Uncertainty-Guided Diffusion Model for Biomedical Image Segmentation

Ju Wang[1], Ying Wang[1], Xiaochen Yuan[1,*] and Guoheng Huang[2]
[1]Faculty of Applied Sciences, Macao Polytechnic University, Macao, China
[2]School of Computer Science and Technology, Guangdong University of Technology, Guangzhou, Guangdong, China
Email: p2423907@mpu.edu.mo, p2412918@mpu.edu.mo, xcyuan@mpu.edu.mo, kevinwong@gdut.edu.cn

***Abstract*—Biomedical image segmentation is crucial to precisely diagnose lesions on medical images. Recently, diffusion probabilistic models (DPMs) have proven to be effective for medical segmentation tasks. However, most diffusion-based methods apply uniform guidance across the entire image, which treats all regions equally during optimization and neglects regional differences in prediction difficulty. To address this issue, we propose UGD-Seg, a two-stage uncertainty-guided diffusion segmentation framework. In the first stage, an evidential network is employed to estimate a pixel-level uncertainty map while extracting multi-scale features from the input image. In the second stage, the multi-scale features are integrated and decoupled into certain and uncertain components under uncertainty guidance, which are then used to guide a latent diffusion model (LDM) for progressive segmentation. Extensive experiments on the ISIC2016 and CVC-ClinicDB datasets demonstrate that UGD-Seg consistently outperforms state-of-the-art methods, achieving more accurate boundary delineation and more stable segmentation predictions.**

***Index Terms*—Medical image segmentation, Diffusion model, Uncertainty estimation**

I. Introduction

Biomedical image segmentation provides essential support for disease diagnosis by enabling precise localization of lesion regions. Recent advances in deep learning have led to substantial improvements in the performance of medical segmentation. It is well-known that Convolutional Neural Networks (CNNs) [1]–[3] and Vision Transformer (ViT) models [4]–[6] have been successfully applied to medical image analysis, benefiting from their strong capability to learn discriminative features from medical images.

Beyond CNN-based and ViT-based methods, Diffusion Probabilistic Models (DPMs) have emerged as a novel paradigm for medical image segmentation tasks, owing to their effectiveness in modeling data distribution. Diffusion-based medical segmentation models can be formulated as a conditional image generation task, where image features are incorporated as conditioning signals to guide the diffusion process toward predicting segmentation masks. Recently, a growing number of studies have focused on designing diverse feature extractors to provide conditional guidance for Denoising Diffusion Probabilistic Model (DDPM) [7] in segmentation tasks [8]–[10]. However, DDPM operating in pixel space requires many denoising steps, which leads to slow and computationally expensive inference. To address this problem, Latent Diffusion Model (LDM) [11] has been introduced to perform diffusion in a low-dimensional latent space, significantly accelerating inference. In LDM-based segmentation, feature extraction and conditional guidance remain essential components [12]–[14]. Nevertheless, despite the differences in model formulation, most diffusion-based segmentation approaches rely on a uniform guidance strategy. This design ignores the fact that different regions exhibit varying levels of segmentation difficulty, which can be revealed by uncertainty estimation models [15]. Notably, uncertain regions often need additional focused refinement, whereas certain regions offer stable semantic cues and should remain consistent to preserve structural integrity. Applying a uniform treatment to these two conflicting regions may result in suboptimal optimization.

To account for regional differences in prediction confidence within the guidance signal, we propose UGD-Seg, a two-stage uncertainty-guided diffusion framework for biomedical image segmentation built upon an LDM. In the first stage, an evidential network is introduced to estimate uncertainty, explicitly characterizing region-wise segmentation difficulty through a pixel-level uncertainty map. In the second stage, an Uncertainty-gated Feature Decoupling Module (UFDM) is proposed to decouple uncertain and certain representations from a unified feature space by leveraging the estimated uncertainty map. The resulting uncertain and certain features are then integrated into the segmentation prediction at each diffusion step, serving as region-specific guidance for the diffusion model. In this way, the UFDM enables the diffusion model to focus refinement on ambiguous regions while preserving stable structures in confident regions, leading to more accurate and stable segmentation results. We validate UGD-Seg on ISIC2016 [16] and CVC-ClinicDB [17] datasets. UGD-Seg surpasses previous state-of-the-art (SOTA) methods, demonstrating its effectiveness and generalization ability.

