Diffusion-based Pre-training for Label-efficient Abdominal Multi-organ Segmentation

Yongzhi Huang
School of Artificial Intelligence,
Beijing University of Posts and Telecommunications
Beijing, China
yhuang@bupt.edu.cn

Jinxin Zhu
School of Artificial Intelligence,
Shenzhen Technology University
Shenzhen, China
ronkoc@outlook.com

Haseeb Hassan
School of Artificial Intelligence,
Shenzhen Technology University
Shenzhen, China
haseeb@sztu.edu.cn

Jingyu Li
School of Artificial Intelligence,
Shenzhen Technology University
Shenzhen, China
lijingyu@sztu.edu.cn

Liyilei Su
School of Artificial Intelligence,
Shenzhen Technology University
Shenzhen, China
suliyilei@sztu.edu.cn

Bingding Huang
School of Artificial Intelligence,
Shenzhen Technology University
Shenzhen, China
huangbingding@sztu.edu.cn

2025 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) | 979-8-3315-1557-7/25/$31.00 ©2025 IEEE | DOI: 10.1109/BIBM66473.2025.11356434

***Abstract*—Accurate multi-organ segmentation in Computed Tomography (CT) images is critical for computer-aided diagnosis systems. However, existing supervised methods heavily rely on costly, high-quality labeled data. To address this, we propose a label-efficient segmentation method for abdominal organs in CT images, leveraging knowledge transfer from a pre-trained diffusion model. Specifically, we pre-train a denoising diffusion model on 207,029 unlabeled 2D CT slices to capture anatomical patterns, which is then fine-tuned on limited labeled data for abdominal organ segmentation. During fine-tuning, two strategies—linear probing and decoder fine-tuning—are employed to adapt the model for segmentation while preserving learned representations. Quantitative results demonstrate that the pre-trained diffusion model can generate diverse and realistic 256x256 CT images (FID: 11.32, sFID: 46.93, F1-score: 73.1%). Moreover, our method achieves competitive performance on the FLARE 2022 dataset for organ segmentation, particularly excelling in limited labeled data scenarios. With only 10% and 1% labeled data, our method achieves DSCs of 78.51% and 71.56% on 13 abdominal organs, respectively. Remarkably, with only four labeled 2D slices, our method still achieves a DSC of 51.81%, highlighting the efficacy of our method in alleviating the reliance of supervised learning on large-scale labeled data.**

***Index Terms*—Medical imaging processing, Abdominal organ segmentation, Label-efficient learning, Diffusion models, Pre-trained models.**

I. Introduction

Medical image segmentation is critical for computer-aided diagnosis systems, such as accurate diagnosis and treatment planning [1]. Although supervised deep learning methods have significantly advanced this field [2], they rely heavily on large-scale, high-quality annotations, which are costly to obtain in clinical practice [3]. The scarcity and inconsistency of annotations, along with cross-domain and noisy-label challenges arising from multi-center, multi-site data acquisition and heterogeneous scanners, motivate label-efficient learning methods that aim to achieve accurate segmentation from limited annotated data [4].

Self-supervised learning provides a promising solution by exploiting abundant unlabeled data. In addition to contrastive [5] and masked reconstruction objectives [6], recent advancements in generative pre-training, through models such as Generative Adversarial Networks (GANs) and Denoising Diffusion Probabilistic Models (DDPMs), have shown strong potential for learning robust and transferable representations and synthesizing diverse domain-specific patterns [7], [8]. However, despite recent progress in diffusion-based pre-training [9]–[13], its application to medical image segmentation, particularly in label-efficient abdominal organ segmentation, remains largely unexplored. This is mainly attributed to the domain gap between natural and medical images, which demands modality-specific pre-training from scratch, and the lack of effective transfer strategies to bridge the gap between diffusion-based generative models and semantic segmentation tasks.

To address this gap, we propose a diffusion-based framework that pre-trains a DDPM on unlabeled CT data and transfers it to label-efficient multi-organ segmentation. Our main contributions are: (1) pre-training a DDPM from scratch on 207,029 unlabeled 2D CT slices, generating diverse and realistic 256×256 CT images, with a Fréchet Inception Distance (FID) of 11.32, spatial FID (sFID) of 46.93, and F1-score of 73.1%; (2) introducing lightweight, end-to-end fine-tuning strategies that adapt the pretrained DDPM for multi-organ segmentation tasks, achieving competitive performance on the MICCAI FLARE2022 benchmark; and (3) demonstrating superior performance in limited labeled data scenarios. With 10% and 1% labeled data, our method achieves DSCs of 78.51% and 71.56%, respectively, outperforming state-of-the-art methods. Remarkably, with only four labeled CT slices, the model still generates anatomically consistent segmentation masks, achieving a DSC of 51.81%.

