

DiffSeg: A Segmentation Model for Skin Lesions Based on Diffusion Difference

Zhihao Shuai^a, Yinan Chen^{a,1}, Shunqiang Mao^{a,1}, Yihan Zhou^{a,1} and Xiaohong Zhang^{a,*}

^aSchool of Big Data and Software Engineering, Chongqing University

Abstract. Weakly supervised medical image segmentation (MIS) using generative models is crucial for clinical diagnosis. However, the accuracy of the segmentation results is often limited by insufficient supervision and the complex nature of medical imaging. Existing models also only provide a single outcome, which does not allow for the measurement of uncertainty. In this paper, we introduce DiffSeg, a segmentation model for skin lesions based on diffusion difference which exploits diffusion model principles to extract noise-based features from images with diverse semantic information. By discerning difference between these noise features, the model identifies diseased areas. Moreover, its multi-output capability mimics doctors' annotation behavior, facilitating the visualization of segmentation result consistency and ambiguity. Additionally, it quantifies output uncertainty using Generalized Energy Distance (GED), aiding interpretability and decision-making for physicians. Finally, the model integrates outputs through the Dense Conditional Random Field (DenseCRF) algorithm to refine the segmentation boundaries by considering inter-pixel correlations, which improves the accuracy and optimizes the segmentation results. We demonstrate the effectiveness of DiffSeg on the ISIC 2018 Challenge dataset, outperforming state-of-the-art U-Net-based methods.

1 Introduction

In the frontier exploration of medical image analysis, image segmentation technology is particularly crucial, serving as an essential foundation for achieving accurate diagnosis and effective treatment. As a foundational task in intelligent diagnostics, MIS is dedicated to precisely extracting key structures or regions from complex medical images, thereby providing robust support for doctors' disease diagnoses.

In recent years, deep learning methods have received widespread attention in MIS tasks and have achieved significant accomplishments. These methods are largely based on Fully Convolutional Networks (FCN), particularly the U-Net model, which adopts an encoder-decoder structure. It effectively merges low-level and high-level information through skip connections, thereby significantly improving the accuracy of segmentation results. The advent of U-Net and its numerous variants has further propelled the development of MIS technology, enabling deep neural networks to process more complex and high-dimensional data, providing a powerful tool for medical image analysis. The advancement of these technologies not only enhances the performance of segmentation tasks but also offers more precise support for clinical diagnosis and treatment.

Existing MIS models often rely on a large amount of manually annotated data when dealing with complex images. Additionally, these models typically only provide a single output result, which has limited referential value. In light of this, the development of algorithms and technologies that can more effectively utilize unannotated or sparsely annotated data, as well as the design of optimization strategies for segmentation results under uncertainty [1], has become a critical issue that urgently needs to be addressed in the field of MIS.

Considering the outstanding performance of diffusion models in the field of natural image processing, we propose a model for medical image segmentation that utilizes the differences in diffusion based on the Denoising Diffusion Probabilistic Models (DDPM). Although the input and output of diffusion models are strongly correlated with noise, their internal feature maps still possess discernible semantic information. Semantic segmentation can be achieved by extracting the semantic information hidden within the diffusion noise, which can be obtained by learning the differences between various noises.

The main contributions of this paper can be summarized as follows:

First, we develop an end-to-end framework from the original image to the optimized segmentation result based on diffusion difference. The model can extract hidden semantic information by learning the differences between various noises, achieving semantic segmentation under weak supervision. It is less susceptible to interference from other factors in the original image and does not depend on a complex annotation process.

Second, we propose a method for measuring the aleatoric uncertainty of segmentation results, which can simulate various segmentation solutions caused by unclear target structures and differences in doctors' experience. This method can intuitively display the consistent areas, ambiguous areas, and prediction confidence of the segmentation results and optimize the model's prediction results, thereby providing more assistance to doctors in practice.

Third, we define a segmentation result optimization method based on the multi-output capability of the diffusion model. This strategy employs DenseCRF algorithm to smooth out noise and local details. It aims to enhance the accuracy of segmentation and the interpretability of the model. This approach not only mitigates the issue of accidental uncertainty stemming from ambiguities in doctors' annotations but also contributes to improving the overall performance of the model.

The rest of the paper is organized as follows: In Section 2, we introduce the related work. In Section 3, we elaborate on our model architecture and optimization strategies. In Section 4, we first introduce the data preprocessing methods we adopted, followed by a de-

* Corresponding Author. Email: xhongz@cqu.edu.cn.