In conclusion, our contributions are:

1) We propose UGD-Seg, a two-stage uncertainty-guided diffusion framework built upon an LDM for medical image segmentation, which overcomes the limitation of uniform guidance that ignores spatial differences in segmentation uncertainty.
2) We design an Uncertainty-gated Feature Decoupling Module (UFDM) that decouples uncertain and certain features and offers region-specific guidance for the diffu-

*Corresponding author: xcyuan@mpu.edu.mo

sion model.

3) Experimental results on two datasets demonstrate that UGD-Seg achieves mIoU improvement of 3.28% and 15.52% over the average of competing methods on ISIC2016 and CVC-ClinicDB, respectively.

II. Related Work

A. Diffusion-based methods for medical image segmentation

Denoising Diffusion Probabilistic Model (DDPM) is a generative model built on a Markov chain, which learns a reverse denoising process to progressively transform Gaussian noise into high-quality realistic data samples. Recently, DDPM has been explored for medical image segmentation. As a pioneering work in this study, MedSegDiff [9] was the first to apply DDPM to medical segmentation by proposing dynamic condition encoding and a feature frequency parser to reduce the impact of high-frequency noise. However, MedSegDiff relies on CNN-based encoders, and the potential of ViT for conditional feature extraction remains unexplored. Therefore, based on MedSegDiff, MedSegDiffv2 [10] further incorporates a transformer into the diffusion-based framework, proposing a spectrum-space transformer to bridge the gap between semantic conditions and anchor conditions. Nevertheless, due to the slow inference speed and high computational cost of DDPMs, LDM has been considered as a more efficient alternative [11]. LSegDiff is the first attempt to employ LDM for medical segmentation. To further improve inference efficiency, SD-Seg [13] employs a single-step reverse process together with a concatenated latent fusion technique to reduce the number of sampling steps. Nevertheless, these works primarily assume spatially uniform guidance for the diffusion model, neglecting that decision difficulty varies across different regions. In contrast, UGD-Seg introduces a pixel-level uncertainty map that indicates the prediction difficulty, guiding the diffusion model to pay more attention to challenging regions.

B. Uncertainty Estimation and Boundary Refinement

Uncertainty estimation provides a way to characterize the reliability of model predictions. Recently, there has been an increasing interest in uncertainty estimation for deep neural network. Common uncertainty estimation methods include evidential deep learning (EDL), Bayesian neural network (BNN), ensemble method, and test-time data augmentation (TTA) [18]. Among them, Bayesian learning with Monte Carlo Dropout [19] is widely adopted for uncertainty estimation. However, it requires multiple stochastic forward passes at test time, causing the high computational cost. In addition, ensemble-based uncertainty estimation [20] combines the outputs of multiple independently trained models, which further increases memory and computation overhead. Similarly, TTA [21] relies on repeated inference under different input perturbations, thus also leading to extra cost. Consequently, the above methods cannot estimate uncertainty with a single forward pass. This limitation motivates the exploration of EDL, which enables efficient uncertainty quantification without costly multiple forward passes. As stated by Zou et al. [22], they propose an end-to-end model that assumes the class probabilities of the segmentation follow a Dirichlet distribution and learns its distributional parameters directly from data. For better interpretability and robustness to noise in medical images, Li et al. [15] introduced a self-supervised uncertainty estimation framework with gradient-based and noise-based uncertainty supervision. Motivated by this work, we follow their design to estimate uncertainty more robustly under the ambiguous boundary and diverse noisy pattern in medical image.

III. Methodology

In this paper, we present a two-stage framework called UGD-Seg for precise segmentation of medical lesions, as illustrated in Fig. 1. In the first stage, we adopt the evidential network proposed by Li et al. [15] to quantify the uncertainty of the input images. It produces an uncertainty map A_{uc} that captures regional uncertainties caused by ambiguous and noisy regions. In the second stage, we extract the uncertainty map A_{uc} and multi-scale encoder features (f_1 to f_5) from the evidential network. The multi-scale features are first integrated through a Multi-scale Feature Fusion Module (MFFM) to obtain a unified representation f_{fu}. Subsequently, f_{fu} and the uncertainty map A_{uc} are sent into UFDM, which separates the feature space into an uncertain feature f'_{uc} and certain feature f'_{ce}. The decoupled features $\{f'_{ce}, f'_{uc}\}$ are then concatenated to construct the conditional guidance signal for the LDM.

