UniSegDiff: Boosting Unified Lesion Segmentation via a Staged Diffusion Model

Yilong Hu[1], Shijie Chang[1], Lihe Zhang[1] (✉), Feng Tian[2], Weibing Sun[2], and Huchuan Lu[1]

[1] Dalian University of Technology, Dalian, China
zhanglihe@dlut.edu.cn

[2] Department of Urology, Affiliated Zhongshan Hospital of Dalian University, Dalian, China

Abstract. The Diffusion Probabilistic Model (DPM) has demonstrated remarkable performance across a variety of generative tasks. The inherent randomness in diffusion models helps address issues such as blurring at the edges of medical images and labels, positioning Diffusion Probabilistic Models (DPMs) as a promising approach for lesion segmentation. However, we find that the current training and inference strategies of diffusion models result in an uneven distribution of attention across different timesteps, leading to longer training times and suboptimal solutions. To this end, we propose UniSegDiff, a novel diffusion model framework designed to address lesion segmentation in a unified manner across multiple modalities and organs. This framework introduces a staged training and inference approach, dynamically adjusting the prediction targets at different stages, forcing the model to maintain high attention across all timesteps, and achieves unified lesion segmentation through pre-training the feature extraction network for segmentation. We evaluate performance on six different organs across various imaging modalities. Comprehensive experimental results demonstrate that UniSegDiff significantly outperforms previous state-of-the-art (SOTA) approaches. The code is available at https://github.com/HUYILONG-Z/UniSegDiff.

Keywords: Diffusion model · Unified Lesion Segmentation · Staged training and inference.

1 Introduction

Lesion segmentation is a critical task in medical image analysis. However, existing neural network models are typically designed for specific imaging modalities and lesion tasks [13, 31, 30, 21], which limits their broader applicability. Therefore, developing a unified model capable of handling multiple imaging modalities and lesion types is essential. In medical imaging, boundary ambiguity often arises in both images and labels [17]. To address this, we use Diffusion Probabilistic Models (DPMs) [9] for medical lesion segmentation, as they incorporate randomness in modeling and can capture complex distributions. However, we observed that directly applying diffusion models to lesion segmentation tasks leads to longer convergence times, inference times, and suboptimal results, due to the uneven attention distribution across different timesteps. Through an in-depth analysis of the characteristics exhibited by diffusion models during training, we identified the root cause of the issue and developed a targeted staged diffusion framework, which was then applied to unified lesion segmentation.

When diffusion models are applied to segmentation tasks, they typically consist of two parts: the conditional feature extraction network and the denoising network [3]. The former encodes the image into conditional features to guide the latter in denoising training. The training and inference process is described as a Markov chain consisting of T timesteps. As t increases, the original mask x_0 is gradually corrupted by noise ϵ until it becomes pure Gaussian noise. The denoising network learns the ability to generate reconstructions by predicting ϵ or x_0 from the noisy mask x_t. Our observations indicate that predicting ϵ requires more training time to converge compared to predicting x_0. This is because, when the prediction target is ϵ, the model finds it easier to learn the distribution of noise from noisy masks x_t at larger timesteps than from those at smaller timesteps. As a result, the model tends to focus more on the latter (low-noise x_t). However, during inference, the model starts with pure Gaussian noise at the highest timestep and gradually denoises. The steps with larger timesteps are crucial in shaping the basic structure of the segmentation mask, which requires additional training for the model to converge at higher-noise timesteps. When the prediction target is x_0, the model tends to focus more on noisy masks at larger timesteps. Although the model can converge more quickly, it fails to model the noisy masks at smaller timesteps adequately, leading to poor performance. The upper-left part of Figure 1 shows the average gradient distribution of the model across timesteps for different prediction targets, where higher values indicate greater attention from the model during that phase.

Moreover, applying diffusion models to unified lesion segmentation introduces new challenges. Different lesion images vary greatly in imaging modalities, lesion morphology, and other aspects, while the masks are simple binary images. This causes inevitable confusion of features from different lesions when using the lesion images as conditional guidance for denoising, leading to a mismatch between the conditional features and the denoising features.