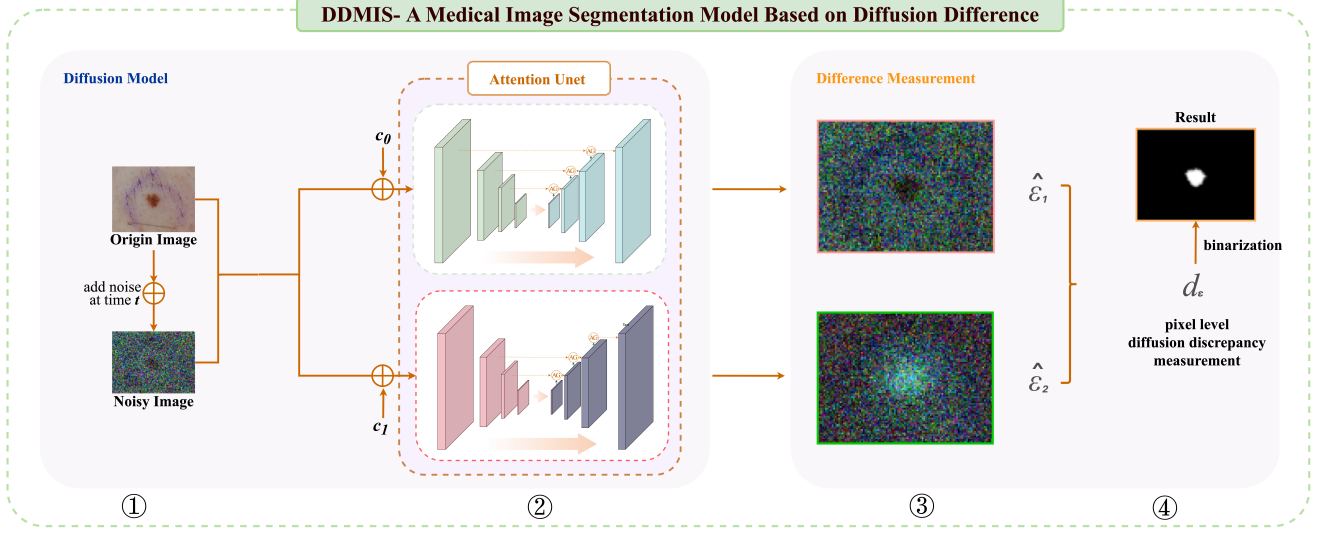


Figure 1. The four steps of DiffSeg. First, add noise to the original image t times. Second, input the original image, noisy image, and category labels into the model. Third, guide the model to generate different noises through two category labels. Fourth, measure the difference between the two types of noise. description of the experimental setup and comparison methodologies, and finally present the experimental results. In Section 5, we compare our methods with previous studies and discuss the limitations of this paper as well as potential future research directions. Finally, we conclude the paper in Section 6.

2 RELATED WORKS

Research on MIS methods has a long history of development. Classic algorithms mainly include thresholding, edge detection, and region-based segmentation algorithms. Yet, these methods lack learning capabilities and highly depend on manually set parameters, quickly being surpassed by numerous methods in the deep learning field, such as fully convolutional networks [14], [16], [20], [7] and the U-Net proposed by Ronneberger et al. [15] However, since deep neural networks require a substantial amount of labeled data for training and given the complex semantics and cumbersome annotation process of medical images, the further development of deep learning algorithms in this field has been limited.

Weakly supervised image segmentation algorithms, which reduce dependence on annotations, have received widespread attention. Currently, many scholars in this field have conducted extensive research using single models or collaborative training algorithms, achieving significant results. For example, Bai et al. [2] combined conditional random fields’ post-processing methods with self-training algorithms for ventricular MRI segmentation tasks. Similarly, Tang et al. [22] used the level set method for post-processing refinement of pseudo-labels. Additionally, Rajchl et al. [19] used self-training methods with additional weak annotations at the boundary level to assist the supervision process. Zhou et al. [29] defined an additional student model based on collaborative training methods to learn fused pseudo-labels. Peng et al. [17] used the mean predictions of multiple models as pseudo-labels and introduced adversarial examples to capture differences between models.

In recent years, generative MIS technology has improved the advantages of strong segmentation modeling ability, good generalization, high efficiency, and low accuracy dependence on phosphate, and has become one of the core technologies to solve the problem of

weakly supervised medical image segmentation. Jin et al. designed a Conditional Generative Adversarial Network (CGAN) [9] to learn the attribute distribution of lung nodules in 3D space, improving the performance of the Progressive Holistically Nested Network (P-HNN) model for pathological lung segmentation in CT scans. Zhao et al. proposed a cascaded generative adversarial network with a deep supervision discriminator (Deep-supGAN) [28] for skeletal segmentation, generating high-quality CT images from MRI for the supervisor to segment bone structures, enhancing segmentation accuracy. Dong et al. introduced a domain adaptation framework based on generative adversarial networks for measuring the cardiothoracic ratio (CTR) [5], incorporating discriminators to learn domain-invariant feature representations and generate accurate segmentations for unlabeled datasets. Yu et al. proposed a Shape Consistency Generative Adversarial Network (SC-GAN) [26] for coronary artery segmentation based on knowledge transfer, enhancing the accuracy of coronary artery segmentation. Liang et al. introduced a generative learning framework for semantic segmentation (GMMSeg) [13], establishing a Gaussian Mixture Model through expectation maximization to obtain class-conditional densities, showcasing the advantages of generative classifiers in large-scale visual tasks.

Simultaneously, diffusion models have emerged as a new generative model with powerful generation capabilities, achieving significant accomplishments, especially in the image domain. Guillaume, Jakob, and others proposed a semantic image editing method based on diffusion models [4], which automatically generates a mask by comparing the differences in the results of diffusion models under different text prompts. This method effectively annotates targets from images, providing new ideas and methods for the medical image segmentation field.

Furthermore, in the medical image segmentation field, due to subjective standards of doctors and objective cognitive differences among experts, a single image may have various segmentation solutions. These differences caused by insurmountable objective factors are referred to as Aleatoric Uncertainty [25]. However, current work tends to output a single segmentation result, ignoring the uncertainty of the predictions. Doctors hope that the model can provide a quantification of the uncertainty of the results, so they can focus their ef-

forts where the model's uncertainty is high, reducing repetitive work. Nowadays, different methods have been proposed to quantify the uncertainty in deep learning models, including evaluating uncertainty using Bayesian deep networks [6], estimating arbitrary uncertainty through network output [10], and using deep ensembles for simple and scalable uncertainty prediction [12], but the development has been somewhat limited.

