Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2024.0429000

Unsupervised Anatomical Feature Learning via Diffusion Models: Enhanced Medical Image Segmentation with Denoising Diffusion Probabilistic Models

AKSHAT G[1], DIVYANSH GUPTA[1], SHALEEN BHATNAGAR[1] (Senior Member, IEEE) , SHILPA ANKALAKI[1], AND TUSAR KANTI MISHRA[1] (Senior Member, IEEE)

Manipal Institute of Technology Bengaluru, Manipal Academy of Higher Education, Manipal, India

Corresponding author: Shaleen Bhatnagar (shaleen.bhatnagar@manipal.edu)

This work was supported by Manipal Academy of Higher Education (Open Access Funding).

ABSTRACT The process of acquiring pixel and voxel level annotations is a bottleneck when it comes to medical image segmentation as it is extremely expensive. Furthermore, it requires expert diagnosis to manually delineate anatomical structures across three-dimensional volumes. Although traditional neural network architectures like U-net are affective, they operate as black boxes that learn local texture patterns. These models lack awareness of global anatomical structures. This leads to potential catastrophic failures in boundary delineation and poor generalization when quantity of labeled data is limited. This study proposes utilizing unsupervised Denoising Diffusion Probabilistic Models (DDPMs) to extract anatomical features. The proposed framework trains a DDPM on 21 unlabeled abdominal CT scans to learn structural representations through the iterative denoising process, which later transfers the learned encoder weights to perform a down-stream segmentation task. The framework is evaluated using data from the BTCV multi-organ dataset. Diffusion pretraining significantly improved liver segmentation where Dice increased from 0.75 ± 0.36 to 0.93 ± 0.16 ($p < 5.33\times10^{-26}$, with 0.529 Cohen's d), reduced 95th-percentile Hausdorff Distance (HD95) by 45% (11.76 to 6.50 mm), and decreased Average Surface Distance (ASD) by 66% (4.38 to 1.51 mm). For kidney segmentation, Dice improved from 0.90 ± 0.19 to 0.95 ± 0.10 ($p < 4.01\times10^{-11}$, with 0.271 Cohen's d), with 37% HD95 reduction. Multi-organ pooled performance showed the most dramatic improvement, where Dice increased from 0.78 ± 0.23 to 0.95 ± 0.07, and representing a 68% reduction in variance with 74% improvement in boundary precision. Furthermore, frozen encoder models achieved >80% of fine-tuned performance without seeing any segmentation labels, demonstrating the existence of learned anatomical priors. In low-data regimes, diffusion-pretrained models maintained robust performance with only 50% (Dice: 0.92 liver, 0.94 kidney), 25% (Dice: 0.90 liver, 0.81 kidney), and even 10% of labeled data (Dice: 0.89 liver, 0.71 kidney), substantially outperforming random initialization baselines. Using unlabeled images for diffusion-based pretraining helps embed robust, interpretable anatomical features into encoder network prior to any human supervision. The performance retention when compared to low-data scenarios, and the graceful degradation even with only 10% of labelled data confirms that the learned representations greatly improve segmentation accuracy, boundary precisions and model robustness. The diffusion pretraining effectively helps transform U-Nets from texture matching networks to anatomy aware systems.

INDEX TERMS Diffusion models, medical image segmentation, unsupervised learning, transfer learning, abdominal CT imaging, U-Net, DDPM, anatomical feature learning.

I. INTRODUCTION

MODERN medical artificial intelligence relies heavily on expert annotated datasets. For medical image segmentation, pixel level segmentation of anatomical structures and pathologies are required, which in turn creates a severe bottleneck [1]. Unlike image classification where labels can be crowd-sourced, medical segmentation required trained radiologists to manually trace out and segment organ bound-

aries across three dimensional volumes, one slice at a time. A single CT scan containing anywhere between 100 to 300 slices may require an expert radiologist 30 to 60 minutes of annotation time, with costs often exceeding 100 dollars per volume [1]–[3].

Inter labeler variability introduces additional challenges, previous studies have shown that even expert radiologists disagree on precise boundary delineation [4], [5], particularly for organs with diffused edges like liver. This variability in turn propagates into the training data, resulting in inconsistent signals that limit model performance and reliability [6], [7]. In a resource constrained clinical setting, or for rare pathologies, obtaining sufficient annotated data becomes practically infeasible, such scenarios are where automated assistance would provide the greatest clinical value [5], [7].