To address these challenges, we propose a new framework called UniSegDiff. First, we divide different timesteps into three stages and dynamically set prediction targets: the Rapid Segmentation Stage (predicting x_0), the Probabilistic Modeling Stage (predicting both x_0 and ϵ), and the Denoising Refinement Stage (predicting ϵ), ensuring the model maintains high attention across all timesteps. Next, the conditional feature extraction network is pre-trained for segmentation on the unified lesion dataset and frozen during denoising training. This transforms lesion images from different modalities into distributions similar to the masks, reducing feature confusion between different lesions and better utilizing the conditional features to guide denoising. Finally, to fully leverage the inherent randomness modeled by the diffusion model, we use staged inference to quickly generate multiple segmentation results for fusion, obtaining the optimal solution.

Our UniSegDiff achieves state-of-the-art performance on six lesion segmentation tasks across different modalities, as well as on the unified lesion segmentation task composed of these datasets.

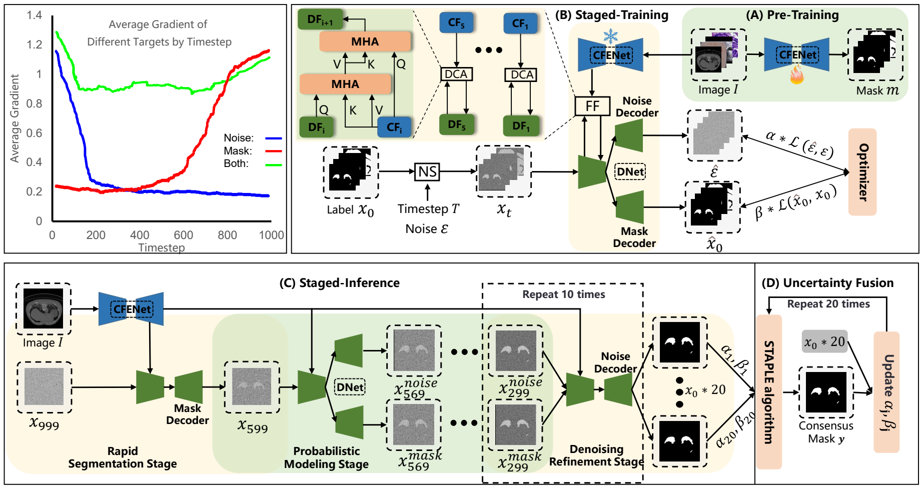

Fig. 1: The top-left corner shows the relationship between the model's average gradient and timestep for different targets. (A) Pre-Training and (B) Stage-Training represent the training process of UniSegDiff, while (C) Staged-Inference and (D) Uncertainty Fusion illustrate the inference process. NS denotes Noise Schedule, FF stands for Features Fusion, DCA refers to Dual Cross Attention, and CF, DF represent Conditional Feature and Denoising Feature, respectively.

2 Approach

2.1 Overall Architecture

The focus of this paper is on designing a diffusion framework for unified lesion segmentation, so our model architecture is kept simple, as shown in Figure 1. It consists of 2.5 UNet networks and a features fusion module. One UNet serves as the conditional feature extraction network (CFENet), while the remaining UNets function as the denoising network (DNet). During a single training step, CFENet extracts conditional features $CF_i \quad (i = 1 \sim 5)$ from the input images. The encoder of DNet takes the noisy masks x_t, added by the noise scheduler, as input and progressively receives CF_i as conditional guidance. Finally, the two decoders of DNet separately learn to model the ϵ and the x_0.