Moreover, research on correcting segmentation results using the output uncertainty of models has also made some progress in fields like medical image segmentation. Yarin Gal et al. proposed Bayesian SegNet [11], which can correct segmentation results through Bayesian inference on the model's output uncertainty. Ren et al. reduce the occurrence of false positives or false negatives in segmentation results by considering the uncertainty of model outputs [18]. Zhang et al. optimized segmentation results by fusing segmentation outcomes from different modalities based on multi-modal data fusion segmentation results [27].

3 METHOD

3.1 Differentiation metrics

We propose the following method:

Given a dataset X that includes annotations for two categories, where c_0 represents the "healthy" annotation and c_1 represents the "unhealthy" annotation, as illustrated in Figure (1). A trained diffusion model f_θ can perform weakly-supervised semantic segmentation using only these category annotations. By employing variational inference techniques [8] to optimize the likelihood function, we can derive the loss function:

$$L = E_{(x,c)} E_{\bar{\alpha}_t} \|f_\theta(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\varepsilon, c, \bar{\alpha}_t) - \varepsilon\| \quad (1)$$

The Gaussian noise variance scales are set as $\{\beta_t \in (0, 1)\}_{t=1}^T$, with $\bar{\alpha}_t = \prod_{i=1}^t 1 - \beta_i$. Equation (1) is equivalent to maximizing the variational lower bound of the likelihood function [8].

Specifically, since the diffusion model $f_\theta(x_0, c, \bar{\alpha}_t)$ is conditioned on the category label embeddings c , switching the category label embeddings c during the generation process can steer the original image towards different image categories, thereby resulting in different output noises. The pixel-level differences between these two noises contain semantic information:

$$y = \text{Bin}(|f_\theta(x_t, c_0, \bar{\alpha}_t) - f_\theta(x_t, c_1, \bar{\alpha}_t)|, \delta) \quad (2)$$

where $\text{Bin}(\cdot, \delta)$ denotes the process of binarizing the model's output by setting a threshold θ . By measuring the differences in noise levels between c_0 and c_1 and binarizing the difference results, the segmentation outcome y corresponding to the location of malignant lesions for the categories can be obtained.

3.2 Aleatoric Uncertainty Metrics

To address the aleatoric uncertainties in the segmentation results, we quantify them from two perspectives: size of ambiguity areas and value of aleatoric uncertainty.

Aiming for this objective, we first leverage the characteristics of the diffusion model by sampling at different time steps n , achieving multiple output segmentation results, denoted as $Y \in \{y_1, y_2, \dots, y_t, \dots, y_n\}$.

Subsequently, we use the mean and variance as mathematical methods to represent the consistent and ambiguous areas within the

segmentation results, respectively. The mean of the set of segmentation results represents the model's average prediction for a given input image, as it reflects the most consistent part across all model outputs and can be viewed as the "consistent area" of model predictions. The variance measures the deviation of each segmentation result from the mean, indicating the variability among predictions. Therefore, a larger variance suggests greater differences in model predictions within that area, indicating stronger ambiguity. These areas may require additional attention in medical diagnostics, as the model's predictive confidence in these regions is lower, and different experts may have varying interpretations during image segmentation.

The specific formulas for mean and variance are as follows:

$$Y_{Coherence} = \bar{y} \quad (3)$$

$$Y_{Ambiguity} = \frac{1}{N} \sum_{i=1}^N (y_i - \bar{y})^2 \quad (4)$$

in which, y_i represents the segmentation outcomes at different noise addition levels.

Next, utilizing the concept of the Generalized Energy Distance (GED), we quantify the uncertainty of outputs from the result set, expressed by the formula:

$$D_{GED}^2 = \frac{2}{N^2} \sum_{i < j} d(y_i, y_j) - \frac{1}{N^2} \sum_{i=1}^N d(y_i, y_i) - \frac{1}{N^2} \sum_{j=1}^N d(y_j, y_j) \quad (5)$$

where $d(\cdot)$ is a distance function that represents the similarity between two segmentation results, and $y_i, y_j \in Y$. To intuitively measure the pixel-level prediction differences from the true annotations, we opt for the Euclidean distance for computation. The formula can be articulated as:

$$d(P, Q) = \sqrt{\sum_{i=1}^{w \times h} (p_i - q_i)^2} \quad (6)$$

Herein, p_i and q_i respectively denote the pixel points in two segmentation results. Since the simple Euclidean distance measure is influenced by the image size, the formula introduces a factor of the image size to normalize the distance.

$$d(P, Q) = \sqrt{\frac{\sum_{i=1}^{w \times h} (p_i - q_i)^2}{w \times h}} \quad (7)$$

Through the calculations described above, we can obtain the Generalized Energy Distance $D_{GED}^2 \in [0, \frac{\sqrt{2}}{2}]$, which serves as a metric for quantifying the differences between two probability distributions. It can be used to measure the distance between different prediction results within a set of segmentation outcomes. By calculating the GED, we can more accurately assess the uncertainty of the segmentation results and the diversity of model predictions. Larger values of GED indicate greater differences between segmentation results, whereas smaller values suggest more consistency among the results, further eliminating the interference caused by noise external to the target.