A. Uncertainty Estimation for Biomedical Images

Biomedical images often contain ambiguous regions, which introduce high uncertainty and hinder segmentation models from producing confident and stable predictions in these areas. To address this issue, one of the key components in our framework is to estimate the uncertainty for each pixel. Following the work of [15], we adopt EDL to quantify predictive uncertainty. We adopt an evidential network f_θ to parameterize the evidence for each pixel. For the i-th pixel, its categorical probability vector p_i is modeled as a Dirichlet distribution:

$$\mathrm{Dir}(p_i|\alpha_i) = \frac{1}{\mathrm{B}(\alpha_i)} \prod_{k=1}^{K} p_{ik}^{\alpha_{ik}-1} \tag{1}$$

where $\alpha_i = e_i + 1$ represents the Dirichlet parameters, $\mathrm{B}(\boldsymbol{\alpha}_i)$ denotes the multivariate Beta function that normalizes the Dirichlet distribution, $e_i = f_\theta(X)$ is the predicted evidence for the input image X and K is the number of classes.

Consequently, the predictive uncertainty for each pixel u_i can be obtained by $u_i = \frac{K}{\sum_{k=1}^{K} \alpha_{ik}}$, yielding an uncertainty map $A_{uc} = \{u_i\}_{i=1}^{H\times W}$. Under this formulation, pixels with low evidence naturally correspond to higher uncertainty values, which aligns well with the ambiguous and noisy regions commonly observed in biomedical images.

B. Multi-scale Feature Fusion Module

The evidential network not only outputs the uncertainty map A_{uc}, but also offers a hierarchy of multi-scale features

