Supplementary Materials for A Single Pixel is All You Need: Weakly Supervised Medical Image Segmentation using Discrete Denoising Diffusion Models

Mehmet Demirel
KIOS CoE, University of Cyprus
1 Panepistimiou Avenue, Nicosia, Cyprus
demirel.mehmet@ucy.ac.cy

Christos Kyrkou
KIOS CoE, University of Cyprus
1 Panepistimiou Avenue, Nicosia, Cyprus
kyrkou.christos@ucy.ac.cy

1s. Datasets

To evaluate our proposed model, we conduct experiments on three public datasets for medical image segmentation. These datasets cover different imaging modalities and anatomical structures, allowing for a thorough assessment of our method's performance and generalizability.

ISIC 2017 Skin Lesion Dataset: The ISIC 2017 dataset [4] is a large-scale benchmark for the analysis of skin lesions from dermoscopic images. It is designed for the task of binary skin lesion segmentation, which is a critical step towards automated melanoma detection. The dataset is officially divided into a training set of 2,000 images, a validation set of 150 images, and a testing set of 600 images.

ACDC Cardiac Dataset: The Automated Cardiac Diagnosis Challenge (ACDC) dataset [1] contains MRI scans for cardiac segmentation. The dataset includes exams from 100 patients, with a series of short-axis slices covering the heart from base to apex. The task is to segment three key cardiac structures: the Right Ventricle (RV), the Left Ventricle (LV), and the Myocardium (Myo). Following the common experimental setup [2], we split the dataset into 70 cases (1,930 axial slices) for training, 10 cases for validation, and 20 cases for testing.

Synapse Multi-organ Dataset: The Synapse dataset [2] is a benchmark for abdominal organ segmentation. It consists of 30 abdominal CT scans. Each CT volume contains between 85 and 198 slices. The dataset is annotated for 8 abdominal organs: Aorta (AT), Gallbladder (GB), Spleen (SP), Left Kidney (LK), Right Kidney (RK), Liver (LR), Pancreas (PC), and Stomach (SM). Following the standard protocol established by [2], we use 18 scans (2,212 axial slices) for training and the remaining 12 scans for testing.

2s. Implementation Details

Our framework is built upon a 4-level conditional U-Net that serves as the denoising network, conditioned on both sinusoidal timestep embeddings and contextual features extracted from a pretrained VGG16 backbone [11]. We set the total number of diffusion timesteps to T=200, with a linear noise schedule where β_t increases from 10^{-4} to 0.02. The model is trained for 50 epochs using the AdamW optimizer with a fixed learning rate of 10^{-4} and a batch size of 8. To address the extreme class imbalance, we employ a heavily-weighted Focal Loss objective with a focusing parameter $\gamma = 2.0$ and a foreground balancing weight of $\alpha = 0.999$. For supervision, the ground-truth map x_0 is constructed using a single, randomly (discrete uniform distribution) sampled pixel annotation per foreground class from the full segmentation masks. All conditioning images are resized to 224×224 pixels, and we apply random horizontal flipping for data augmentation. During inference, we generate $N = 100$ point cloud samples for each input image by starting with different random initial noise maps. The entire framework is implemented in PyTorch and trained on a single NVIDIA A100 GPU.

Table 1s. Ablation study results on the ISIC 2017 dataset, comparing our D3PM approach against a standard U-Net. Both models were trained under the same conditions, using a single positive pixel and a heavily-weighted Focal Loss.