3.3 Segmentation results optimization

In the field of image segmentation, the Conditional Random Field (CRF) algorithm performs outstandingly [24], and it is often used in

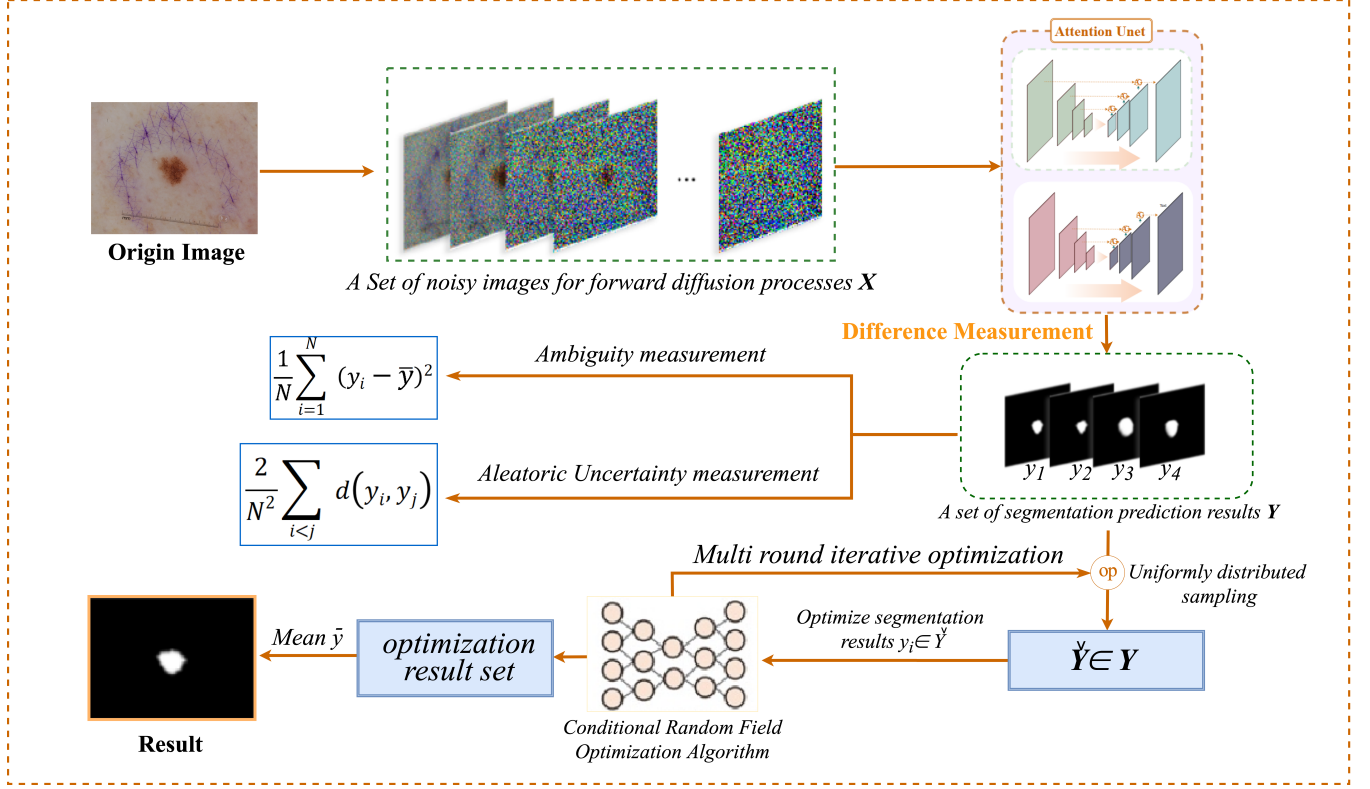


Figure 2. Strategies for measuring ambiguity and uncertainty in multi-output results and optimizing outcomes.

the post-processing stage of MIS to refine and optimize preliminary segmentation results. Therefore, for multiple output results, we adopt an improved version of the DenseCRF to optimize the results. DenseCRF considers the relationship between each pixel in the image and all other pixels, enabling it to capture finer details and a broader context. This allows it to provide more precise segmentation results in some cases.

The optimization process is as follows:

First, we initialize the set of segmentation results Y , where initially, the probability of each element being selected is set to an equal probability of $1/n$.

Second, the set of segmentation results undergoes K iterations. In each iteration, a subset $\tilde{Y}$ is randomly sampled from Y with a given probability. This step introduces randomness to explore different combinations of segmentation results in Y , thereby discovering and utilizing the complementary information among these results.

Third, DenseCRF is applied to each segmentation result in the sampled subset $\tilde{Y}$ to filter out noise and improve segmentation quality.

The core of DenseCRF is an energy function $E(\mathbf{x})$, as shown in formula 12. This function consists of two parts: a data term (Unary Potential), which can be directly obtained from the noise difference calculated by the formula (6) indicating the initial confidence of each pixel belonging to each category; and a smoothness term (Pairwise Potential), which depends on the relationship between pairs of pixels. This is usually defined using the intensity or position information of pixels in the original image, aiming to assign similar pixels to the same category while maintaining the clarity of edges. In this experiment, we design an appearance kernel based on the principle that nearby pixels with similar colors are likely to belong to the same class. The smoothness kernel is used to remove small isolated areas [21].