2.2 Train and Infer Stage

UniSegDiff divide the training process into three stages, each with distinct primary prediction objectives designed to ensure the neural network maintains high attention throughout all training steps. As shown in part (B) of Figure 1, in the Rapid Segmentation Stage ($599 < t$), the primary prediction target is the original mask x_0, as predicting the noise ϵ distribution is much simpler than predicting the original mask x_0 at this stage. Additionally, since the distribution difference of noisy masks x_t at different time steps is minimal during this phase, all time steps are set to the maximum value ($t = 999$). Surprisingly, this not only accelerated the convergence speed but also improved segmentation accuracy. In the Probabilistic Modeling Stage ($299 < t \leq 599$), the noise and mask information are more balanced, allowing the diffusion model to fully utilize its learning capability. At this stage, both prediction targets are given equal weight. In the Denoising Refinement Stage ($t \leq 299$), the primary prediction target is the noise ϵ. Similar to the Rapid Segmentation Stage, the distribution difference of noisy masks x_t at different time steps is minimal, so all time steps are set to the minimum value ($t = 0$). The loss function of UniSegDiff is as follows:

$$\mathcal{L}_{total} = \alpha \mathcal{L}_n + \beta(\mathcal{L}_{dice} + \mathcal{L}_{ce}). \tag{1}$$

The loss function consists of the noise prediction loss ($\mathcal{L}n$) and the original mask prediction loss ($\mathcal{L}dice + \mathcal{L}_{ce}$), weighted accordingly. The weight coefficients α and β are dynamically adjusted across different stages: in the Rapid Segmentation Stage ($\alpha : \beta = 1 : 3$), the Probabilistic Modeling Stage ($\alpha : \beta = 1 : 1$), and the Denoising Refinement Stage ($\alpha : \beta = 3 : 1$). This dynamic weighting scheme is consistent with our staged training approach.

The inference process is shown in part (C) of Figure 1, the initial time step of DNet is set to $t = 999$, with the input X_{999} being pure Gaussian noise. After obtaining the conditional features CF_i from CFENet, the mask prediction branch directly samples X_{999} to X_{599} in a single step. The subsequent sampling follows the DDIM method [24], with a step interval of 30. After sampling X_{599} ten times in each of the two decoder branches, the results are X_{299}^{mask} and X_{299}^{noise}. Finally, the two noisy masks at $t = 299$ are each sampled ten times by the noise prediction branch, with each step directly sampling from X_{299} to X_0. The 20 generated masks form a set of results, which are then prepared for subsequent uncertainty fusion. The entire sampling process is completed.

2.3 Pre Train and Condition Injection

To achieve unified lesion segmentation based on diffusion models, it is essential to eliminate the mismatch between conditional features and denoising features across different lesions. This requires the model to handle images from multiple modalities simultaneously and smoothly inject the conditional features of each lesion into the corresponding denoising features of the DNet. To this end, as shown in part (A) of Figure 1, we pre-train the CFENet on the unified dataset for the segmentation task and freeze it during the DNet training. This ensures that images from different modalities are transformed into distributions similar to the masks before being injected into DNet, narrowing the distribution gap between modalities and guiding DNet with the same set of features. This provides an appropriate prediction range for DNet, enabling it to refine and generate optimal results. During the stepwise injection of conditional features, we integrate them using the DCA (Dual Cross-Attention) module. The DCA module consists of two cascaded cross-attention blocks, with conditional features and noise mask features alternating as queries.

2.4 Uncertainty Fusion

For a lesion image I, we obtain a set of masks $z_j \quad (j = 1 \sim 20)$ through multiple samplings. To improve the model's accuracy and robustness, as shown in part (D) of Figure 1, we use the STAPLE [27] algorithm to iteratively generate a consensus mask y. The confidence values α_j and β_j for each z_j are initialized to 0.9 and 0.1, respectively, representing the probabilities of correctly labeling the target and incorrectly labeling the background as the target. For each pixel i, the initial value of y_i belongs to the target is set to 50%. The posterior probability of y_i is calculated using the 20 masks z_j through Equation.2. Then, using the maximum likelihood function from Equation.3, α_j and β_j are updated based on new y_i. This update process is repeated 20 times to obtain the final consensus mask y.