This study was supported by Shenzhen Science and Technology Program (KJZD20240903095605007) and Shenzhen Medical Research Fund (D250402003). Corresponding authors: Bingding Huang and Liyilei Su.

II. Related work

A. Pre-training Generative Models for Semantic Segmentation

Generative models synthesize data by learning the underlying distributions of unlabeled samples. GANs [14] have dominated image synthesis for the past decade, enabling diverse computer vision applications. Recently, diffusion models have emerged as a powerful alternative, surpassing GANs in stability and quality [15]–[18]. The potential of generative pre-training for segmentation was first explored in GAN-based models [11]–[13], which exploit latent representations to generate pseudo labels and capture joint image–label distributions. This motivates exploring whether diffusion-based pre-training can similarly benefit segmentation. Recent studies confirmly demonstrated that diffusion pre-training transfers effectively to both natural [8]–[10] and medical images [19], [20], often outperforming GAN-based approaches.

Diffusion-based segmentation methods can be grouped into three paradigms. (1) *Representation projection:* using classifiers (SVMs or MLPs) to map pre-trained diffusion features to segmentation masks [8], [12], [13], though they require task-specific hyper-parameters tuning. (2) *Conditional generation:* integrating segmentation priors into the DDPM sampling chain [10], [19], but inference is slow due to iterative denoising. (3) *Self-supervised fine-tuning:* pre-training via DDPM followed by aligning with post-fine-tuning practice yet limited by the lack of modality-specific checkpoints for medical imaging [9].

B. Label-efficient Organ Segmentation in Medical Imaging

Deep supervised learning has revolutionized medical image segmentation, led by FCNs [21] and U-Net [22], inspiring variants [23]–[27] tailored to anatomical or modality-specific constraints. However, manual annotations remain costly and expertise-intensive [3], motivating label-efficient solutions [4].

DDPM-based pre-training contributes to label-efficient segmentation through two complementary aspects. First, the denoising process serves as generative pre-training, analogous to contrastive [5], [28] or masked reconstruction learning [6], where noise prediction implicitly encodes semantic structure [8]. Second, the generative capacity of DDPMs enables synthetic data augmentation for semi-supervised frameworks such as teacher–student [29], self-training [30], and pseudo-label refinement [31], facilitating efficient segmentation under limited annotations.

III. Methods

As shown in Fig. 1, our framework is divided into two stages: **(A)** the upstream DDPM pre-training and **(B)** the downstream multi-organ segmentation stage. In the pre-training stage, a U-Net is trained to predict the added noise from augmented images based on a denoising diffusion objective. In the segmentation stage, the same U-Net architecture with shared parameters is fine-tuned using annotated CT images for multi-organ segmentation tasks.

A. Pre-training Models with DDPM

1) Diffusion Model: DDPM consists of two key components: the forward diffusion process and the reverse sampling process. In the diffusion process, a Markov chain progressively adds Gaussian noise to input images $\mathbf{x}_0$. In contrast, the sampling process gradually reconstructs the images by denoising from an initial random Gaussian noise. DDPM aims to learn a data distribution $p_\theta(\mathbf{x}_0)$ that approximates the given data distribution $q(\mathbf{x}_t)$ by reversing the noising process.

The forward process is defined as:

$$q(\mathbf{x}_t \mid \mathbf{x}_{t-1}) := \mathcal{N}\left(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t \mathbf{I}\right) \tag{1}$$

For a fixed variance schedule $\beta_1, \ldots, \beta_t$, this Markov process enables direct sampling of noisy images $\mathbf{x}_t$ at arbitrary time steps directly from input images $\mathbf{x}_0$ using the following closed-form expression:

$$q(\mathbf{x}_t \mid \mathbf{x}_0) := \mathcal{N}\left(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I}\right) \tag{2}$$

where $\alpha_t := 1 - \beta_t, \bar{\alpha}_t := \prod_{s=1}^{t} \alpha_s$.

The reverse process starts from Gaussian noise $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and iteratively predicts noisy samples $\mathbf{x}_{t-1}$ using a learned model:

$$p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t) := \mathcal{N}\left(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t)\right) \tag{3}$$

$$\boldsymbol{\mu}_\theta(\mathbf{x}_t, t) = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)\right) \tag{4}$$

Here, $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)$ is a noise predictor network implemented using a U-Net architecture in this work.

2) Network Architecture: We use a modified U-Net for both diffusion and segmentation, termed as the predictor and segmentor U-Net. Two backbones are evaluated: U-Net-B and U-Net-L, with ResBlock widths of 128 and 256. For segmentation, the segmentor U-Net adds a lightweight head, yielding three variants: U-Net-B/h128, U-Net-B/h256, and U-Net-L/h256, where "/h" denotes the hidden width in the head.

