rather than at the pixel level. However, these methods suffer from unstable generation owing to long stochastic transitions. In addition, a segmentation mask is often a continuous set of pixels. Therefore, small discontinuous masks can hardly exist. Typical denoising-based diffusion models [4] may produce unexpected discontinuous masks due to large noise addition.

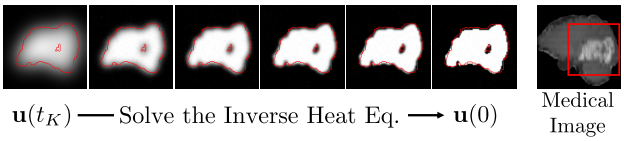

Fig. 1 Mask generation by IHDM conditioned on a medical image. The boundary is sharpened step by step from a coarse initial mask. Sampling does not remove noise as in the regular diffusion model, but rather restores the high-frequency components of the image.

This study proposes a diffusion-based model segmentation method to capture contextual features to avoid undiagnosed diseases. We propose a segmentation method using the inverse heat dissipation model (IHDM) [9]. IHDM is inspired by diffusion models and generates samples by modeling inverse heat dissipation. During diffusion, IHDM gradually averages the input images rather than converting them to noise. The diffusion process is equivalent to the operation of gradually reducing the high-frequency components of an image. Sampling was performed by modeling the inverse heat dissipation using a deep learning model. IHDM generates the image by restoring the high-frequency components of the image and gradually sharpening the boundaries, as shown in Fig. 1. Mask generation by restoring high-frequency components does not produce unexpected discontinuous masks. The proposed method uses coarse segmentation as input and reconstructs it with short stochastic transitions to generate a mask with high accuracy. Sampling from

Manuscript received March 23, 2023.
Manuscript revised July 31, 2023.
Manuscript publicized August 22, 2023.
[†]The authors are with Graduate School of Engineering Science, Osaka University, Toyonaka-shi, 560–8531 Japan.
a) E-mail: kashihara@hopf.sys.es.osaka-u.ac.jp
b) E-mail: matsubara@sys.es.osaka-u.ac.jp
DOI: 10.1587/transinf.2023EDL8017

the initial mask allows the model to focus only on correcting the details of the mask.

2. Method

2.1 Inverse Heat Dissipation Model

IHDM [9] is a diffusion-based generative model consisting of diffusion and reverse processes. In the diffusion process, the pixel values of the image gradually approach the average value of the input image and information is lost. In the reverse process, the model generates samples through the inverse heat equation. The heat equation, expressed by the following linear partial differential equation, defines the diffusion process.

$$\frac{\partial}{\partial t}u(x, y, t) = \Delta u(x, y, t). \tag{1}$$

Here, $u(x, y, t)$ is a continuous two-dimensional image plane for each color channel and Δ is the Laplace operator. Neumann boundary conditions ($\partial u/\partial x = \partial u/\partial y = 0$) define the image boundary conditions. In the diffusion process, as time t approaches infinity, pixel values become the average of the intensities of the input image. The heat equation is reversible if assumed to have infinitely continuous states. However, for discrete values in real-world problems, this becomes an ill-posed problem.

The heat equation in Eq. (1) can be represented in the evolution equation form of $u(x, y, t) = \mathcal{F}(t)u(x, y, t)|_{t=t_0}$, where $\mathcal{F}(t)$ is an evolution operator. We solve this equation using the eigenbasis of the Laplace operator. Under Neumann boundary conditions, a cosine basis is used for the eigenbasis. An image of finite resolution can be represented by a grid whose spectrum is bounded by Nyquist frequency. In a finite-dimensional evolution model, the image $\mathbf{u}(t_k)$ exhibits a frequency decay at time t_k controlled by the discrete time sequence $t_1, t_2, \ldots, t_K$. In the following, we use $\mathbf{u}_k$ as a simplified notation for $\mathbf{u}(t_k)$. The evolution operator is defined as $\Delta = \mathbf{V}\Lambda\mathbf{V}^\top$ in the finite eigendecomposition: where $\mathbf{V}^\top$ is the projection matrix on a cosine basis and Λ is the negative squared frequency that corresponds to the frequency decay. Thus, the image with frequency decay at time t_k with a finite resolution is represented as follows:

$$\mathbf{u}_k = \mathbf{F}(t_k)\mathbf{u}_0 = \exp(\mathbf{V}\Lambda\mathbf{V}^\top t_k)\mathbf{u}_0, \tag{2}$$

where $\mathbf{F}(t_k)$ is the evolution operator for a finite dimension and $\mathbf{u}_0$ is the initial state. Instead of using the Markov chain obtained by Eq. (2), we add the noise and sample as $\hat{\mathbf{u}}_k \sim \mathcal{N}(\mathbf{u}_k, \sigma^2\mathbf{I})$.