$$E(\mathbf{x}) = \sum_i \psi_u(x_i) + \sum_{i < j} \psi_p(x_i, x_j) \quad (8)$$

$$\psi_u(x_i) = -\log P(x_i) \quad (9)$$

$$\psi_p(x_i, x_j) = \mu(x_i, x_j) \left[\underbrace{w_1 \exp \left(-\frac{|p_i - p_j|^2}{2\theta_\alpha^2} - \frac{|l_i - l_j|^2}{2\theta_\beta^2} \right)}_{\text{appearance kernel}} + \underbrace{w_2 \exp \left(-\frac{|p_i - p_j|^2}{2\theta_\gamma^2} \right)}_{\text{smoothness kernel}} \right] \quad (10)$$

in which, $P(x_i)$ denotes the probability that pixel i belongs to label x_i , $\mu(x_i, x_j)$ is a penalty term for label inconsistency, which is 1 when $x_i \neq x_j$ and 0 otherwise. p_i and p_j represent the position coordinates of pixels i and j respectively, I_i and I_j represent the corresponding color intensity or feature vectors, w_1 , w_2 , θ_α , θ_β , and θ_γ model parameters used to adjust the influence of the Gaussian kernel.

The objective of DenseCRF is to achieve energy minimization, a process accomplished through the mean field algorithm. In this process, the category label of each pixel is adjusted in every iteration to reduce the overall energy.

$$E^{(t+1)}(\mathbf{x}) = \frac{1}{z_i} \exp \left(-\psi_u(x_i) - \sum_{j \neq i} \sum_{x_j} E^{(t)}(\mathbf{x}) \psi_p(x_i, x_j) \right) \quad (11)$$

In this formulation, $E^{(t)}(\mathbf{x})$ represents the probability distribution of the label x_i for pixel i at iteration step t , where Z_i is the normalization constant that ensures $E(\mathbf{x})$ is a probability distribution. This step is iterated until $E(\mathbf{x})$ converges or the maximum number of iterations is reached, marking the end of DenseCRF optimization.

Subsequently, the optimized segmentation results are averaged and then binarized to obtain the average prediction $\bar{y}_k$ for that iteration step. This step combines multiple optimized results to reduce bias and interference in individual predictions.

After completing all K iterations, further averaging is performed on the average predictions $\bar{y}_k$ obtained from all iteration steps, with the calculation formula being:

$$Y_{final} = \frac{1}{K} \sum_{k=1}^K \bar{y}_k \quad (12)$$

Through these operations, we can synthesize information obtained across all iterations, smooth out the segmentation results, and reduce noise and errors, thereby achieving a more accurate and stable final segmentation prediction.

This method allows us to utilize the information about the uncertainty in model predictions to iteratively refine and improve the segmentation results.

4 EXPERIMENTS

4.1 Model Training

In this experiment, we utilize the ISIC2018 challenge dataset [3], [23], which contains a large number of high-resolution images of skin lesions and their corresponding annotation information. After collecting the data, we conduct thorough data organization, which included 2500 images in the training set, 100 in the validation set, and 1000 in the test set. Furthermore, we use the annotations corresponding to the training set as masks to remove the lesion areas in the disease images, considering them as healthy skin, to aid the model in learning and acquiring features of healthy images.

To enhance the model's training effectiveness, we improve the data input method by adding various preprocessing techniques such as blurring, rotation, and sharpening to the input data.

During the model training process, based on Python 3.9.13 and PyTorch 2.0.1 frameworks, and conducted on an RTX 4070, we perform multiple experiments to identify well-performing hyperparameters. These included an image size of 128×128 , a noise addition count $t = 100$, a batch size of 4, a learning rate of 0.0005, and a training duration of 500 epochs. We train two diffusion models based on U-Net, with a total of 13.8M parameters for each training session, averaging a duration of 7 minutes and 11 seconds.

In the model testing phase, for the multiple outputs of the diffusion model, we set 10 time steps between 60 and 150, at intervals of 10, to quantify the consistency areas of segmentation results, ambiguity areas, calculate the GED values, and the subsequent optimization processes.

During the optimization process, we set three iterations, each time selecting 4 results from the outputs of 10 time steps for optimization calculations. Finally, we perform a second averaging calculation on the optimized results after three iterations.

4.2 Index Evaluation

The evaluation indicators we use in this experiment include Dice Coefficient, Jaccard Index, Precision and Recall. The calculation formula

as are as follows:

$$\text{Dice} = \frac{2 \times |X \cap Y|}{|X| + |Y|} \quad (13a)$$

$$\text{Jaccard} = \frac{|X \cap Y|}{|X \cup Y|} \quad (13b)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (13c)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (13d)$$

Herein, X represents the segmentation results predicted by the model, while Y represents the actual segmentation. Positive cases typically refer to the target areas that the model needs to identify and segment, namely, the lesion areas. Conversely, negative cases refer to areas other than the positive ones, namely, the non-target areas.

These metrics are greatly beneficial in assessing the model's performance. The value of the Dice coefficient ranges from 0 to 1, with higher values indicating greater similarity between the predicted segmentation and the actual segmentation, that is, better model performance. The Jaccard index also has values ranging from 0 to 1, with higher values indicating a higher degree of overlap between the predicted and actual segmentations, meaning better model performance. High precision means fewer false positives, that is, the model is more accurate in marking positive cases. A high recall rate implies that the model can capture more actual positive cases, thereby reducing missed detections.

5 RESULTS & DISCUSSION

5.1 Medical Image Segmentation

The figure below showcases several distinct examples from our experimental process. (see Fig. 3) It is observed that our experimental model can efficiently differentiate between skin and other interfering factors, such as hair and bubbles, which are disturbances during the image acquisition process. In the process of denoising with the diffusion model, it can treat interfering factors as noise and separate them from the skin, thereby focusing more on the differences between lesions and normal skin when measuring noise diversity. This approach results in image segmentation outcomes that are closer to the ground truth.