B. Transferring to Multi-organ Segmentation Tasks

1) Transfer Strategy: As shown in Fig. 1B, the segmentor U-Net is initialized from the best-performing predictor U-Net checkpoints (250k for U-Net-B, 300k for U-Net-L). The diffusion timestep t is treated as a tunable hyper-parameter influencing ResBlock initialization. The output layer is replaced by a classification head with convolutional and normalization layers trained from scratch. The network is then fine-tuned on labeled segmentation tasks.

2) Fine-tuning strategies: We evaluate three fine-tuning strategies (Fig. 2): linear probing (LP), decoder fine-tuning (FT-dec), and training from scratch (FS). LP freezes the pretrained backbone and updates only a linear classifier. FT-dec unfreezes the decoder blocks for better adaptation while keeping the encoder fixed. FS trains the segmentor U-Net from random initialization, serving as a baseline to evaluate the benefits of DDPM pre-training.

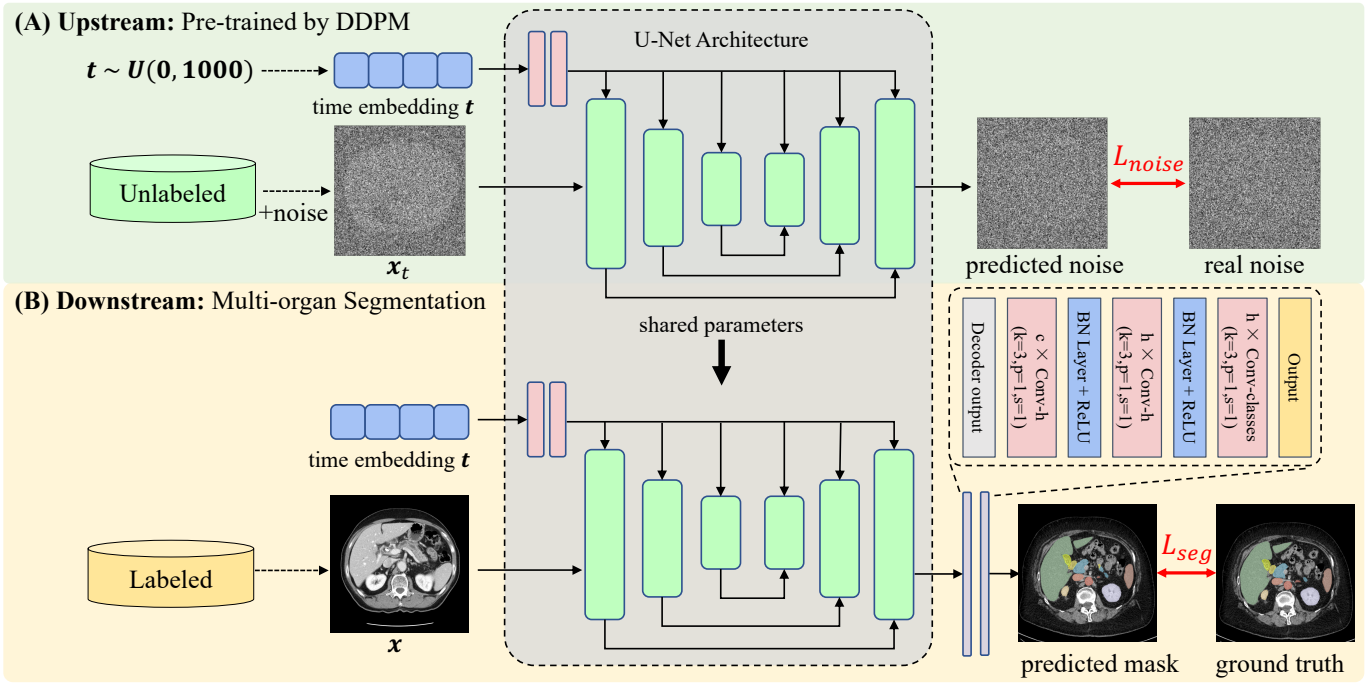

Fig. 1. The proposed framework includes **(A)** the upstream DDPM pre-training and **(B)** the downstream multi-organ segmentation.

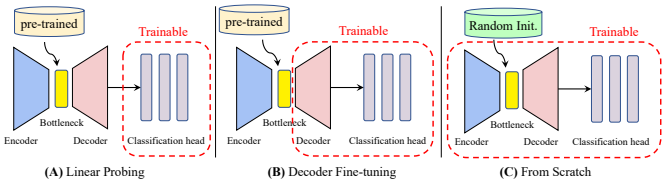

Fig. 2. Fine-tuning strategies for downstream multi-organ segmentation.