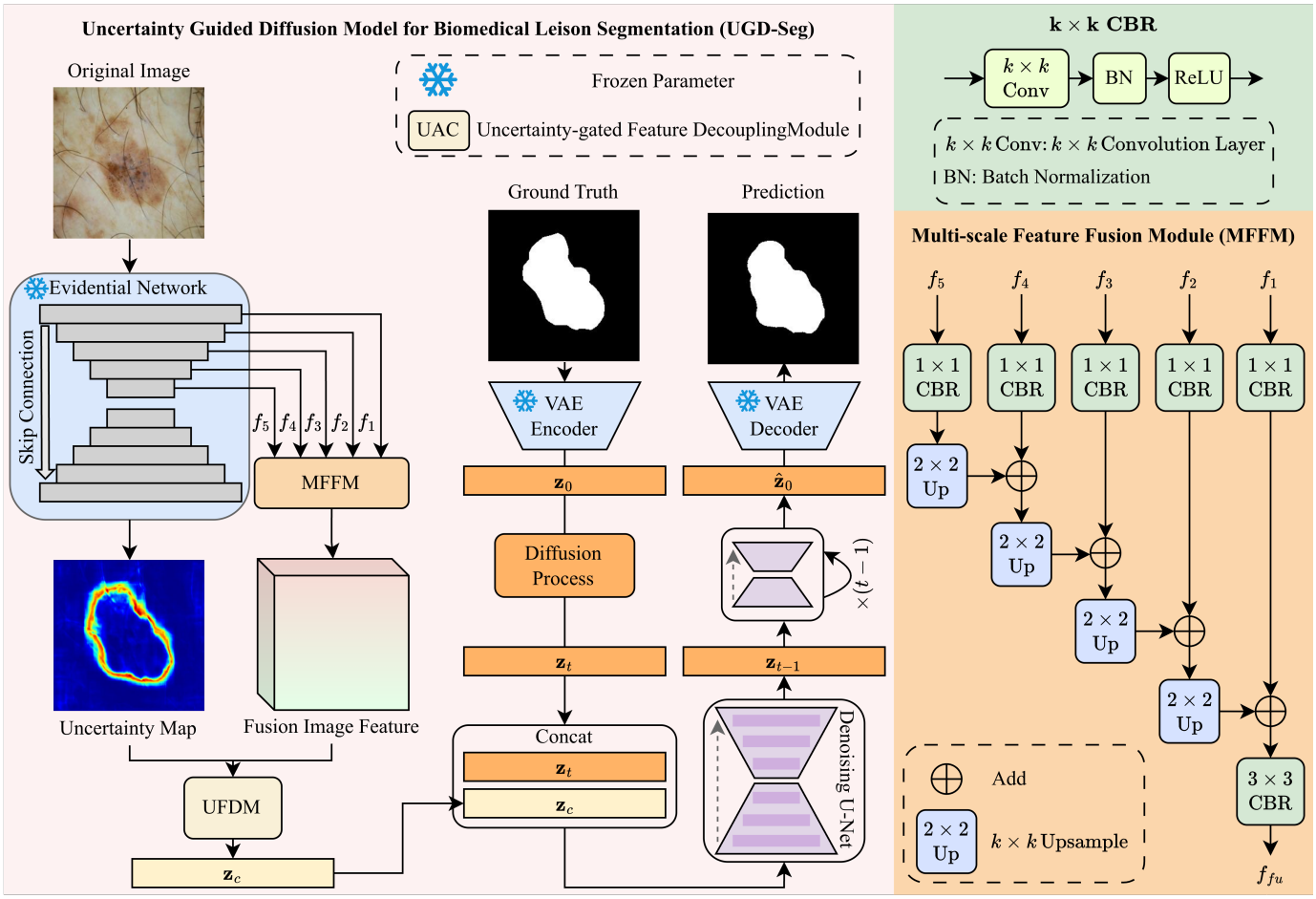

Fig. 1. Overall framework of the proposed UGD-Seg.

$(f_1, f_2, \ldots f_5)$ that encode abundant contextual information at different spatial resolutions. As shown in Fig 1, we introduce MFFM to integrate multi-scale encoder features from the evidential network into a unified feature space.

In the proposed MFFM, each feature map f_i is first processed by a 1×1 CBR block to align channel dimension and refine the feature representations. Specifically, a $k \times k$ CBR block consists of a $k \times k$ convolution followed by batch normalization and ReLU activation function. Subsequently, beginning with the deepest level feature f_5, the refined feature applies a 2×2 bilinear upsampling operation and an element-wise addition with the next shallower refined feature, repeatedly. The operation can be formulated as:

$$\tilde{f}_i = \mathrm{CBR}_{1\times1}(f_i), \quad i = 1, 2, \ldots, 5 \tag{2}$$

$$g_i = \begin{cases} \tilde{f}_i, & i = 5 \\ \tilde{f}_i \oplus \mathrm{Up}_{2\times2}(g_{i+1}), & i = 1, 2, 3, 4 \end{cases} \tag{3}$$

where $\oplus$ denotes element-wise addition. Finally, a 3×3 CBR block is applied to obtain the fused features f_{fu}.

C. Uncertainty-Gated Feature Decoupling Module

In medical image segmentation, uncertain regions often have a serious impact on segmentation accuracy. Meanwhile, certain regions provide reliable contextual cues. When modeled within a unified feature space, demands for semantic consistency and boundary sharpness impose conflicting constraints. This internal competition often results in over-smoothed predictions and unstable boundary localization. Therefore, we design the UFDM to decouple uncertain and certain features, which is illustrated in Fig. 2.

Building upon the uncertainty map A_{uc} derived from the evidential network, we design an uncertainty-gated mechanism that employs A_{uc} as a soft gate. By applying A_{uc} to the fused feature f_{fu} from MFFM, we first formulate the uncertainty gate G_u as:

$$G_u = \sigma\big(\mathrm{Conv}_{1\times1}(A_{uc})\big), \tag{4}$$

where $\sigma(\cdot)$ is the sigmoid function. The 1×1 convolution layer is used to adaptively project the raw uncertainty signal into a latent gating representation.

Next, the uncertainty gate G_u is utilized to separate f_{fu} into two distinct components: the uncertainty-aware feature

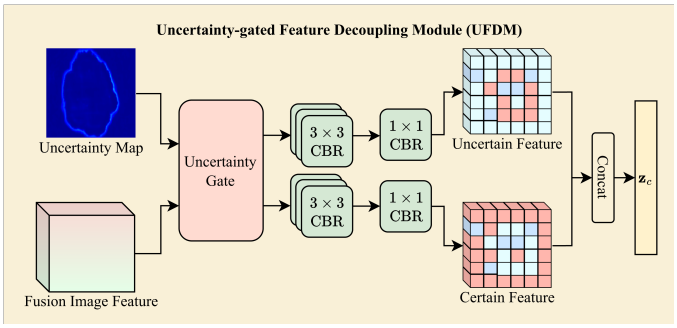

Fig. 2. The overview architecture of UFDM.

f_{uc} and certainty-preserving feature f_{ce}. This operation can be defined as:

$$f_{uc} = G_u \odot f_{fu} \tag{5}$$

$$f_{ce} = (1 - G_u) \odot f_{fu} \tag{6}$$

where $\odot$ denotes element-wise multiplication.