Particularly in the areas where skin meets hair and at the lesion borders, our model demonstrated superior performance compared to the traditional U-Net model. This strategy enhances the model's robustness and improves its capability to capture details, which is especially important when dealing with highly complex and variable data such as medical images.

5.2 Aleatoric Uncertainty Measurement

After completing the segmentation task, we utilize the model's multi-output capability to generate multiple different output results at various time steps. The specific output results are illustrated in the figure below:

By employing the formula (3) and (4), we calculate the mean and variance of the aforementioned ten output results. This allows us to identify the consistency and ambiguity regions of the segmentation results, which are specifically displayed in the following figure.

Additionally, according to the measurement method provided by the formula (5), the uncertainty value of the segmentation result can

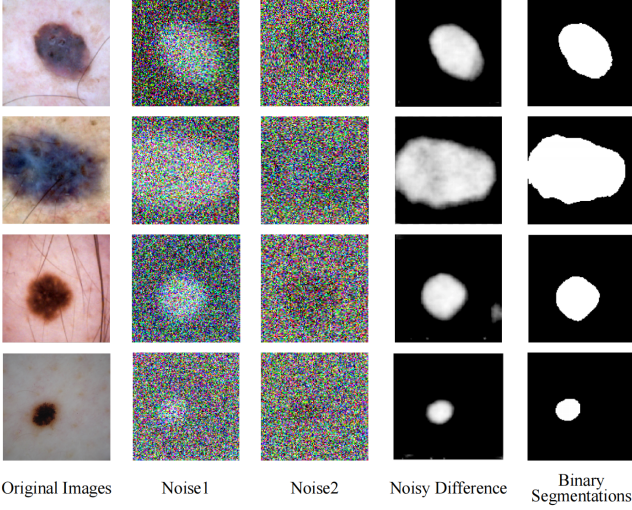


Figure 3. The experimental process includes the original images, noise under different semantics, noise difference grayscale images, and binarized segmentation.

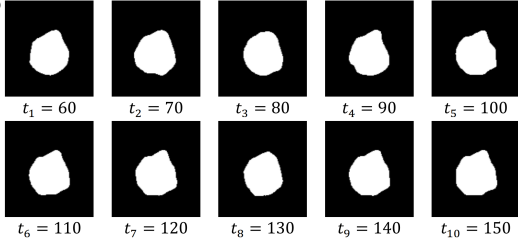


Figure 4. Segmentation results obtained under different noise addition iterations (from 60 to 150). Through testing, on the entire test set, the average uncertainty of our segmentation results is 0.13.

These results allow doctors to visually observe the model’s segmentation performance. In practical applications, doctors can use this information to focus on the ambiguous regions within the segmentation results, providing effective decision support for clinical diagnosis and further research. Moreover, our approach to uncertainty measurement with a defined range yields numerical results within a specific range, offering clear and effective assistance for subsequent optimization research.

5.3 Optimization for Multi-output

By analyzing the multiple output results generated by the model, we employ the DenseCRF algorithm to optimize these results. The specific outcomes are illustrated in the figure below.

Comparing the multiple output results in the figure 4, it is evident that through iterative optimization steps, we further smooth the segmentation results, reducing noise and errors to achieve more accurate and stable final segmentation predictions. This process not only enhanced the precision of the segmentation results but also made full use of the uncertainty information from the model’s predictions, offering a new avenue for precise medical image analysis.

After foregoing processes, we obtain the following results, as detailed in Table 1. Additionally, we list the outcomes of some segmentation models based on U-Net on the ISIC 2018 dataset.

From the table data, it is evident that compared to other U-Net-based models, our experimental model demonstrates superior performance in terms of the Dice coefficient and Precision.

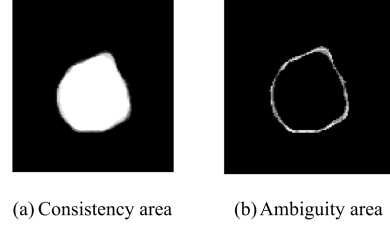


Figure 5. After calculation, the consistent and ambiguous regions are obtained.



Figure 6. After calculation, the consistent and ambiguous regions are obtained.

Method	Dice	Precision	Recall	Jaccard
U-Net	0.838	0.863	0.868	0.724
DU-Net	0.852	0.894	0.867	0.749
RU-Net	0.85	0.892	0.855	0.743
AttnU-Net	0.86	0.861	0.884	0.746
DiffSeg	0.864	0.897	0.882	0.735

6 CONCLUSION

In this study, we introduce a novel MIS model, known as DiffSeg, specifically designed to address two major challenges faced by traditional segmentation models in the context of processing highly complex medical images: One is the excessive dependency on large-scale annotated datasets; and the other is the difficulty in providing reliable uncertainty measurements. By introducing a denoising diffusion probability model, we skillfully leverage the model’s capability to extract features (i.e., noise) from images with different semantics and utilize the differences in image noise to achieve precise semantic segmentation under weak supervision conditions. This approach significantly improves segmentation accuracy and demonstrates the practicality of generative models in the field of MIS. Experimental results show that the DiffSeg model outperforms existing U-Net-based segmentation methods on the ISIC 2018 dataset, particularly in terms of the Dice coefficient and recall rate.