IV. Experiments and Results

A. Dataset and Implementation Details

Dataset. Experiments are conducted on the MICCAI FLARE22 dataset containing 2,000 unlabeled and 50 labeled abdominal CT scans across 13 organs.[1] The first 1,000 unlabeled scans (207,029 2D slices) are used for DDPM pre-training, while the 50 labeled scans are split 4:1 for fine-tuning (3,879 training and 915 test slices).

Preprocessing. All scans are resampled to a unified voxel spacing, with intensities clipped to the 0.5–99.5 percentile, normalized to $[-1, 1]$. Each 3D scan is decomposed into transverse 2D slices resized to 256×256 (bilinear for images, nearest-neighbor for labels). The horizontal flipping is applied in DDPM pre-training.

Experiment Settings. The model is implemented in PyTorch and trained on NVIDIA A100 GPUs. DDPM pre-training uses Gaussian noise with a cosine variance schedule over 1,000 steps (β_1=0.0001, β_T=0.02), optimized via Adam for 300k iterations with EMA (momentum 0.9999) and a decaying learning rate from 2×10^{-4} to 2×10^{-5}. For segmentation, we use a weighted sum of cross-entropy and smoothed Dice loss (w=0.5, ϵ=10^{-5}). All fine-tuning strategies use Adam with weight decay 10^{-3} for the head and 10^{-4} elsewhere; the head learning rate is $10\times$ higher than the backbone. Training runs for 30k iterations in the fully supervised and 10k in the label-efficient setting.

Evaluation Metrics. For generative tasks, the DDPM-generated CT images are evaluated by FID [32], sFID [33], precision, recall, and F1-score [34], computed against 2D slices from 2000 real unlabeled FLARE22 scans. For segmentation tasks, segmentation results are evaluated by Dice similarity coefficient (DSC) and Jaccard index (JI), computed against segmentation labels.

B. CT Image Synthesis Performance

We evaluate the quality of synthetic CT images generated from U-Net-B and U-Net-L models, both trained for 300k iterations and evaluated every 50k iterations. As shown in Table I, U-Net-B achieves better overall performance than U-Net-L in the generative task, suggesting that increasing model size does not necessarily improve generation quality. The smaller model converges earlier, reaching stable performance after 150k iterations and the best results at 250k iterations, while U-Net-L requires more training to reach stable performance with its best results at 300k iterations. Although some metrics are slightly lower, the overall quality remains competitive. Representative images generated by U-Net-B at 250k iterations are shown in Fig. 3, which exhibits realistic anatomic structures and plausible intensity distributions of abdominal organs.

[1]https://flare22.grand-challenge.org/

TABLE I
EVALUATION ON GENERATED CT IMAGES OF A RESOLUTION OF 256×256.

		50k	100k	15k	200k	250k	300k
U-Net-B	FID	31.4092	16.3188	12.0206	11.5218	**11.3162**	12.0449
	sFID	72.5776	56.0745	48.8117	47.5353	**46.9282**	47.4179
	Precision	0.489	0.6755	0.758	0.7995	0.796	**0.803**
	Recall	0.4035	0.5635	0.585	0.63	**0.676**	0.659
	F1-score	0.4422	0.6144	0.6604	0.7047	**0. 7311**	0.7239
U-Net-L	FID	38.8874	38.5144	24.4172	13.4527	13.479	**12.2408**
	sFID	83.7842	85.2128	63.9002	52.5795	53.1192	**50.3049**
	Precision	0.3765	0.434	0.5985	0.7135	**0.736**	0.724
	Recall	0.319	0.535	0.588	0.599	0.595	**0.642**
	F1-score	0.3454	0.4792	0.5932	0.6513	0.658	**0.6805**

TABLE II
QUANTITATIVE RESULTS WITH DIFFERENT FINE-TUNING STRATEGIES (LP: LINEAR PROBING, FT-DEC: DECODER FINE-TUNING, FS: FROM SCRATCH).