Since 2015, the U-Net architecture and its variants have been leading choice of use for medical image segmentation tasks [8]. This is because of their ability to achieve competitive results on benchmarking datasets [8], [9]. However, these models have a few fundamental limitations [9], [10]. A randomly initialized standard U-Net learns individual features which are optimized to minimize pixel-wise classification loss on training data. This optimization process helps the network in identifying local texture patterns, edge orientations, intensity gradients, and co-occurrence statistics that correlate with organ boundaries in the training set.

This texture-focused learning reveals three significant failure modes:

- **Boundary failures:** Models usually struggle with segmenting boundaries precisely in case of organs with irregular shapes or variable contrasts. A high 95th-percentile Hausdorff Distance (HD95) suggests that although overall segmentation may seem reasonable, inaccuracies such as missing lobes, incorrect attachments, or disconnected regions may be present at the boundaries [9], [11].
- **Lack of global spatial context:** Convolution inherently is a local operation. While the skip connections from the U-Net help propagate features, randomly initialized encoders lack any prior understanding of anatomical topology, such as livers occupy specific regions, kidneys come in pairs, or that specific organs follow predictable size ranges [12], [13].
- **Poor performance in low-data regimes:** When ground truth for a specific task is limited, texture-based discrimination becomes unreliable. Models will tend to overfit spurious correlations in limited training examples. This results in failures to generalize to unseen anatomical variations, imaging protocols or certain patient demographics [9], [14].

This study proposes a paradigm shift where rather than training segmentation models from random initialization, a DDPM [15] is deployed to learn unsupervised anatomical representations. Diffusion models have recently been adopted for generative modeling, and producing synthetic images by learning to reverse a gradual noise corruption process [16]–[21]. However, the objective of this study focuses on leveraging the reverse diffusion process not to generate synthetic images, but to force an encoder to learn the underlying structural patterns of abdominal anatomy.

To predict and remove noise at intermediate steps, the network must learn anatomy such as organ shapes, positions, relative arrangements. The model must also learn boundary characteristics of organs, and background with respect to each other along with global spatial relationships and size constraints. The learning process in DDPM occurs without manual annotation, and the training signal comes purely from the data distribution itself. After pretraining the model with images, the encoder weights are transferred to downstream segmentation task, where even with limited labeled data, the encoder weights can be fine tuned with anatomically aware features for precise boundary prediction.

The main contributions of this study are:

1) **Statistically significant improvements in spatial boundary precision:** The results inferred from the study demonstrate that diffusion pretraining produces statistically substantial and reproducible improvements in segmentation accuracy across multiple metrics like Dice [22], IoU [23] and predominantly in boundary precision metrics like HD95 [24] and ASD [25]. Statistical analysis using Wilcoxon signed-rank tests with Bonferroni correction confirms significance at ($p < 10^{-10}$) [26], with medium effect sizes (Cohen's d = 0.271–0.529) [27].
2) **Variance reduction and improved robustness:** Diffusion pretraining reduces prediction variance by 47–68%, inferred with improved generalization across diverse anatomical configurations and imaging conditions.
3) **Evidence for learned anatomical priors:** Frozen encoder experiments demonstrate that DDPM-pretrained encoders retain over 80% of the performance of fully fine-tuned models for liver, despite never being trained on segmentation label. This provides direct evidence that the unsupervised pretraining phase learns meaningful anatomical structure, not just generic image features.
4) **Multi-organ generalization:** This study shows that single diffusion pretraining benefits multiple anatomically distinct organs (liver and kidney), with multi-organ models achieving the best overall performance. This suggests that diffusion models learn generalizable abdominal anatomy representations rather than organ-specific patterns.

II. RELATED WORK

A. DEEP LEARNING IN MEDICAL IMAGE SEGMENTATION

Fully Convolutional Networks(FCNs) [28]–[32] were one of the earliest applications of deep-learning for medical image segmentation. This field was later redefined by the introduction of the U-Net architecture [8]. The skip connections first hypothesized in the U-Net combine high-resolution spatial

information from the encoder with semantic features from the decoder. This concept standardized well with the requirements of precise boundary localization in medical imaging.