	Dice Score (%)
U-Net$_{\text{FL}}$	6.81
Proposed (Ours)	75.37

3s. Ablation on the Generative Framework

To demonstrate the critical role of the Discrete Denoising Diffusion Model (D3PM) framework in our proposed approach, we conducted a crucial ablation study. The goal of this experiment was to isolate the contribution of our generative modeling paradigm from that of the heavily-weighted Focal Loss. To achieve this, we replaced our D3PM with a standard U-Net architecture while keeping the training conditions identical. This involved providing supervision from a single positive pixel per object and using the same heavily-weighted Focal Loss objective. The results are pre-

sented in Table 1s. The standard U-Net trained with Focal Loss (U-Net$_{\text{FL}}$) achieves a Dice score of only 6.81%. This extremely poor performance indicates an almost complete model collapse. Despite the heavy weighting on the foreground class, the deterministic nature of the U-Net struggled to propagate information from a single supervisory point. The model was compelled to learn a trivial solution where it predicted the background class for nearly every pixel because the training objective was dominated by the overwhelming number of background pixels. In comparison, our proposed D3PM-based method achieves a Dice score of 75.37%. This significant difference in performance highlights that the success of our method is not solely attributable to the choice of loss function. Instead, it is the fundamental reframing of segmentation as a conditional generative task that prevents collapse. The stochastic nature of the diffusion process, combined with the Focal Loss, encourages the model to learn a distribution of plausible foreground locations, manifesting as an emergent point cloud. This learned geometric prior is far more robust than the direct, deterministic prediction attempted by the standard U-Net. This experiment demonstrates that the generative and stochastic properties of our D3PM framework are essential for learning meaningful representations from such extremely sparse supervision.

4s. Ablation on the Number of Inference Samples (N)

Our proposed inference strategy (Sec 3.3 in the main paper) relies on aggregating N stochastically generated point cloud samples to form the final dense segmentation mask. The choice of N presents a trade-off between segmentation accuracy and computational cost, as a higher N requires more forward passes of the learned reverse diffusion process, leading to longer inference times.

The results, presented in Figure 1s shows that for very low values of N, the resulting segmentation quality is poor. This is because the aggregation of only a few samples results in poor spatial coverage of the target object. Consequently, the final mask generated via the majority vote is fragmented and incomplete. As the number of aggregated samples increases, the Dice score rises steeply. This rapid improvement demonstrates the effectiveness of our aggregation strategy, where the consensus from multiple samples successfully fills in these gaps and converges toward a more complete and accurate segmentation.

However, the plot also reveals a distinct point of diminishing returns. Beyond a certain number of samples, the performance curve begins to plateau, with further increases in N yielding only marginal gains in segmentation accuracy. This observation indicates that a sufficient number of samples have been aggregated to form a consensus mask. Given this trade-off, we selected a value for $N = 100$, as it achieves near-maximal performance without incurring the unnecessary computational overhead of aggregating a much larger set of samples. This choice ensures an effective balance between accuracy and inference efficiency for all our experiments.

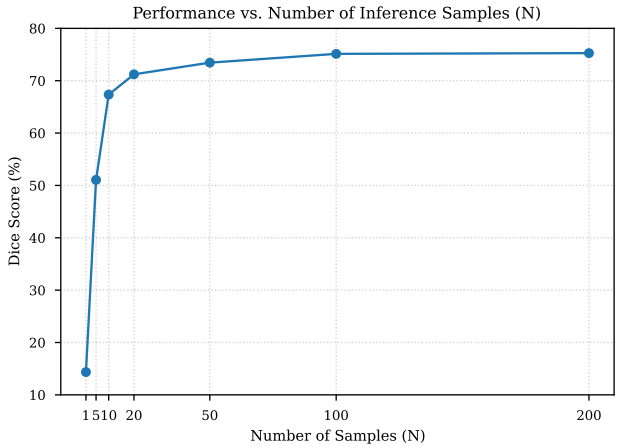

Figure 1s. Performance against the number of inference samples (N) on the ISIC 2017 dataset.

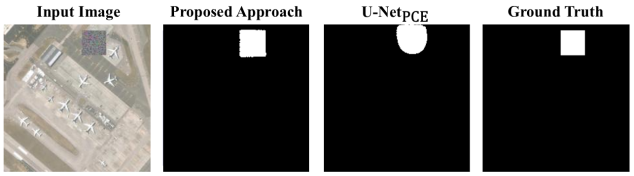

Figure 2s. Qualitative comparison for adversarial patch detection on the AID dataset.