Both f_{uc} and f_{ce} are further processed by a series of 3×3 CBR blocks and a 1×1 CBR block to refine their representation, respectively. Through the proposed UFDM, the uncertain feature f'_{uc} is encouraged to emphasize ambiguous and noisy regions. And certain feature f'_{ce} focuses on stable and semantically confident regions. This separation allows the LDM to use f'_{ce} and f'_{uc} adaptively to balance robustness and boundary precision, leading to more accurate and stable segmentation results.

D. Uncertainty-guided Latent Diffusion Model

The diffusion model adopted in our method is LDM, which is computationally efficient by operating in a low-dimensional latent space. With the uncertain feature f'_{uc} and certain feature f'_{ce} obtained from UFDM, the LDM performs segmentation through an iterative denoising process under the joint guidance of both features. Specifically, certain feature provides stable global guidance. In addition, the uncertain feature is exploited to make a progressive boundary refinement in high-uncertainty regions.

As illustrated in Fig. 1, given a segmentation mask y, the LDM uses an encoder $\mathrm{Enc}(\cdot)$ to map it into a latent representation $z_0 = \mathrm{Enc}(y)$. A forward diffusion process then gradually adds Gaussian noise to z_0 via a Markov chain:

$$q(z_t \mid z_{t-1}) = \mathcal{N}\Big(z_t; \sqrt{1-\beta_t}\, z_{t-1},\, \beta_t\, \mathbf{I}\Big)\,, \tag{7}$$

where $\beta_t \in (0, 1)$ is a predefined noise schedule and $\mathbf{I}$ is the identity matrix. By composing the transitions, the noised latent representation at timestep t can be formulated as:

$$z_t = \sqrt{\bar{\alpha}_t}\, z_0 + \sqrt{1-\bar{\alpha}_t}\, \epsilon, \quad \epsilon \sim \mathcal{N}(0, \mathbf{I}) \tag{8}$$

where $\bar{\alpha}_t = \prod_{s=1}^{t}(1-\beta_s)$. During the reverse diffusion process, the LDM iteratively denoises the latent variable from z_t to z_{t-1}. To inject uncertainty-guided conditional information, the uncertain and certain features are concatenated to form a conditional representation:

$$z_c = \mathrm{concat}(f'_{uc}, f'_{ce}). \tag{9}$$

At each denoising step, z_c is concatenated with the latent representation and fed into the denoising U-Net ϵ_θ, which predicts the noise:

$$\hat{\epsilon} = \epsilon_\theta(z_t, t \mid z_c). \tag{10}$$

The reverse transition from z_t to z_{t-1} is modeled as:

$$p_\theta(z_{t-1} \mid z_t, z_c) = \mathcal{N}\big(z_{t-1};\, \mu_\theta(z_t, t, z_c),\, \sigma_t^2 \mathbf{I}\big)\,, \tag{11}$$

where σ_t^2 is fixed by the noise schedule, and the mean $\mu_\theta(\cdot)$ is computed from the predicted noise as:

$$\mu_\theta(z_t, t, z_c) = \frac{1}{\sqrt{1-\beta_t}}\left(z_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\,\hat{\epsilon}\right), \tag{12}$$

which allows sampling z_{t-1} from $p_\theta(z_{t-1} \mid z_t, z_c)$.

After T reverse steps, the final denoised latent $\hat{z}_0$ is mapped back to pixel space by the decoder $\mathrm{Dec}(\cdot)$ to obtain the predicted segmentation:

$$y_{\text{pred}} = \mathrm{Dec}(\hat{z}_0). \tag{13}$$

For the segmentation task, our proposed UGD-Seg is primarily trained using the standard denoising objective of diffusion models, denoted as the noise prediction loss $\mathcal{L}_{\text{noise}}$. In addition, we introduce a reconstruction loss $\mathcal{L}_{\text{seg}}$ to minimize the discrepancy between the predicted latent $\hat{z}_0$ and the ground-truth latent z_0, which further stabilizes the latent recovery during training. The $\mathcal{L}_{\text{seg}}$ can be formulated as:

$$\mathcal{L}_{\text{seg}} = \frac{1}{N}\|z_0 - \hat{z}_0\|_2^2 \tag{14}$$

where $\|\cdot\|$ denotes the ℓ_2 norm.