The reverse process of the heat equation is a Markov chain starting from the image $\mathbf{u}_{t_K}$, defined by a joint probability distribution, as follows:

$$p_\theta(\mathbf{u}_{0:K}) = \mathbf{u}_K \prod_{i=1}^{K} \mathcal{N}(\mathbf{u}_{i-1}|\mu_\theta(\mathbf{u}_i, i), \delta^2\mathbf{I}), \tag{3}$$

where μ_θ is the model that predicts a slightly less blurred image $\mathbf{u}_{k-1}$ from blurred image $\mathbf{u}_k$. We add normal noise with variance $\delta^2\mathbf{I}$ and sample stochastically to improve the sampling performance and manipulate the frequency information in the image $\mathbf{u}_0$.

The model is trained by minimizing $\|\mu_\theta(\hat{\mathbf{u}}_k, k) - \mathbf{u}_{k-1}\|_2^2$., which is obtained by transforming and simplifying the negative log likelihood. The diffusion process in IHDM is not a Markov chain, and diffusion and noise addition are performed separately. Therefore, denoising score matching could not be performed, and the loss function used in IHDM directly minimizes the error in the mean of the distribution.

2.2 Proposed Method

We propose a method to generate a mask that fits the details by refinement of the coarse segmentation; the proposed method uses a diffusion-based model to generate a mask through a reverse process. To predict the segmentation mask, the medical image is conditioned on the mask as prior information, thereby providing the model with information for segmentation. This image-conditioning method is used in many diffusion-based model segmentation methods [10]–[12]. In practice, the model takes the concatenation of the mask and the medical image as input, and predicts the mask for the next step. In image generation with IHDM, the sum of the pixel values at each step does not change from step to step. As a result, the size of the mask is limited, and thus masks are generated in stable sizes and shapes. As the initial state, the proposed method uses $\mathcal{M}_{\text{prior}}$, a coarse mask generated by another segmentation model. The initial mask must have a coarse shape and should not require detailed accuracy. Therefore, it is possible to use a mask from a segmentation model at a low training cost. The proposed method uses coarse masks to guide the generation trajectory, thereby allowing the model to focus on generating details. Figure 2 shows the mask generation flow using the proposed method. The mask was reconstructed by applying a k-step diffusion process to the initial mask, followed by a k-step reverse process. Therefore, the stochastic transitions of the mask generation can be represented as follows:

$$p_\theta(\mathbf{u}_{0:k}) = \mathbf{u}_k \prod_{i=1}^{k} p_\theta(\mathbf{u}_{i-1}|\mathbf{u}_i), \tag{4}$$

$$= \mathbf{F}(t_k)\mathcal{M}_{\text{prior}} \prod_{i=1}^{k} \mathcal{N}(\mathbf{u}_{i-1}|\mu_\theta(\mathbf{u}_i, i), \delta^2\mathbf{I}), \tag{5}$$

where $\delta^2\mathbf{I}$ is the variance matrix, a hyperparameter that controls the frequency features of the generated image. The mask is generated using a joint probability distribution defined by a Markov chain. The transitions are defined per image and not per pixel. Therefore, it is possible to obtain a relationship between the pixels. Diffusion-based models with joint probability distributions can generate masks that can capture the context of an image. The number of reconstruction steps k can be considered as the confidence level of the initial mask. For an initial mask with low confidence, the shape of the mask can be changed significantly by performing a large number of reconstruction steps. On the other hand, it is possible to generate a mask with only the details adjusted by taking a few reconstruction steps for an initial

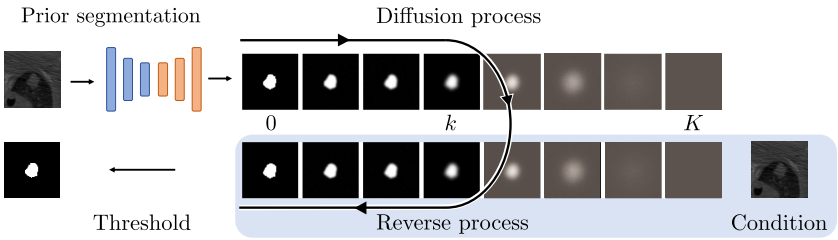

Fig. 2 Segmentation using IHDM. The proposed method receives a coarse mask from another segmentation model. The proposed method performs k-step reconstruction on the initial mask $\mathcal{M}_{\text{prior}}$ to produce a mask that fits the details. The mask is generated as the contours become gradually sharper.

Table 1 Comparison of segmentation methods. The values in the table indicate the mean and standard deviation.