5s. Application to Other Domains

To demonstrate the generalizability and robustness of our proposed framework beyond the scope of medical image segmentation, we evaluate its performance on an entirely different task. For this we used adversarial patch detection. This supplementary experiment is included to illustrate the broad applicability of our single-pixel supervision strategy in other visual domains.

In this application, the objective is to densely segment adversarial patches placed within images. Similar to medical lesions or organs, defining the exact dense boundaries of these patches is a time-consuming annotation process. Furthermore, selecting representative "background" (clean) pixels can be ambiguous depending on the complexity of the underlying scene. We tackle this by applying our proposed generative framework, providing only a single positive point annotation (1 click) per adversarial patch.

We compare our proposed approach against the standard U-Net$_{PCE}$ baseline. As established in our main methodol-

Table 2s. Full quantitative results on the ISIC 2017 dataset, corresponding to Table 1 in the main paper. We report Dice Score (%) as mean ± std.

Methods	Dice Score (%)
Full Supervision	$80.73_{\pm 0.14}$
U-Net$_{PCE}$ [7]	$71.79_{\pm 0.72}$
AIL [3]	$68.60_{\pm 0.41}$
S2L [5]	$69.65_{\pm 0.71}$
ScribbleVC [6]	$69.87_{\pm 1.51}$
USTM [8]	$70.60_{\pm 0.52}$
U-Net$_{PCE+CRF}$ [13]	$72.26_{\pm 1.54}$
Proposed Method	$75.03_{\pm 0.54}$

Table 3s. Full quantitative results on the ACDC dataset, corresponding to Table 2 in the main paper. We report Dice Score (%) as mean ± std.

Method	RV	Myo	LV	Average Dice (%)
Full Supervision	$89.53_{\pm 0.48}$	$87.14_{\pm 0.32}$	$93.36_{\pm 0.67}$	$90.01_{\pm 0.49}$
U-Net$_{PCE}$ [7]	$62.78_{\pm 3.23}$	$70.56_{\pm 1.61}$	$89.60_{\pm 0.90}$	$74.31_{\pm 1.88}$
AIL [3]	$33.81_{\pm 4.31}$	$32.03_{\pm 2.93}$	$42.05_{\pm 4.15}$	$35.96_{\pm 0.99}$
ScribbleVC [6]	$46.42_{\pm 2.70}$	$44.26_{\pm 5.97}$	$62.82_{\pm 6.23}$	$51.16_{\pm 4.96}$
U-Net$_{PCE+CRF}$ [13]	$49.02_{\pm 0.46}$	$67.12_{\pm 9.27}$	$88.48_{\pm 2.63}$	$68.20_{\pm 3.90}$
S2L [5]	$63.35_{\pm 1.52}$	$70.27_{\pm 2.41}$	$88.52_{\pm 1.49}$	$74.05_{\pm 0.87}$
USTM [8]	$62.51_{\pm 4.15}$	$72.61_{\pm 2.92}$	$89.20_{\pm 1.98}$	$74.77_{\pm 3.00}$
Proposed Method	$70.27_{\pm 1.08}$	$72.64_{\pm 0.77}$	$87.74_{\pm 0.54}$	$76.88_{\pm 0.73}$

Table 4s. Full quantitative results on the Synapse Multi-organ dataset, corresponding to Table 3 in the main paper. We report Dice Score (%) as mean ± std.