Besides, a cosine similarity loss $\mathcal{L}_{\text{orth}}$ is used to enforce orthogonality between the uncertain and certain features, thereby reducing feature interference and improving their functional separation during uncertainty-guided diffusion. The loss is defined as follows:

$$\mathcal{L}_{\text{orth}} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{f_{uc}^{'(i)} \cdot f_{ce}^{'(i)}}{\|f_{uc}^{'(i)}\|_2\, \|f_{ce}^{'(i)}\|_2}\right)^2 \tag{15}$$

where i indexes spatial locations (or feature vectors) and N denotes the total number of feature elements.

Therefore, the total loss is:

$$\mathcal{L}_{total} = \mathcal{L}_{\text{noise}} + \mathcal{L}_{\text{seg}} + \mathcal{L}_{\text{orth}} \tag{16}$$

IV. Experiments

A. Datasets and metrics

The proposed UGD-Seg is evaluated by two public datasets, consisting of ISIC2016 [16] and CVC-ClinicDB [17]. The ISIC2016 dataset comprises 1279 dermoscopic pathology images of skin lesions. The CVC-ClinicDB colonoscopy dataset includes 612 images for polyp segmentation tasks. To validate

TABLE I
PERFORMANCE COMPARISON WITH OTHER SOTA METHODS ON ISIC2016 AND CVC-CLINICDB.

Methods	ISIC2016						CVC-ClinicDB					
	mIoU ↑	mDSC ↑	HD95 ↓	Prec. ↑	Recall ↑	F1 ↑	mIoU ↑	mDSC ↑	HD95 ↓	Prec. ↑	Recall ↑	F1 ↑
U-Net [1]	78.49	86.58	26.71	88.98	88.10	88.53	59.97	69.75	49.16	71.03	75.48	73.19
UNeXt [3]	81.67	88.84	17.61	90.34	90.39	90.36	62.44	69.66	36.04	71.42	73.08	72.24
Rolling-UNet [23]	82.65	89.43	15.43	90.37	**91.15**	90.75	69.88	77.25	31.91	76.75	83.91	80.17
TransFuse [4]	84.10	90.42	**14.46**	**93.20**	90.57	**91.86**	69.62	78.45	36.45	76.11	92.53	83.52
TransUNet [6]	82.28	89.73	15.05	90.74	90.86	90.80	84.30	89.36	16.98	88.43	**95.39**	**91.77**
MedSegDiffV2 [10]	78.28	86.25	21.35	90.60	87.44	88.99	82.52	89.31	**11.23**	**92.43**	89.06	90.71
Ours	**84.53**	**91.15**	14.51	92.96	90.09	91.50	**86.98**	**91.91**	17.60	89.80	93.51	91.61

the performance of our framework, we use standard medical image segmentation metrics such as mean Intersection over Union (mIoU) [24], mean Dice Score (mDSC) [25], HD95 [26], Precision (Prec.), Recall, and F1 Score (F1).

B. Implementation Details

All experiments were conducted on NVIDIA GeForce RTX4090D GPU. During training, all input images were resized to 256×256 pixels. Data augmentation strategies including random horizontal flipping and vertical flipping were applied in our method. The batch size was set to 16, and the Adam optimizer was used for model optimization with an initial learning rate of 1×10^{-4}. During inference, a DDIM [27] sampler was adopted to accelerate the sampling process, with the number of diffusion timesteps set to 50.

C. Comparison with other State-of-the-Art Methods

To comprehensively evaluate the effectiveness of the proposed UGD-Seg, we compare it with several state-of-the-art deep learning segmentation methods, including U-Net [1], UNext [3], Rolling-UNet [23], TransUNet [6], TransFuse [4] and MedSegDiff V2 [10].

1) Results on ISIC2016 dataset: Table I presents the quantitative comparison of the different methods on ISIC2016. Benefiting from the proposed uncertainty-guided diffusion framework, our method achieves the superior overall performance, increasing the average mIoU by 3.28% and the mDSC by 2.61% compared to the competing approaches. In particular, UGD-Seg attains better boundary accuracy, as reflected by its competitive HD95 score. The visual segmentation result of our model is shown in Fig. 3. As illustrated, the segmentation results produced by UGD-Seg exhibit fine-grained boundary prediction, especially in regions with high-uncertainty contours.

2) Results on CVC-ClinicDB: Table I shows the quantitative comparison of the different methods on CVC-ClinicDB. As shown, our method achieves the improved overall performance in terms of mIoU and mDSC, demonstrating its generalizability across different datasets. Although some methods perform better on specific metrics, UGD-Seg maintains a more balanced performance across all evaluation metrics.