Furthermore, we explore a measurement method for the aleatoric uncertainty of segmentation results by utilizing the model’s multi-output capability to quantify the consistency and ambiguity regions of the segmentation results. We also define an uncertainty calculation method with a determined range using GED. Based on the multi-output results, we optimize the segmentation results using the DenseCRF algorithm. This strategy not only improves the accuracy of the segmentation results but also provides intuitive decision support for doctors, helping to reduce repetitive labor in practical applications.

In summary, the DiffSeg model proposed in this study offers a new research direction for the application of generative models in the field of medical image segmentation. Our future work will focus on further enhancing the model’s generalization ability to achieve higher accuracy and practicality in a broader range of medical image segmentation tasks. We hope our research will inspire more researchers and promote the development of advanced medical image analysis technologies.

7 ACKNOWLEDGMENTS

Project 202310611112 supported by National Training Program of Innovation and Entrepreneurship for Undergraduates.

References

- [1] M. Abdar, F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, P. Fieguth, X. Cao, A. Khosravi, and U. R. Acharya. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information fusion*, 76: 243–297, 2021. URL <https://www.sciencedirect.com/science/article/pii/S1566253521001081>. Publisher: Elsevier.
- [2] W. Bai, O. Oktay, M. Sinclair, H. Suzuki, M. Rajchl, G. Tarroni, B. Glocker, A. King, P. M. Matthews, and D. Rueckert. Semi-supervised learning for network-based cardiac MR image segmentation. In *Medical Image Computing and Computer-Assisted Intervention-MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part II* 20, pages 253–260. Springer, 2017. URL https://link.springer.com/chapter/10.1007/978-3-319-66185-8_29.
- [3] N. C. F. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler, and A. Halpern. Skin lesion analysis toward melanoma detection: A challenge at the 2017 International symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC). In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 168–172, Apr. 2018. doi: 10.1109/ISBI.2018.8363547. URL <https://ieeexplore.ieee.org/abstract/document/8363547>. ISSN: 1945-8452.
- [4] G. Couairon, J. Verbeek, H. Schwenk, and M. Cord. DiffEdit: Diffusion-based semantic image editing with mask guidance, Oct. 2022. URL <http://arxiv.org/abs/2210.11427>. arXiv:2210.11427 [cs].
- [5] N. Dong, M. Kampffmeyer, X. Liang, Z. Wang, W. Dai, and E. Xing. Unsupervised domain adaptation for automatic estimation of cardiothoracic ratio. In *Medical Image Computing and Computer Assisted Intervention-MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part II* 11, pages 544–552. Springer, 2018. URL https://link.springer.com/chapter/10.1007/978-3-030-00934-2_61.
- [6] Y. Gal and Z. Ghahramani. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In *Proceedings of The 33rd International Conference on Machine Learning*, pages 1050–1059. PMLR, June 2016. URL <https://proceedings.mlr.press/v48/gal16.html>. ISSN: 1938-7228.
- [7] R. M. Haralick and L. G. Shapiro. Image segmentation techniques. *Computer vision, graphics, and image processing*, 29(1): 100–132, 1985. URL <https://www.sciencedirect.com/science/article/pii/S0734189X85901537>. Publisher: Elsevier.
- [8] J. Ho, A. Jain, and P. Abbeel. Denoising Diffusion Probabilistic Models. In *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc., 2020. URL <https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584afd0967f1ab10179ca4b-Abstract.html>.
- [9] D. Jin, Z. Xu, Y. Tang, A. P. Harrison, and D. J. Mollura. CT-Realistic Lung Nodule Simulation from 3D Conditional Generative Adversarial Networks for Robust Lung Segmentation. In A. F. Frangi, J. A. Schnabel, C. Davatzikos, C. Alberola-López, and G. Fichtinger, editors, *Medical Image Computing and Computer Assisted Intervention - MICCAI 2018*, pages 732–740, Cham, 2018. Springer International Publishing. ISBN 978-3-030-00934-2. doi: 10.1007/978-3-030-00934-2_81.
- [10] A. Kendall and Y. Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017. URL <https://proceedings.neurips.cc/paper/2017/hash/2650d6089a6d640c5e85b2b88265dc2b-Abstract.html>.
- [11] A. Kendall, Y. Gal, and R. Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7482–7491, 2018. URL http://openaccess.thecvf.com/content_cvpr_2018/html/Kendall_Multi-Task_Learning_Using_CVPR_2018_paper.html.
- [12] B. Lakshminarayanan, A. Pritzel, and C. Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/9ef2ed4b7fd2c810847ffa5fa85bce38-Abstract.html.
- [13] C. Liang, W. Wang, J. Miao, and Y. Yang. GMMSeg: Gaussian Mixture based Generative Semantic Segmentation Models. *Advances in Neural Information Processing Systems*, 35:31360–31375, Dec. 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/cb1c4782f159b55380b4584671c4fd88-Abstract-Conference.html.