$$P(y_i = 1|\{z_{ij}\}) = \frac{\prod_{j=1}^{20} P(z_{ij}|y_i = 1, \alpha_j, \beta_j)P(y_i = 1)}{\sum_{y'_i \in \{0,1\}} \prod_{j=1}^{20} P(z_{ij}|y'_i, \alpha_j, \beta_j)P(y'_i)} \tag{2}$$

$$\hat{\alpha}_j, \hat{\beta}_j = \frac{\sum_{i=1}^{n} P(y_i = \theta|\{z_{ij}\})}{\sum_{i=1}^{n} P(z_{ij} = 1|y_i = \theta, \alpha_j, \beta_j)}, \quad \theta \in \{0, 1\} \tag{3}$$

3 Experiments

3.1 Datasets and Implementation Details

We selected six publicly available and widely used lesion segmentation datasets from different organs and modalities to form a unified lesion segmentation dataset. The details are provided in Table 1. For colon polyp segmentation, we follow the setting in Spider [29], combining five datasets to increase the challenge. Each dataset was randomly split into four equal parts for 4-fold cross-validation. For evaluation, we used two common metrics: mean Intersection over Union (mIoU) and mean Dice Similarity Coefficient (mDice). Detailed experimental setup, including the platform and hyperparameter settings, can be found in Table 2.

Table 1: The dataset information of the six lesion segmentation tasks.

Task	Dataset	Modality	Images
Wet-AMD	AMD-SD [10]	OCT	3049
Brain-Tumor	BTD [5, 6]	MR-TI	3064
Adenocarcinoma	EBHI-Seg [22]	Pathology image	795
Colon Polyp	Five datasets [25, 26, 23, 11, 7]	Endoscopy image	2248
Lung Infection	COVID-19 [12, 1]	CT	1277
Breast Lesion	BUSI [2]	Ultrasound	647

Table 2: Implementation Details

Category	Details
Framework	PyTorch
Hardware	4 × 3090 GPUs
Image Resolution	256 × 256
Optimizer	AdamW
Lr Scheduler	CosineAnnealingLR
Initial Lr	$1e^{-4}$
Total Epochs	300
Batch Size	64

3.2 Evaluation

Comparison with State-of-the-Arts To validate the effectiveness of UniSegDiff, we compared it with SOTA discriminative segmentation methods [19, 4, 15, 20, 18] and diffusion-based segmentation methods [28, 14, 8] on both the unified lesion segmentation task and six individual lesion segmentation tasks. The quantitative results are presented in Table 3. UniSegDiff consistently outperforms all models across both single-task and unified tasks. In the unified lesion segmentation task, all models showed a significant performance decline. However, thanks to the pre-training of CFENet, UniSegDiff reduced the distribution gap between datasets, resulting in no noticeable performance drop in the unified lesion segmentation task. As a result, it outperformed other methods by a considerable margin.
Defect analysis All methods showed a significant performance drop on the Lung Infection task during unified lesion segmentation. After examining the dataset, we found that this was due to a large number of masks being empty (approximately one-third of the dataset). We will clean the data and re-validate the results in future work.

3.3 Ablation Study

In this section, we examine how different denoising methods influence segmentation performance, as well as training and inference speed. We also analyze the impact of threshold selection in our proposed staged training method and the contribution of each component in the network. All experiments were conducted on the unified lesion segmentation task. Due to space limitations in the table, we only present the average values of the metrics across all datasets for the unified segmentation task, without showing the standard deviation.
Effectiveness of denoising methods Table 4 compares different denoising training strategies for the diffusion model. In traditional uniform denoising, predicting noise takes significantly more training epochs to converge. Direct original mask prediction accelerates convergence but still requires at least 100 inference steps for satisfactory results [24]. One-step denoising [16] achieves faster training and inference but performs better for original mask prediction than for

Table 3: The quantitative comparisons across various lesion segmentation tasks. From left to right in Table 3, the six tasks are those listed in Table 1. The values following ± represent the standard deviation.