Strategy	Step	Dice Coef.(%)			Jaccard Index (%)		
		B/h128	B/h256	L/h256	B/h128	B/h256	L/h256
FS	-	80.59	79.41	83.83	74.72	73.78	**77.16**
LP	100	26.86	26.29	10.63	20.24	19.86	7.81
	200	28.85	20.41	13.84	22.19	13.9	10.78
	300	18.72	22.45	17.98	13.76	16.4	13.61
	400	28.23	23.72	17.98	22.34	17.93	13.61
FT-dec	0	**86.91**	**85.21**	79.76	**80.38**	**78.66**	74.17
	100	77.98	79.29	78.32	72.13	73.31	72.43
	200	77.15	77.29	78.71	71.22	71.21	72.71
	300	83.73	76.5	**83.84**	76.92	70.55	76.87
	400	76.94	81.28	78.86	70.54	74.04	72.8
	500	81.54	76.71	77.01	74.42	70.33	70.68
	600	80.25	73.87	77.8	72.65	67.3	71.78
	700	72.48	73.03	76.3	65.69	66.08	69.83
	800	68.95	68.98	74.44	61.79	61.9	67.86
	900	63.94	64.18	74.71	57.86	57.73	67.92
	1000	64.47	69.31	74.64	58.07	62.48	67.65

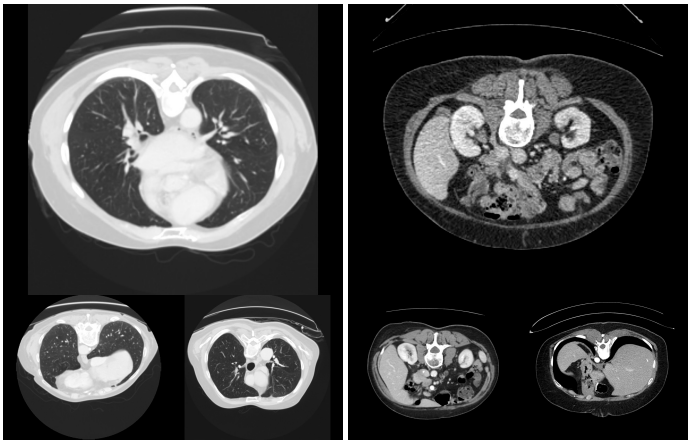

Fig. 3. Samples generated by DDPM in lung view(W:1400, L:-500, left) and abdominal view (W:350, L:40, right).

C. Ablation Experiments for Multi-organ Segmentation

We conducted ablation studies on the FLARE22 dataset to evaluate the effects of model scale, fine-tuning strategy, and initial diffusion step, as summarized in Table II. The FS strategy achieves stable segmentation with all variants exceeding 70% DSC, indicating that the shared U-Net backbone remains effective after removing diffusion-specific components. In contrast, the LP strategy performs poorly across all scales, with most configurations failing to converge, suggesting that representations learned from DDPM pre-training are not directly transferable without sufficient adaptation. In contrast, introducing the FT-dec strategy yields notable improvements, especially for U-Net-B/h128, which surpasses FS by 6.3% DSC and 5.5% JI. Among all settings, the best results are achieved when initializing at diffusion step 0, suggesting that early-stage diffusion features provide the most beneficial pre-training knowledge for abdominal organ segmentation.

D. Multi-organ Segmentation Performance on FLARE Dataset

1) Baseline Methods: We evaluate advanced baseline methods for multi-organ segmentation under two experimental settings: fully supervised and label-efficient.

In the fully supervised setting, we compare our method with DeepLabV3+ [35], U-Net [22] and its variants (ResU-Net [23], U-Net++ [24], Attention U-Net [25], UNETR [26], Swin UNETR [27]), as well as nnU-Net [36] and diffusion-based DDPM-Seg [8]. For fair comparison, we use the official or standard implementations: DeepLabV3+, ResU-Net, and U-Net++ are initialized with ImageNet pre-trained weights[2], while Attention U-Net, UNETR, and Swin UNETR are implemented in the MONAI framework.[3]

In the label-efficient setting, we evaluate performance using 10% (388 slices), 1% (39 slices), and 0.1% (4 slices) of labeled data. Subsets for 1% and 10% are randomly sampled, while 0.1% is manually curated to ensure coverage of all organ categories. DDPM-Seg is reproduced using the same settings as the original paper but evaluated only under 0.1% and 1% due to its high memory demand (>210 GB RAM for 10%). nnU-Net is evaluated under 1% and 10% settings but not 0.1% due to its requirement for full 3D scans in training.

2) Results on Fully Supervised and Label-efficient Settings: Table III summarizes the results under both fully supervised (100%) and label-efficient settings. In the fully supervised setting, ImageNet-pretrained models (DeepLabV3+, ResU-Net, U-Net++) and transformer-based architectures (Attention U-Net, UNETR, Swin UNETR) perform suboptimally on CT segmentation, with DSCs below 80%, suggesting the domain gap between natural and medical images. Among our variants, FT-dec consistently outperforms FS and LP, confirming the effectiveness of DDPM-based generative pre-training, though nnU-Net remains slightly stronger when trained with 100% labeled data.