Method	SP	RK	LK	GB	PC	LR	SM	AT	Average Dice (%)
Full Supervision	$84.42_{\pm 0.41}$	$85.08_{\pm 0.36}$	$80.04_{\pm 0.54}$	$48.71_{\pm 1.10}$	$50.50_{\pm 2.59}$	$86.39_{\pm 0.45}$	$69.18_{\pm 0.57}$	$86.19_{\pm 1.19}$	$73.81_{\pm 0.21}$
U-Net$_{PCE}$ [7]	$57.67_{\pm 6.24}$	$54.53_{\pm 2.64}$	$51.97_{\pm 2.05}$	$29.61_{\pm 6.52}$	$28.98_{\pm 1.53}$	$68.64_{\pm 1.70}$	$46.42_{\pm 4.45}$	$42.76_{\pm 4.15}$	$47.57_{\pm 1.31}$
AIL [3]	$41.08_{\pm 1.41}$	$34.11_{\pm 3.10}$	$27.05_{\pm 0.86}$	$17.91_{\pm 3.81}$	$24.65_{\pm 0.87}$	$60.67_{\pm 0.86}$	$39.43_{\pm 2.40}$	$16.39_{\pm 4.21}$	$32.66_{\pm 1.99}$
ScribbleVC [6]	$54.82_{\pm 5.83}$	$59.51_{\pm 0.65}$	$52.53_{\pm 4.68}$	$31.22_{\pm 0.90}$	$29.90_{\pm 1.78}$	$67.43_{\pm 2.27}$	$42.59_{\pm 0.58}$	$36.47_{\pm 5.87}$	$46.81_{\pm 0.28}$
S2L [5]	$57.87_{\pm 1.32}$	$60.49_{\pm 2.93}$	$51.36_{\pm 6.67}$	$23.69_{\pm 5.37}$	$26.79_{\pm 0.33}$	$69.39_{\pm 0.44}$	$45.54_{\pm 0.58}$	$39.85_{\pm 1.15}$	$46.87_{\pm 0.16}$
USTM [8]	$60.09_{\pm 1.05}$	$59.98_{\pm 4.88}$	$51.75_{\pm 5.03}$	$33.66_{\pm 0.91}$	$27.24_{\pm 0.16}$	$68.75_{\pm 1.25}$	$46.44_{\pm 1.06}$	$37.92_{\pm 1.11}$	$48.23_{\pm 0.18}$
U-Net$_{PCE+CRF}$ [13]	$71.43_{\pm 0.46}$	$52.06_{\pm 2.09}$	$56.63_{\pm 0.60}$	$20.86_{\pm 9.40}$	$29.06_{\pm 6.92}$	$73.64_{\pm 1.54}$	$45.39_{\pm 2.93}$	$43.12_{\pm 3.34}$	$49.03_{\pm 3.32}$
Proposed Method	$65.50_{\pm 1.31}$	$70.82_{\pm 1.09}$	$70.63_{\pm 1.55}$	$31.17_{\pm 0.82}$	$29.47_{\pm 0.46}$	$74.10_{\pm 1.72}$	$46.71_{\pm 0.88}$	$60.77_{\pm 1.62}$	$56.15_{\pm 1.03}$

Table 5s. Quantitative results on adversarial patch detection.

Method	Supervised Method	User Clicks	Dice (%)	IoU (%)
U-Net$_{PCE}$	1 Pos + 1 Neg	2	78.51	64.35
Proposed	1 Pos	1	91.86	84.95

ogy, the U-Net$_{PCE}$ requires at least one positive and one negative point (2 user clicks) to prevent total model collapse. Specifically, for evaluation, we utilized the AID dataset [14] and simulated patch attacks using randomly placed Gaussian patches.

The quantitative results are presented in Table 5s. Our proposed method achieves an impressive Dice score of 91.86% and an Intersection over Union (IoU) [9] of 84.95%. This marks a substantial improvement of 13.35% in Dice and 20.60% in IoU over the U-Net$_{PCE}$ baseline, despite our method requiring half the annotation effort. Furthermore, the baseline reaches only a 78.51% Dice score even with access to both foreground and background supervisory signals. Qualitative results for the adversarial patch detection task are presented in Figure 2s, illustrating our method's superior ability to capture sharp geometric boundaries compared to the baseline approach.