Methods	**WA**		**BT**		**ADC**		**CP**		**LI**		**BL**	
	mDice	mIoU	mDice	mIoU	mDice	mIoU	mDice	mIoU	mDice	mIoU	mDice	mIoU
Individual Lesion Segmentation Tasks												
UNet	86.8 ±0.51	77.6 ±0.69	81.8 ±0.8	73.2 ±0.63	91.6 ±0.47	85.2 ±0.56	82.6 ±0.44	75.5 ±0.52	73.1 ±1.83	65.4 ±1.64	74.9 ±0.72	65.5 ±0.82
TransUNet	86.0 ±0.53	76.4 ±0.70	81.1 ±1.37	72.2 ±1.25	91.5 ±0.39	85.0 ±0.62	84.6 ±0.61	77.9 ±0.62	74.2 ±1.73	66.4 ±1.72	78.6 ±0.54	69.6 ±0.36
RollingUNet	84.1 ±1.41	74.1 ±1.80	78.7 ±0.78	69.0 ±0.83	91.2 ±1.20	84.5 ±2.07	85.0 ±0.55	78.5 ±0.70	76.4 ±1.69	68.8 ±1.56	78.5 ±0.54	69.5 ±0.40
MedNeXt	86.8 ±0.54	77.6 ±0.65	83.2 ±0.78	74.7 ±0.65	92.1 ±0.37	86.2 ±0.41	88.9 ±0.51	83.2 ±0.51	75.4 ±1.63	68.0 ±1.41	80.0 ±0.42	71.5 ±0.43
EMCAD	84.3 ±0.39	73.9 ±0.49	82.8 ±0.52	74.1 ±0.69	93.0 ±0.2	87.2 ±0.27	86.9 ±0.32	81.3 ±0.32	67.4 ±3.4	59.9 ±3.33	78.5 ±1.94	69.4 ±2.0
Medsegdiff-V2	86.7 ±0.59	77.5 ±0.89	80.9 ±0.98	71.9 ±1.23	91.1 ±0.69	84.5 ±0.91	85.2 ±0.26	78.5 ±0.50	77.7 ±1.87	70.4 ±1.81	78.9 ±2.0	70.2 ±1.9
cDAL	78.3 ±0.62	65.8 ±0.87	80.0 ±0.71	70.5 ±0.75	67.4 ±0.31	62.1 ±0.28	85.7 ±0.78	79.1 ±0.89	74.6 ±0.79	66.4 ±0.45	76.9 ±0.98	68.0 ±0.91
SDSeg	85.1 ±0.5	75.2 ±0.57	81.2 ±0.8	72.4 ±0.7	91.3 ±0.41	84.9 ±0.45	86.6 ±0.66	80.1 ±0.75	76.6 ±1.11	69.2 ±1.07	78.7 ±1.86	70.1 ±1.66
UniSegDiff	**87.1** ±0.59	**78.1** ±0.79	**84.5** ±0.71	**76.0** ±0.58	**93.0** ±0.26	**87.3** ±0.44	**89.0** ±1.23	**82.9** ±1.22	**79.9** ±1.25	**72.4** ±1.39	**81.9** ±3.1	**73.1** ±3.59
Unified Lesion Segmentation Task												
UNet	84.5 ±0.73	74.2 ±0.87	80.0 ±0.85	71.2 ±0.87	90.2 ±0.57	83.1 ±0.78	79.6 ±0.23	71.1 ±0.25	64.5 ±1.46	56.0 ±1.44	71.1 ±1.2	61.3 ±1.23
TransUNet	82.7 ±0.3	71.6 ±0.37	76.8 ±1.78	67.6 ±1.94	90.3 ±0.82	83.1 ±0.89	79.7 ±1.56	71.2 ±1.61	64.3 ±1.5	56.1 ±1.59	75.1 ±1.19	65.7 ±1.3
RollingUNet	83.4 ±1.43	72.8 ±1.79	78.4 ±1.48	68.8 ±1.65	89.9 ±0.96	82.6 ±1.3	79.9 ±0.37	71.3 ±0.3	56.2 ±1.52	47.6 ±1.59	72.8 ±0.76	62.3 ±0.71
MedNeXt	84.2 ±0.33	73.9 ±0.33	80.6 ±0.78	71.7 ±0.73	90.6 ±0.37	83.6 ±0.52	84.0 ±1.0	76.7 ±1.42	63.8 ±1.33	55.7 ±1.48	76.7 ±2.1	67.5 ±1.91
EMCAD	83.5 ±0.64	72.7 ±0.87	82.9 ±0.66	74.2 ±0.72	91.6 ±0.27	85.9 ±0.47	86.9 ±0.68	81.3 ±0.89	65.2 ±3.1	57.7 ±2.9	77.5 ±0.84	68.6 ±0.86
Medsegdiff-V2	83.5 ±0.71	72.9 ±0.97	78.0 ±1.55	68.7 ±1.69	91.2 ±0.6	84.6 ±0.76	78.1 ±1.4	69.9 ±1.37	57.7 ±1.53	49.2 ±1.52	73.8 ±0.66	64.7 ±0.67
cDAL	76.7 ±0.63	64.0 ±0.69	73.8 ±0.83	64.4 ±0.99	67.3 ±1.26	62.1 ±1.23	82.8 ±0.94	75.6 ±0.97	65.9 ±1.38	57.6 ±1.36	75.5 ±1.84	66.6 ±1.72
SDSeg	67.2 ±0.95	55.6 ±0.97	76.6 ±0.89	66.9 ±0.87	90.9 ±1.41	84.2 ±1.45	86.6 ±0.72	79.9 ±0.75	62.1 ±1.34	54.7 ±1.31	78.3 ±1.78	69.7 ±1.82
UniSegDiff	**86.8** ±0.5	**77.6** ±0.72	**83.3** ±0.29	**74.7** ±0.18	**92.0** ±0.28	**85.9** ±0.3	**88.4** ±0.71	**82.4** ±0.96	**79.4** ±1.25	**71.9** ±1.3	**79.5** ±1.32	**70.4** ±1.29