Method	Initial mask	LIDC-IDRI [13]		STARE [14]		BraTS [15]	
		IoU	Dice	IoU	Dice	IoU	Dice
DeepLabv3+ [16]	-	0.375±0.231	0.459±0.261	0.540±0.112	0.685±0.114	0.518±0.166	0.644±0.179
UNet++ [17]	-	0.427±0.238	0.512±0.262	0.574±0.103	0.716±0.102	0.547±0.149	0.679±0.151
DDPM Seg [10]	-	0.402±0.250	0.474±0.283	0.351±0.123	0.493±0.148	0.250±0.174	0.342±0.214
SegDiff [11]	-	0.402±0.259	0.468±0.289	0.552±0.116	0.694±0.119	0.545±0.173	0.665±0.185
MedSegDiff [12]	-	0.370±0.232	0.450±0.267	0.409±0.102	0.564±0.120	0.483±0.170	0.610±0.182
IHDM Seg (Ours)	DeepLabv3+	0.439±0.233	0.526±0.233	0.562±0.107	0.705±0.105	0.538±0.162	0.664±0.172
IHDM Seg (Ours)	UNet++	**0.460 ± 0.232**	**0.548 ± 0.259**	**0.595 ± 0.095**	**0.735 ± 0.092**	**0.574 ± 0.145**	**0.703 ± 0.145**

mask with high confidence.

3. Experiments

3.1 Datasets

Our experiments used three datasets with different targets In the LIDC-IDRI dataset [13] of lung nodules, we used only nodules diagnosed as ground-glass opacity (GGO). The GGO is challenging to detect and is a barrier to diagnosis. The experiment was performed on 64×64 CTs scans, focusing on the target lung nodules. The STARE dataset [14] is a retinal-segmentation dataset. The original images were resized to 512×512 pixels and divided into 16 segments of 128×128 pixels each. The BraTS dataset [15] is a dataset of brain tumor cases. After the MRI preprocessing, the T1-weighted MRI data were resized to 128×128 pixels. We only used samples from slices containing brain tumors. The image results shown in this study were upsampled for boundary visualization.

3.2 Experimental Setup

In the experiment, the number of steps for IHDM was set to K=300. The reconstruction steps used in the proposed method was set to $k = 100$, the optimal reconstruction steps, through experimentation. The standard deviation of the normal noise added in the diffusion process was set to $\sigma = 0.01$, and that added in the reverse process was set to $\delta = 0.0125$. For the initial masks of the proposed method, we used two methods: UNet++ [17] and DeepLabv3+ [16]. We used the pre-trained ResNet-50 [18] as the backbone model for these two models, and trained 100 epochs on each dataset. We also experimented with denoising diffusion probabilistic model (DDPM) based segmentation methods [10]–[12] for a comparison. For a fair comparison between the proposed method and DDPM-based methods, the same architecture of UNet++ [17] was used for their backbone model. DDPM-based segmentation methods typically use mask ensembles [10]–[12]. The mask ensemble obtains the mask by averaging the masks of the multiple generations. The mask ensemble prevents mask fluctuations due to stochastic processes. The proposed method and DDPM-based methods used mask ensemble in the experiments.

3.3 Experimental Results

The segmentation ability of the methods was evaluated using IoU (Intersection over Union) and Dice coefficient. The experimental results are shown in Table 1 as mean ± standard deviation. The proposed method performed better than the methods used for the initial mask $\mathcal{M}_{\text{prior}}$. In addition, the performance of the proposed method improved, regardless of the dataset and initial masks. In particular, we find that IoU improves by more than 0.03 on the LIDC-IDRI dataset. The GGO nodules are difficult to detect owing to their blurred shading. The proposed method could identify subtle boundaries within the background. The segmentation performance with thin segmentation areas was also significantly improved in the retinal dataset. In such datasets, slight segmentation errors accumulated, leading to significant score degradation. These results suggest that the proposed method is advantageous for detailed segmentation. The improved performance in all the results indicates that the proposed method is generally applicable regardless of the data type and the initial mask.

Figure 3 shows an enlarged view of the mask generated

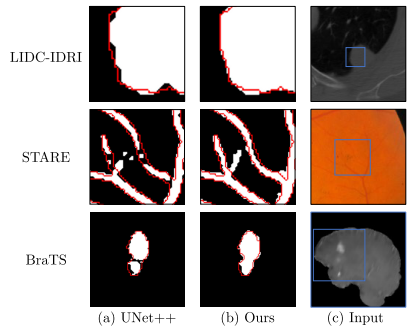

Fig. 3 Enlarged view of the generated masks. The masks of the area enclosed by the blue frame in (c) Input are shown. The red line is the contour of the ground truth. (a) UNet++ [17], (b) Proposed method, (c) Input. The proposed method refines the coarse mask of UNet++ and generates a mask that fits the details.