These results confirm that the generative prior learned by our heavily-weighted Focal Loss and stochastic sample aggregation is not restricted to medical modalities. Our method successfully captures the distinct geometry and distribution of adversarial patches from an extremely localized signal, outperforming traditional partial cross-entropy methods while significantly reducing the human-in-the-loop requirement for robust defense.

6s. Limitations & Future Works

The primary limitation of our method is its inference speed, which stems from the iterative nature of the reverse diffusion process and our strategy of generating N samples per image for aggregation. This was a deliberate trade-off, prioritizing maximal accuracy under extreme weak supervision over inference latency. However, this presents a clear avenue for future work. The inference time can be substantially reduced by incorporating faster sampling techniques, such as Denoising Diffusion Implicit Models (DDIMs) [12], or by adapting our framework to operate in a compressed latent space via Latent Diffusion Models (LDMs) [10]. Investigating these acceleration strategies is a promising direction to enhance the clinical applicability of

our approach without sacrificing the generative power that underpins its accuracy.

References

[1] Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester, Gerard Sanroma, Sandy Napel, Steffen Petersen, Georgios Tziritas, Elias Grinias, Mahendra Khened, Varghese Alex Kollerathu, Ganapathy Krishnamurthi, Marc-Michel Rohé, Xavier Pennec, Maxime Sermesant, Fabian Isensee, Paul Jäger, Klaus H. Maier-Hein, Peter M. Full, Ivo Wolf, Sandy Engelhardt, Christian F. Baumgartner, Lisa M. Koch, Jelmer M. Wolterink, Ivana Išgum, Yeonggul Jang, Yoonmi Hong, Jay Patravali, Shubham Jain, Olivier Humbert, and Pierre-Marc Jodoin. Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: Is the problem solved? *IEEE Transactions on Medical Imaging*, 37(11):2514–2525, 2018. 1

[2] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021. 1

[3] Jingkun Chen, Wenjian Huang, Jianguo Zhang, Kurt Debattista, and Jungong Han. Addressing inconsistent labeling with cross image matching for scribble-based medical image segmentation. *IEEE Transactions on Image Processing*, 2025. 3

[4] Noel CF Codella, David Gutman, M Emre Celebi, Brian Helba, Michael A Marchetti, Stephen W Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, et al. 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. IEEE, 2018. 1

[5] Hyeonsoo Lee and Won-Ki Jeong. Scribble2label: Scribble-supervised cell segmentation via self-generating pseudo-labels with consistency. In *International conference on medical image computing and computer-assisted intervention*, pages 14–23. Springer, 2020. 3

[6] Zihan Li, Yuan Zheng, Xiangde Luo, Dandan Shan, and Qingqi Hong. Scribblevc: Scribble-supervised medical image segmentation with vision-class embedding. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 3384–3393, 2023. 3

[7] Di Lin, Jifeng Dai, Jiaya Jia, Kaiming He, and Jian Sun. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3159–3167, 2016. 3

[8] Xiaoming Liu, Quan Yuan, Yaozong Gao, Kelei He, Shuo Wang, Xiao Tang, Jinshan Tang, and Dinggang Shen. Weakly supervised segmentation of covid19 infection with scribble annotation on ct images. *Pattern Recognition*, 122: 108341, 2022. 3

[9] Shervin Minaee, Yuri Boykov, Fatih Porikli, Antonio Plaza, Nasser Kehtarnavaz, and Demetri Terzopoulos. Image segmentation using deep learning: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(7):3523–3542, 2021. 3

[10] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 3

[11] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1

[12] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. 3

[13] Meng Tang, Federico Perazzi, Abdelaziz Djelouah, Ismail Ben Ayed, Christopher Schroers, and Yuri Boykov. On regularized losses for weakly-supervised cnn segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 507–522, 2018. 3

[14] Gui-Song Xia, Jingwen Hu, Fan Hu, Baoguang Shi, Xiang Bai, Yanfei Zhong, Liangpei Zhang, and Xiaoqiang Lu. Aid: A benchmark data set for performance evaluation of aerial scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(7):3965–3981, 2017. 3