noise, likely due to the network's preference for distribution mapping. While fast, this method sacrifices accuracy. Our staged denoising strategy balances efficiency and accuracy: During training, dynamic prediction targets ensure high attention across all time steps, facilitating rapid convergence and fully leveraging the model's capabilities. During inference, the rapid segmentation and denoising refinement stages perform single-step sampling, achieving accurate segmentation in as few as eleven steps (multiple refinements for mask fusion yield optimal results). This approach is at least 10 times faster than DDIM and 100 times faster than DDPM.

Ablation Study on Threshold Selection The staged training and inference approach we propose is divided into three phases, with the threshold selection between phases being critical. Table 5 presents detailed ablation experiments. In

Table 4: Ablation experiments of the denoising methods.

Denoise Method	Target	Train Epoch	Infer Step	Unified Task mDice	Unified Task mIoU
Uniform	noise	1000	100	81.5	73.4
	mask	300	100	80.9	72.6
One-Step	noise	300	1	75.3	68.2
	mask	300	1	78.6	71.4
Staged	both	300	11	**84.4**	**76.3**

Table 5: Ablation experiments of the threshold selection

High threshold	Low threshold	Unified Task mDice	Unified Task mIoU
700		82.1	74.0
600		83.3	75.6
500		82.7	74.7
600	400	83.9	76.8
600	300	**84.4**	**76.3**
600	200	84.4	76.3

Table 6: An ablation experiments of each component.

Staged	Pre-Tra	DCA	Fusion	Unified Task mDice
				77.0
✓				80.5
✓	✓			83.8
✓	✓	✓		84.4
✓	✓	✓	✓	**85.3**

Table 7: Comparison of training time, inference speed and inference Steps.

Methods	Training Time (hours)	Inference Speed (samples/s)	inference Steps	Unified Task mDice
Medsegdiff	≈ 172	0.24	100	77.1
SDSeg	≈ 43	**13.3**	1	77.0
cDAL	≈ 110	1.18	60	73.7
UniSegDiff	$\approx$ **25**	8.95	11	**84.4**

the experiments, the high threshold was first set to $t = 600$, and then, with the high threshold fixed, different low threshold values were tested. Ultimately, it was found that the optimal low threshold is $t = 300$.

Effectiveness of each component Table 6 presents the ablation experiments for each component proposed in UniSegDiff. The baseline uses uniform sampling to predict x_0 during training. Clearly, while the staged training approach improves segmentation performance, pre-training CFENet for segmentation significantly enhances the model's accuracy in unified lesion segmentation. The DCA module further facilitates feature fusion between CFENet and DNet, while uncertainty fusion leverages the randomness of the diffusion model to further enhance the accuracy and robustness of the segmentation results.