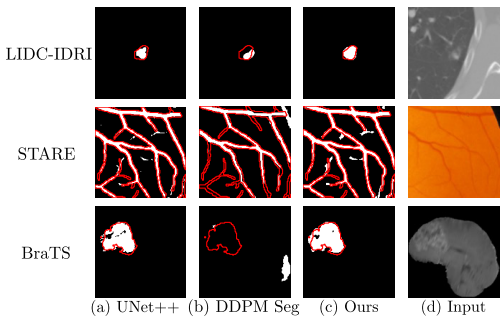

Fig. 4 Visualizations of the segmentation masks for each method. The red line is the contour of the ground truth. DDPM Seg failed to identify a disease for the example shown in the figure of BraTS dataset. (a) UNet++ [17], (b) DDPM Seg [10], (c) Proposed method, (d) Input.

by the proposed method using UNet++ for the initial mask. The proposed method refines the coarse mask of UNet++ and generates a mask that fits the details. Segmentation models such as UNet++ capture targets independently on a pixel-by-pixel basis. In contrast, the diffusion model captures the image in a joint probability distribution, allowing segmentation using the features of the surrounding pixels. Consequently, it is possible to complement the masks of discontinuous missing areas, as shown in the STARE and BRATS results. However, some regions were not segmented sufficiently in detail. One reason is that the average value of the initial mask limited the output mask of IHDM segmentation model. This limitation causes the area of the masked region in the image to be dependent on the initial mask. IHDM segmentation model, which allows for a wide range of mask sizes, is a future challenge.

The proposed method also showed better segmentation performance than DDPM-based segmentation method [10]. Figure 4 shows a visualization of the segmentation masks for each method. The proposed method can predict anomalous regions, whereas DDPM Seg [10] cannot. Long stochastic transitions produce a variety of products in the generation task, but can be an uncertain mask in the segmentation task. The proposed method allows stable segmentation using short stochastic transitions and the initial mask.

4. Conclusion

In this study, we proposed a segmentation method based on IHDM, which is a new types of diffusion-based model. The proposed method reconstructs coarse segmentation as an initial state. Unlike existing methods, we succeeded in using an initial mask to support the generation trajectory. The proposed method is based on a joint probability distribution and succeeds in capturing the features of the entire image rather than pixel-by-pixel. We evaluated the proposed method using different datasets and initial masks and found that all showed improvements. These results indicate that the proposed method can improve performance regardless of the dataset and model used for the initial mask.

Acknowledgments

This work was supported by JST PRESTO (JPMJPR21C7) and Mirai Program (JPMJMI20B8), and JSPS KAKENHI (19H04172, 19K20344, 21H03515).

References

[1] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," MICCAI, pp.234–241, 2015.
[2] S.A. Kamran, K.F. Hossain, A. Tavakkoli, S.L. Zuckerbrod, K.M. Sanders, and S.A. Baker, "RV-GAN: Segmenting retinal vascular structure in fundus photographs using a novel multi-scale generative adversarial network," MICCAI, pp.34–44, 2021.
[3] S. Li, X. Sui, X. Luo, X. Xu, Y. Liu, and R. Goh, "Medical image segmentation using squeeze-and-expansion transformers," IJCAI, 2021.
[4] J. Ho et al., "Denoising diffusion probabilistic models," NeurIPS, 2020.
[5] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," NeurIPS, 2021.
[6] J. Song et al., "Denoising diffusion implicit models," arXiv, 2020.
[7] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," CVPR, pp.10674–10685, 2022.
[8] G. Kim, T. Kwon, and J.C. Ye, "Diffusionclip: Text-guided diffusion models for robust image manipulation," CVPR, pp.2416–2425, 2022.
[9] S. Rissanen et al., "Generative modelling with inverse heat dissipation," arXiv, 2022.
[10] J. Wolleb et al., "Diffusion models for implicit image segmentation ensembles," arXiv, 2021.
[11] T. Amit et al., "Segdiff: Image segmentation with diffusion probabilistic models," arXiv, 2021.
[12] J. Wu et al., "Medsegdiff: Medical image segmentation with diffusion probabilistic model," arXiv, 2022.
[13] C. Fenimore et al., "The lung image database consortium (lidc) and image database resource initiative (idri): A completed reference database of lung nodules on ct scans," Medical Physics, 2011.
[14] A.D. Hoover, V. Kouznetsova, and M. Goldbaum, "Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response," IEEE Trans. Med. Imaging, vol.19, no.3, pp.203–210, 2000.
[15] B.H. Menze et al., "The multimodal brain tumor image segmentation benchmark (brats)," IEEE Trans. Med. Imaging, 2015.

[16] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," ECCV, pp.833–851, 2018.

[17] Z. Zhou, M.M.R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: Redesigning skip connections to exploit multiscale features in image segmentation," IEEE Trans. Med. Imaging, vol.39, no.6, pp.1856–1867, 2011.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," CVPR, pp.770–778, 2016.