D. Ablation Study

We conducted extensive ablation experiments on the ISIC2016 dataset to validate the contribution of UFDM, and

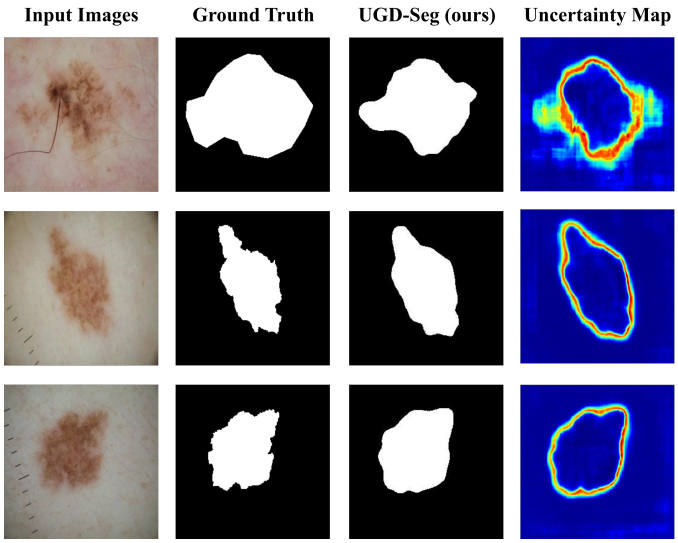

Fig. 3. Visual segmentation results on ISIC2016 of our UGD-Seg.

TABLE II
OUR ABLATION STUDY OF UGD-SEG ON THE ISIC2016. THE RESULTS EVALUATE THE CONTRIBUTIONS OF THE UNCERTAINTY MAP AND THE DECOUPLED CERTAIN/UNCERTAIN FEATURES.

Index	Model	mIoU ↑	mDSC ↑	HD95 ↓
(a)	UGD-Seg	**84.53**	**91.15**	**14.51**
(b)	w/o uncertainty map	83.93	90.65	18.2
(c)	w/o uncertainty feature	17.34	27.89	83.41
(d)	w/o certain feature	84.09	90.77	17.08

the results are summarized in Table II. As shown in Table II(a), the complete UGD-Seg achieves the best overall performance. When the uncertainty map is removed (Table II(b)), performance consistently drops across all metrics, indicating that using only a uniform feature without decoupling certain and uncertain information is insufficient. When removing the uncertain feature branch (Table II(c)) leads to a pronounced degradation, since uncertain feature encodes information from ambiguous boundary regions. Without uncertain feature, the diffusion model lacks effective boundary guidance and tends to produce distorted contours. In contrast, removing the certain feature branch (Table II(d)) results in a milder performance decline, suggesting that uncertain feature mainly drives boundary refinement, whereas certain feature provides stable contextual cues. Overall, these ablation experiments confirm that decou-

pling and jointly leveraging certain and uncertain features yield more accurate and stable segmentation than either using a single branch or mixing them in a unified feature space.

V. Conclusion

In this paper, we proposed a novel uncertainty-guided diffusion model for biomedical image segmentation (UGD-Seg). We adopt an evidential network to estimate uncertainty maps of medical images, which serves as a gate to distinguish uncertain and certain regions. Besides, we introduce a Multi-scale Feature Fusion Module (MFFM) which effectively fuses features from different scales. In addition, by designing an Uncertainty-gated Feature Decoupling Module (UFDM), we can separate the fused feature into uncertain and certain features, enabling the LDM to achieve region-aware refinement for different regions. Experimental results on two publicly available biomedical datasets demonstrate that UGD-Seg consistently improves segmentation performance across standard metrics. In particular, it achieves mIoU improvements of 3.28% and 15.52% and mDSC improvements of 2.61% and 12.95% over the average of competing methods on ISIC2016 and CVC-ClinicDB, respectively.

In future work, we plan to investigate advanced feature decoupling strategies to better separate uncertain and certain representations. We also aim to jointly optimize the uncertainty estimation network and the LDM, potentially improving consistency between uncertainty prediction and diffusion guidance. Finally, we will explore more efficient diffusion sampling strategies to achieve faster yet accurate segmentation.

Acknowledgments

This work was supported by the Science and Technology Development Fund of Macau SAR under grant 0053/2025/RIB2, and the Macao Polytechnic University under grant RP/FCA-12/2025.

References

[1] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[2] S. Wang, L. Li, and X. Zhuang, "Attu-net: attention u-net for brain tumor segmentation," in *International MICCAI brainlesion workshop*. Springer, 2021, pp. 302–311.