Comparison of time efficiency Table 7 presents the efficiency evaluation results for MedSegDiff-V2, SDSeg, cDAL, and UniSegDiff in the unified lesion segmentation task. To ensure a fair comparison, all models were trained on the same server. The results show that UniSegDiff significantly reduced training time and is much faster during inference compared to MedSegDiff-V2 and cDAL. Although it is slower than SDSeg, which samples only once, UniSegDiff achieves segmentation accuracy far superior to other diffusion-based models.

4 Conclusion and Future Work

In this paper, we investigate the characteristics of applying diffusion models to segmentation tasks. Through analysis, we propose a staged diffusion framework for unified lesion segmentation tasks, which includes tailored training strategies, inference methods, and model architecture. To enhance alignment across different types of lesion data, we pre-train the conditional feature extraction network as a segmentation model, significantly improving both inference speed and segmentation accuracy. Our method achieves state-of-the-art performance across multiple lesion segmentation benchmarks. Future work will focus on expanding our dataset to cover more lesion types and extending our approach into a unified framework that supports both 2D and 3D lesion data, with the goal of achieving comprehensive segmentation for all lesion types.

Acknowledgments. This work was supported by Dalian Science and Technology Innovation Foundation under Grant 2023JJ12GX015, and by the National Natural Science Foundation of China under Grant 62276046 and 62431004.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article

References

1. Covid-19 ct lung and infection segmentation dataset. https://medicalsegmentation.com/covid19/ (2020), accessed: June 25, 2025
2. Al-Dhabyani, W., Gomaa, M., Khaled, H., Fahmy, A.: Dataset of breast ultrasound images. Data in brief **28**, 104863 (2020)
3. Amit, T., Shaharbany, T., Nachmani, E., Wolf, L.: Segdiff: Image segmentation with diffusion probabilistic models. arXiv preprint arXiv:2112.00390 (2021)
4. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y.: Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)
5. Cheng, J., Huang, W., Cao, S., Yang, R., Yang, W., Yun, Z., Wang, Z., Feng, Q.: Enhanced performance of brain tumor classification via tumor region augmentation and partition. PloS one **10**(10), e0140381 (2015)
6. Cheng, J., Yang, W., Huang, M., Huang, W., Jiang, J., Zhou, Y., Yang, R., Zhao, J., Feng, Y., Feng, Q., et al.: Retrieval of brain tumors by adaptive spatial pooling and fisher vector representation. PloS one **11**(6), e0157112 (2016)
7. Fan, D.P., Ji, G.P., Zhou, T., Chen, G., Fu, H., Shen, J., Shao, L.: Pranet: Parallel reverse attention network for polyp segmentation. In: MICCAI. pp. 263–273. Springer (2020)
8. Hejrati, B., Banerjee, S., Glide-Hurst, C., Dong, M.: Conditional diffusion model with spatial attention and latent embedding for medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 202–212. Springer (2024)
9. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)