In contrast, our FT-dec strategy shows clear superiority as labeled data decreases. With 10% and 1% labeled data, it achieves DSCs of 78.51% and 71.56%, exceeding nnU-Net by over 5% and DDPM-Seg/c256 by over 10%. Under the extreme 0.1% setting (four labeled slices), most supervised baselines fail to exceed 30% DSC, while our model maintains 51.81% DSC and 44.79% JI, outperforming the only comparable method, DDPM-Seg, by 6–8% in DSC and JI metrics.

[2]https://github.com/qubvel/segmentation_models.pytorch
[3]https://github.com/Project-MONAI/MONAI

TABLE III
COMPARISON OF PERFORMANCE UNDER DIFFERENT DATA RATIOS (BRACKETS SHOW GAPS VS. FULL DATA).

Method	DSC (%)				JI (%)			
	∼0.1%	1%	10%	100%	∼0.1%	1%	10%	100%
DeepLabV3+	20.74 (–47.01)	41.78 (–25.97)	58.71 (–9.04)	67.75	15.13 (–43.06)	34.84 (–23.35)	50.52 (–7.67)	58.19
ResU-Net	21.00 (–56.22)	41.29 (–35.93)	71.62 (–5.6)	77.22	16.06 (–53.24)	35.99 (–33.31)	63.47 (–5.83)	69.30
U-Net++	15.13 (–50.15)	34.22 (–31.06)	64.99 (–0.29)	65.28	11.45 (–45.86)	29.28 (–28.03)	56.58 (–0.73)	57.31
Attention U-Net	28.81 (–48.26)	50.70 (–26.37)	71.78 (–5.29)	77.07	21.93 (–46.71)	42.74 (–25.9)	63.16 (–5.48)	68.64
UNETR	13.87 (–50.85)	33.41 (–31.31)	54.91 (–9.81)	64.72	9.55 (–45.07)	26.31 (–9.47)	45.15 (–9.47)	54.62
Swin UNETR	28.21 (–45.65)	49.57 (–24.29)	70.19 (–3.67)	73.86	21.88 (–42.95)	42.06 (–22.77)	61.43 (–3.40)	64.83
nnU-Net	–	58.69 (–28.70)	73.43 (–13.96)	**87.39**	–	52.03 (–29.69)	66.75 (–14.97)	**81.72**
DDPM-Seg/c128	44.54	59.27	–	–	36.59	51.13	–	–
DDPM-Seg/c256	43.39	60.78	–	–	35.73	52.65	–	–
B/h128 FS (ours)	28.34 (–52.25)	60.07 (–20.52)	69.92 (–10.67)	80.59	23.27 (–51.45)	52.24 (–22.48)	64.26 (–10.46)	74.72
B/h256 FS (ours)	24.68 (–54.73)	58.10 (–21.31)	68.23 (–11.18)	79.41	19.82 (–53.96)	50.94 (–22.84)	62.51 (–11.27)	73.78
L/h256 FS (ours)	28.92 (–54.91)	54.71 (–29.12)	70.46 (–13.37)	83.83	24.01 (–53.15)	47.64 (–29.52)	64.97 (–12.19)	77.16
B/h128 FT-dec (ours)	**51.81** (–35.10)	**71.56** (–15.35)	**78.51** (–8.4)	86.91	**44.79** (–35.59)	**64.21** (–16.17)	**72.43** (–7.95)	80.38
B/h256 FT-dec (ours)	51.17 (–34.04)	70.25 (–14.96)	76.52 (–8.69)*	85.21	44.61 (–34.05)	63.31 (–15.35)	69.86 (–8.80)	78.66
L/h256 FT-dec (ours)	50.35 (–33.49)	69.07 (–14.77)	77.33 (–6.51)	83.84	43.22 (–33.65)	61.93 (–14.94)	71.23 (–5.64)	76.87

TABLE IV
ORGAN-LEVEL DSC SCORES UNDER DIFFERENT LABELED DATA RATIOS (RK: RIGHT KIDNEY, IVC: INFERIOR VENA CAVA, RAG: RIGHT ADRENAL GLAND, LAG: LEFT ADRENAL GLAND, AND LK: LEFT KIDNEY). "–" DENOTES THE DSC SCORE IS LESS THAN 1%.