[3] J. M. J. Valanarasu and V. M. Patel, "Unext: Mlp-based rapid medical image segmentation network," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2022, pp. 23–33.

[4] Y. Zhang, H. Liu, and Q. Hu, "Transfuse: Fusing transformers and cnns for medical image segmentation," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2021, pp. 14–24.

[5] J. M. J. Valanarasu, P. Oza, I. Hacihaliloglu, and V. M. Patel, "Medical transformer: Gated axial-attention for medical image segmentation," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2021, pp. 36–46.

[6] J. Chen, J. Mei, X. Li, Y. Lu, Q. Yu, Q. Wei, X. Luo, Y. Xie, E. Adeli, Y. Wang *et al.*, "Transunet: Rethinking the u-net architecture design for medical image segmentation through the lens of transformers," *Medical Image Analysis*, vol. 97, p. 103280, 2024.

[7] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[8] J. Wolleb, R. Sandkühler, F. Bieder, P. Valmaggia, and P. C. Cattin, "Diffusion models for implicit image segmentation ensembles," in *International conference on medical imaging with deep learning*. PMLR, 2022, pp. 1336–1348.

[9] J. Wu, R. Fu, H. Fang, Y. Zhang, Y. Yang, H. Xiong, H. Liu, and Y. Xu, "Medsegdiff: Medical image segmentation with diffusion probabilistic model," in *Medical Imaging with Deep Learning*. PMLR, 2024, pp. 1623–1639.

[10] J. Wu, W. Ji, H. Fu, M. Xu, Y. Jin, and Y. Xu, "Medsegdiff-v2: Diffusion-based medical image segmentation with transformer," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 38, no. 6, 2024, pp. 6030–6038.

[11] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[12] H. Vu Quoc, T. Tran Le Phuong, M. Trinh Xuan, and S. Dinh Viet, "Lsegdiff: A latent diffusion model for medical image segmentation," in *Proceedings of the 12th International Symposium on Information and Communication Technology*, 2023, pp. 456–462.

[13] T. Lin, Z. Chen, Z. Yan, W. Yu, and F. Zheng, "Stable diffusion segmentation for biomedical images with single-step reverse process," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2024, pp. 656–666.

[14] X. Liu, Y. Zhao, S. Wang, and J. Wei, "Transdiff: medical image segmentation method based on swin transformer with diffusion probabilistic model," *Applied Intelligence*, vol. 54, no. 8, pp. 6543–6557, 2024.

[15] Y. Li, A. Sui, F. Wu, and X. Zhuang, "Uncertainty-supervised interpretable and robust evidential segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2025, pp. 649–658.

[16] D. Gutman, N. C. Codella, E. Celebi, B. Helba, M. Marchetti, N. Mishra, and A. Halpern, "Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (isbi) 2016, hosted by the international skin imaging collaboration (isic)," *arXiv preprint arXiv:1605.01397*, 2016.

[17] J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, D. Gil, C. Rodríguez, and F. Vilariño, "Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians," *Computerized medical imaging and graphics*, vol. 43, pp. 99–111, 2015.

[18] K. Zou, Z. Chen, X. Yuan, X. Shen, M. Wang, and H. Fu, "A review of uncertainty estimation and its application in medical imaging," *Meta-Radiology*, vol. 1, no. 1, p. 100003, 2023.

[19] T. Nair, D. Precup, D. L. Arnold, and T. Arbel, "Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation," *Medical image analysis*, vol. 59, p. 101557, 2020.

[20] A. Mehrtash, W. M. Wells, C. M. Tempany, P. Abolmaesumi, and T. Kapur, "Confidence calibration and predictive uncertainty estimation for deep medical image segmentation," *IEEE transactions on medical imaging*, vol. 39, no. 12, pp. 3868–3878, 2020.

[21] 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*, vol. 338, pp. 34–45, 2019.

[22] K. Zou, X. Yuan, X. Shen, M. Wang, and H. Fu, "Tbrats: Trusted brain tumor segmentation," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2022, pp. 503–513.

[23] Y. Liu, H. Zhu, M. Liu, H. Yu, Z. Chen, and J. Gao, "Rolling-unet: Revitalizing mlp's ability to efficiently extract long-distance dependencies for medical image segmentation," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 38, no. 4, 2024, pp. 3819–3827.

[24] 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*, 2015, pp. 3431–3440.

[25] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, pp. 297–302, 1945.

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

[27] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," *arXiv preprint arXiv:2010.02502*, 2020.