10. Hu, Y., Gao, Y., Gao, W., Luo, W., Yang, Z., Xiong, F., Chen, Z., Lin, Y., Xia, X., Yin, X., et al.: Amd-sd: An optical coherence tomography image dataset for wet amd lesions segmentation. Scientific Data **11**(1), 1014 (2024)
11. Jha, D., Smedsrud, P.H., Riegler, M.A., Halvorsen, P., De Lange, T., Johansen, D., Johansen, H.D.: Kvasir-seg: A segmented polyp dataset. In: MultiMedia modeling: 26th international conference, MMM 2020, Daejeon, South Korea, January 5–8, 2020, proceedings, part II 26. pp. 451–462. Springer (2020)
12. Jun, M., Cheng, G., Yixin, W., Xingle, A., Jiantao, G., Ziqi, Y., Minqing, Z., Xin, L., Xueyuan, D., Shucheng, C., et al.: Covid-19 ct lung and infection segmentation dataset. (No Title) (2020)
13. Kim, R.Y., Oke, J.L., Pickup, L.C., Munden, R.F., Dotson, T.L., Bellinger, C.R., Cohen, A., Simoff, M.J., Massion, P.P., Filippini, C., et al.: Artificial intelligence tool for assessment of indeterminate pulmonary nodules detected with ct. Radiology **304**(3), 683–691 (2022)
14. Lin, T., Chen, Z., Yan, Z., Yu, W., Zheng, F.: Stable diffusion segmentation for biomedical images with single-step reverse process. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 656–666. Springer (2024)
15. Liu, Y., Zhu, H., Liu, M., Yu, H., Chen, Z., Gao, J.: 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, pp. 3819–3827 (2024)
16. Parmar, G., Park, T., Narasimhan, S., Zhu, J.Y.: One-step image translation with text-to-image models. arXiv preprint arXiv:2403.12036 (2024)
17. Rahman, A., Valanarasu, J.M.J., Hacihaliloglu, I., Patel, V.M.: Ambiguous medical image segmentation using diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11536–11546 (2023)
18. Rahman, M.M., Munir, M., Marculescu, R.: Emcad: Efficient multi-scale convolutional attention decoding for medical image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11769–11779 (2024)
19. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. pp. 234–241. Springer (2015)
20. Roy, S., Koehler, G., Ulrich, C., Baumgartner, M., Petersen, J., Isensee, F., Jaeger, P.F., Maier-Hein, K.H.: Mednext: transformer-driven scaling of convnets for medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 405–415. Springer (2023)
21. Ruan, J., Xie, M., Gao, J., Liu, T., Fu, Y.: Ege-unet: an efficient group enhanced unet for skin lesion segmentation. In: International conference on medical image computing and computer-assisted intervention. pp. 481–490. Springer (2023)
22. Shi, L., Li, X., Hu, W., Chen, H., Chen, J., Fan, Z., Gao, M., Jing, Y., Lu, G., Ma, D., et al.: Ebhi-seg: A novel enteroscope biopsy histopathological hematoxylin and eosin image dataset for image segmentation tasks. Frontiers in Medicine **10**, 1114673 (2023)
23. Silva, J., Histace, A., Romain, O., Dray, X., Granado, B.: Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer. International journal of computer assisted radiology and surgery **9**, 283–293 (2014)
24. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)

25. Tajbakhsh, N., Gurudu, S.R., Liang, J.: Automated polyp detection in colonoscopy videos using shape and context information. IEEE TMI **35**(2), 630–644 (2015)
26. Vázquez, D., Bernal, J., Sánchez, F.J., Fernández-Esparrach, G., López, A.M., Romero, A., Drozdzal, M., Courville, A.: A benchmark for endoluminal scene segmentation of colonoscopy images. Journal of healthcare engineering **2017**(1), 4037190 (2017)
27. Warfield, S.K., Zou, K.H., Wells, W.M.: Simultaneous truth and performance level estimation (staple): an algorithm for the validation of image segmentation. IEEE transactions on medical imaging **23**(7), 903–921 (2004)
28. Wu, J., Ji, W., Fu, H., Xu, M., Jin, Y., Xu, Y.: Medsegdiff-v2: Diffusion-based medical image segmentation with transformer. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 6030–6038 (2024)
29. Zhao, X., Pang, Y., Ji, W., Sheng, B., Zuo, J., Zhang, L., Lu, H.: Spider: A unified framework for context-dependent concept understanding. arXiv preprint arXiv:2405.01002 (2024)
30. Zhu, L., Xue, Z., Jin, Z., Liu, X., He, J., Liu, Z., Yu, L.: Make-a-volume: Leveraging latent diffusion models for cross-modality 3d brain mri synthesis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 592–601. Springer (2023)
31. Zhu, Z., Xia, Y., Xie, L., Fishman, E.K., Yuille, A.L.: Multi-scale coarse-to-fine segmentation for screening pancreatic ductal adenocarcinoma. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part VI 22. pp. 3–12. Springer (2019)