Ratio	Methods	Liver	RK	Spleen	Pancreas	Aorta	IVC	RAG	LAG	Gallbladder	Esophagus	Stomach	Duodenum	LK
10%	nnU-Net	90.02	87.92	91.77	43.63	94.05	82.78	**62.68**	60.2	51.37	76.09	69.93	55.33	88.8
	B/h128 FS (ours)	92.28	94.52	93.31	60.65	94.42	87.06	–	–	71.17	82.03	77.65	62.03	93.89
	B/h128 FT-dec (ours)	**95.05**	**95.5**	**94.73**	**73.95**	**94.53**	**89.03**	–	**72.29**	**76.86**	**83.57**	**85.88**	**64.92**	**94.31**
1%	DDPM-Seg/c128	92.11	87.14	87.97	34.27	83.24	70.77	19.58	6.66	46.82	55.01	71.31	26.41	89.17
	DDPM-Seg/c256	92.76	85.96	88.9	35.97	87.8	72.34	10.45	23.42	38.82	57.83	**75.15**	30.0	90.76
	nnU-Net	82.91	84.01	82.23	31.02	82.56	71.34	38.03	33.09	13.06	61.49	65.18	**35.63**	82.4
	B/h128 FS (ours)	88.78	85.17	83.76	30.56	87.69	71.94	37.7	39.33	53.6	43.31	53.18	18.3	87.55
	B/h128 FT-dec (ours)	**94.14**	**93.69**	**89.33**	**41.67**	**93.4**	**83.2**	**55.56**	**53.23**	**66.77**	**63.68**	72.18	30.18	**93.14**
∼0.1%	DDPM-Seg/c128	85.85	78.21	75.37	22.95	76.6	59.51	–	–	20.46	5.96	**44.98**	**36.98**	72.21
	DDPM-Seg/c256	85.31	78.52	78.16	26.65	75.47	55.01	–	–	21.37	9.7	27.65	31.49	74.79
	B/h128 FS (ours)	66.61	54.51	50.26	5.79	55.04	35.46	–	–	21.54	–	4.71	9.61	64.11
	B/h128 FT-dec (ours)	**90.29**	**89.34**	**79.46**	**37.27**	**86.32**	**62.25**	–	**3.61**	**56.26**	**13.95**	38.02	28.49	**87.23**

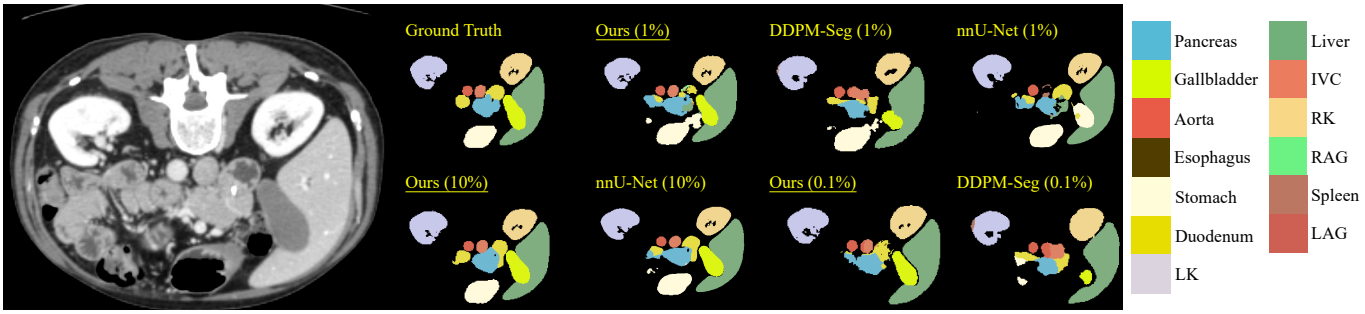

Fig. 4. Visualization of segmentation performance with different fine-tuning strategies across labeled data ratios.

We compare FT-dec with its FS counterpart, which shares the same architecture without DDPM pre-training. This variant exhibits degradation similar to supervised baselines, suggesting that improvements stem from generative pre-training rather than architectural superiority. Organ-level evaluation (Table IV) further shows that these improvements extend to most abdominal organs rather than only larger structures. Additional segmentation visualizations are provided in Fig. 4 to demonstrate that our framework maintains competitive accuracy and anatomical consistency even in extremely low-data regimes.

V. DISCUSSION AND CONCLUSION

Our study presents a DDPM-pretrained framework for label-efficient abdominal organ segmentation. By leveraging diffusion-based generative pre-training, the model learns anatomical priors from unlabeled CT data and effectively transfers them to downstream segmentation tasks. This bridges generative diffusion modeling with semantic segmentation, enabling high performance in label-scarce settings without additional inference cost. Experiments on the FLARE22 dataset verify that the proposed FT-dec strategy achieves superior

results compared to fully supervised methods, underscoring the potential of diffusion pre-training for scalable and annotation-efficient medical image analysis.

Despite these advantages, several limitations remain. The LP strategy underperforms from limited use of intermediate features, and future work may explore multi-layer integration for richer hierarchical semantics. Second, extending diffusion pre-training to higher-resolution or 3D volumetric CT data is still challenging. Finally, while the diffusion step strongly impacts performance, its underlying mechanism in learning transferable representations remains to be fully explored.