- [14] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.
- [15] F. Milletari, N. Navab, and S.-A. Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. Ieee, 2016. ISBN 1-5090-5407-3.
- [16] N. Otsu. A threshold selection method from gray-level histograms. *Automatica*, 11(285-296):23–27, 1975. URL https://dspace.tul.cz/bitstream/handle/15240/49062/_otsu1979.pdf?sequence=4.
- [17] J. Peng, G. Estrada, M. Pedersoli, and C. Desrosiers. Deep co-training for semi-supervised image segmentation. *Pattern Recognition*, 107: 107269, 2020. URL <https://www.sciencedirect.com/science/article/pii/S0031320320300741>. Publisher: Elsevier.
- [18] P. O. Pinheiro, T.-Y. Lin, R. Collobert, and P. Dollár. Learning to Refine Object Segments. In B. Leibe, J. Matas, N. Sebe, and M. Welling, editors, *Computer Vision - ECCV 2016*, volume 9905, pages 75–91. Springer International Publishing, Cham, 2016. ISBN 978-3-319-46447-3 978-3-319-46448-0. doi: 10.1007/978-3-319-46448-0_5. URL http://link.springer.com/10.1007/978-3-319-46448-0_5. Series Title: Lecture Notes in Computer Science.
- [19] M. Rajchl, M. C. Lee, O. Oktay, K. Kamnitsas, J. Passerat-Palmbach, W. Bai, M. Damodaram, M. A. Rutherford, J. V. Hajnal, and B. Kainz. Deepcut: Object segmentation from bounding box annotations using convolutional neural networks. *IEEE transactions on medical imaging*, 36(2):674–683, 2016. URL <https://ieeexplore.ieee.org/abstract/document/7739993/>. Publisher: IEEE.
- [20] P. K. Sahoo, S. Soltani, and A. K. Wong. A survey of thresholding techniques. *Computer vision, graphics, and image processing*, 41(2): 233–260, 1988. URL <https://www.sciencedirect.com/science/article/pii/0734189X88900229>. Publisher: Elsevier.
- [21] J. Shotton, J. Winn, C. Rother, and A. Criminisi. TextonBoost for Image Understanding: Multi-Class Object Recognition and Segmentation by Jointly Modeling Texture, Layout, and Context. *International Journal of Computer Vision*, 81(1):2–23, Jan. 2009. ISSN 0920-5691, 1573-1405. doi: 10.1007/s11263-007-0109-1. URL <http://link.springer.com/10.1007/s11263-007-0109-1>.
- [22] M. Tang, S. Valipour, Z. Zhang, D. Cobzas, and M. Jagersand. A deep level set method for image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 14, Proceedings 3*, pages 126–134. Springer, 2017. URL https://link.springer.com/chapter/10.1007/978-3-319-67558-9_15.
- [23] P. Tschandl, C. Rosendahl, and H. Kittler. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*, 5(1):180161, Aug. 2018. ISSN 2052-4463. doi: 10.1038/sdata.2018.161. URL <https://www.nature.com/articles/sdata2018161>. Publisher: Nature Publishing Group.
- [24] R. Vemulapalli, O. Tuzel, M.-Y. Liu, and R. Chellapa. Gaussian conditional random field network for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3224–3233, 2016. URL http://openaccess.thecvf.com/content_cvpr_2016/html/Vemulapalli_Gaussian_Conditional_Random_CVPR_2016_paper.html.
- [25] G. Wang, W. Li, M. Aertsen, J. Deprest, S. Ourselin, and T. Vercauteren. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. *Neurocomputing*, 338:34–45, Apr. 2019. ISSN 0925-2312. doi: 10.1016/j.neucom.2019.01.103. URL <https://www.sciencedirect.com/science/article/pii/S0925231219301961>.
- [26] F. Yu, J. Zhao, Y. Gong, Z. Wang, Y. Li, F. Yang, B. Dong, Q. Li, and L. Zhang. Annotation-Free Cardiac Vessel Segmentation via Knowledge Transfer from Retinal Images. In D. Shen, T. Liu, T. M. Peters, L. H. Staib, C. Essert, S. Zhou, P.-T. Yap, and A. Khan, editors, *Medical Image Computing and Computer Assisted Intervention - MICCAI 2019*, pages 714–722, Cham, 2019. Springer International Publishing. ISBN 978-3-030-32245-8. doi: 10.1007/978-3-030-32245-8_79.
- [27] Y. Zhang, J. Yang, J. Tian, Z. Shi, C. Zhong, Y. Zhang, and Z. He. Modality-Aware Mutual Learning for Multi-modal Medical Image Segmentation. In M. de Bruijne, P. C. Cattin, S. Cotin, N. Padoy, S. Speidel, Y. Zheng, and C. Essert, editors, *Medical Image Computing and Computer Assisted Intervention - MICCAI 2021*, pages 589–599, Cham, 2021. Springer International Publishing. ISBN 978-3-030-87193-2. doi: 10.1007/978-3-030-87193-2_56.
- [28] M. Zhao, L. Wang, J. Chen, D. Nie, Y. Cong, S. Ahmad, A. Ho, P. Yuan, S. H. Fung, H. H. Deng, J. Xia, and D. Shen. Craniomaxillofacial Bony Structures Segmentation from MRI with Deep-Supervision Adversarial

Learning. In A. F. Frangi, J. A. Schnabel, C. Davatzikos, C. Alberola-López, and G. Fichtinger, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, volume 11073, pages 720–727. Springer International Publishing, Cham, 2018. ISBN 978-3-030-00936-6 978-3-030-00937-3. doi: 10.1007/978-3-030-00937-3_82. URL https://link.springer.com/10.1007/978-3-030-00937-3_82. Series Title: Lecture Notes in Computer Science.

- [29] Y. Zhou, Y. Wang, P. Tang, S. Bai, W. Shen, E. Fishman, and A. Yuille. Semi-supervised 3D abdominal multi-organ segmentation via deep multi-planar co-training. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 121–140. IEEE, 2019. URL <https://ieeexplore.ieee.org/abstract/document/8658899/>.