REFERENCES

[1] S. Zhang and D. Metaxas, "On the challenges and perspectives of foundation models for medical image analysis," *Medical Image Analysis*, vol. 91, p. 102996, 2024.

[2] R. Azad, E. K. Aghdam, A. Rauland, Y. Jia, A. H. Avval, A. Bozorgpour, S. Karimijafarbigloo, J. P. Cohen, E. Adeli, and D. Merhof, "Medical image segmentation review: The success of u-net," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[3] D. Karimi, H. Dou, S. K. Warfield, and A. Gholipour, "Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis," *Medical Image Analysis*, vol. 65, p. 101759, 2020.

[4] W. Shen, Z. Peng, X. Wang, H. Wang, J. Cen, D. Jiang, L. Xie, X. Yang, and Q. Tian, "A survey on label-efficient deep image segmentation: Bridging the gap between weak supervision and dense prediction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 8, pp. 9284–9305, 2023.

[5] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.

[6] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, 2022, pp. 16 000–16 009.

[7] G. Kim, T. Kwon, and J. C. Ye, "Diffusionclip: Text-guided diffusion models for robust image manipulation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 2426–2435.

[8] D. Baranchuk, A. Voynov, I. Rubachev, V. Khrulkov, and A. Babenko, "Label-efficient semantic segmentation with diffusion models," in *International Conference on Learning Representations*, 2022.

[9] E. A. Brempong, S. Kornblith, T. Chen, N. Parmar, M. Minderer, and M. Norouzi, "Denoising pretraining for semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4175–4186.

[10] A. Graikos, N. Malkin, N. Jojic, and D. Samaras, "Diffusion models as plug-and-play priors," *Advances in Neural Information Processing Systems*, vol. 35, pp. 14 715–14 728, 2022.

[11] A. Bielski and P. Favaro, "Emergence of object segmentation in perturbed generative models," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[12] N. Tritrong, P. Rewatbowornwong, and S. Suwajanakorn, "Repurposing gans for one-shot semantic part segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 4475–4485.

[13] D. Li, J. Yang, K. Kreis, A. Torralba, and S. Fidler, "Semantic segmentation with generative models: Semi-supervised learning and strong out-of-domain generalization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8300–8311.

[14] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[15] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *International Conference on Machine Learning*. PMLR, 2015, pp. 2256–2265.

[16] Y. Song and S. Ermon, "Generative modeling by estimating gradients of the data distribution," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[17] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020.

[18] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," *Advances in Neural Information Processing Systems*, vol. 34, pp. 8780–8794, 2021.

[19] 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.

[20] W. H. Pinaya, P.-D. Tudosiu, J. Dafflon, P. F. Da Costa, V. Fernandez, P. Nachev, S. Ourselin, and M. J. Cardoso, "Brain imaging generation with latent diffusion models," in *MICCAI DGM4MICCAI Workshop*. Springer, 2022, pp. 117–126.

[21] 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.

[22] 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.

[23] F. I. Diakogiannis, F. Waldner, P. Caccetta, and C. Wu, "Resunet-a: A deep learning framework for semantic segmentation of remotely sensed data," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 162, pp. 94–114, 2020.

[24] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE Transactions on Medical Imaging*, vol. 39, no. 6, pp. 1856–1867, 2019.

[25] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz *et al.*, "Attention u-net: Learning where to look for the pancreas," *arXiv preprint arXiv:1804.03999*, 2018.

[26] A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H. R. Roth, and D. Xu, "Unetr: Transformers for 3d medical image segmentation," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2022, pp. 574–584.

[27] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. R. Roth, and D. Xu, "Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images," in *MICCAI BrainLes Workshop*. Springer, 2022, pp. 272–284.

[28] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 9650–9660.

[29] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[30] Q. Xie, Y. Li, N. He, M. Ning, K. Ma, G. Wang, Y. Lian, and Y. Zheng, "Unsupervised domain adaptation for medical image segmentation by disentanglement learning and self-training," *IEEE Transactions on Medical Imaging*, vol. 43, no. 1, pp. 4–14, 2022.

[31] X. Chen, Y. Yuan, G. Zeng, and J. Wang, "Semi-supervised semantic segmentation with cross pseudo supervision," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 2613–2622.

[32] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[33] C. Nash, J. Menick, S. Dieleman, and P. Battaglia, "Generating images with sparse representations," in *International Conference on Machine Learning*. PMLR, 2021, pp. 7958–7968.

[34] T. Kynkäänniemi, T. Karras, S. Laine, J. Lehtinen, and T. Aila, "Improved precision and recall metric for assessing generative models," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[35] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818.

[36] 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.