Subsequent work resulting from U-Net has explored numerous architectural variations. V-Net extended U-Net to 3D volumetric segmentation with residual connections [33]. Attention U-Net incorporated attention mechanisms to help the network focus on relevant spatial regions [34]. nnU-Net systematized hyperparameter selection and preprocessing, achieving strong performance across diverse medical imaging tasks through careful engineering rather than architectural novelty [35]. More recently, Vision Transformers have been adapted for medical segmentation, with architectures like Swin-UNETR demonstrating that self-attention mechanisms can capture long-range dependencies that convolutions miss [36], [37].

Despite these advances, a common limitation for these models is the fact that they require substantial amounts of labeled training data and learn features from random initialization [38]–[41].

B. SELF-SUPERVISED REPRESENTATION LEARNING

The goal of Self-Supervised Learning (SSL) [42], [43] is to learn useful representation from unlabeled data by defining a pretext task. This pretext task provides sensory input without the requirement of manual annotation. Learning methods like SimCLR and Moco have achieved success in computer vision and contrastive learning tasks. They do this by learning representations that maximize agreement between different augmented views of the same image while also minimizing the similarity between two different augmented views of different images [44]–[46].

Several SSL approaches have been explored prior for medical imaging. Autoencoders learn compressed representations by reconstructing input images, forcing the bottleneck to capture essential information [47], [48]. Contrastive learning has been applied using domain-specific augmentations like random cropping, rotation, and intensity transformations [44]. Masked Autoencoders (MAE), inspired by BERT in NLP, randomly mask patches of images and train networks to reconstruct the missing regions [49], [50].

However, these methods have limitations for dense prediction tasks. Contrastive learning optimizes for image-level representations, potentially discarding fine-grained spatial information crucial for segmentation [51]. Standard autoencoders with MSE reconstruction loss may learn to reproduce textures and intensities without capturing higher-level anatomical structure. MAE's discrete masking strategy can miss subtle boundary information [52].

Alternatively, the gradual denoising process of diffusion requires learning of hierarchical representations at different scales, from fine to global textures. This denoising objective explicitly encourages the network to understand spatial relationships and structural coherence, which makes it particularly advantageous for segmentation pretraining.

C. DIFFUSION MODELS FOR FEATURE EXTRACTION

Denoising Diffusion Probabilistic Models (DDPMs) [15] are used to generate high-quality image data, quickly exceeding GANs in sample quality and training stability [53]–[55]. This fundamental approach is based on the principle of gradually corrupting data with noise over numerous iterations and timesteps, followed by training a neural network to reverse this process.

Recent work have started exploring the application of diffusion models for purposes beyond generation. Evidence suggests that the features acquired from diffusion models are effectively transferable to classification tasks involving natural images [56]. Furthermore, it has been observed that diffusion models implicitly learn sematic segmentation during the generation process [57]. In medical imaging, diffusion models have been directly applied to the segmentation tasks, using the reverse process to iteratively enhance segmentation [58].

The current study expands upon these foundations while also diverging in different key aspects. Instead of utilizing diffusion for comprehensive segmentation, which is computationally expensive at inference time, an unsupervised pretraining method is used, focusing completely on extracting only the learned encoder weights.

This approach provides the benefits of diffusion's comprehensive learned representations while ensuring the effectiveness of conventional segmentation architectures during deployment. Furthermore, this study focuses on limited data, critical boundary precision requirements, and the need for interpretable anatomically grounded features.

III. METHODS

A. DATASET AND PREPROCESSING PIPELINE

1) Dataset Description

The methodology was evaluated on the Beyond the Cranial Vault (BTCV) multi-organ abdominal CT segmentation dataset [59], a widely-used benchmark in medical image segmentation research. Sample images from the dataset are shown in Fig. 1. The dataset comprises 30 abdominal CT scans from patients undergoing radiotherapy treatment, with expert annotations for 13 anatomical structures including liver, kidneys, spleen, stomach, gallbladder, pancreas, and various vascular structures. Images were acquired using clinical protocols with varying slice thickness (2.5–5.0 mm), in-plane resolution (0.54–0.98 mm), and scanner manufacturers, providing realistic heterogeneity representative of clinical deployment scenarios [60].

2) Preprocessing Pipeline

A domain-specific preprocessing pipeline, as demonstrated in Fig. 2, was utilized to improve soft tissue visualization. In addition to image-level processing, patient-level preprocessing was performed to ensure consistency across volumetric samples. Algorithm 1 summarizes the full preprocessing pipeline, including both image-level and patient-level operations.

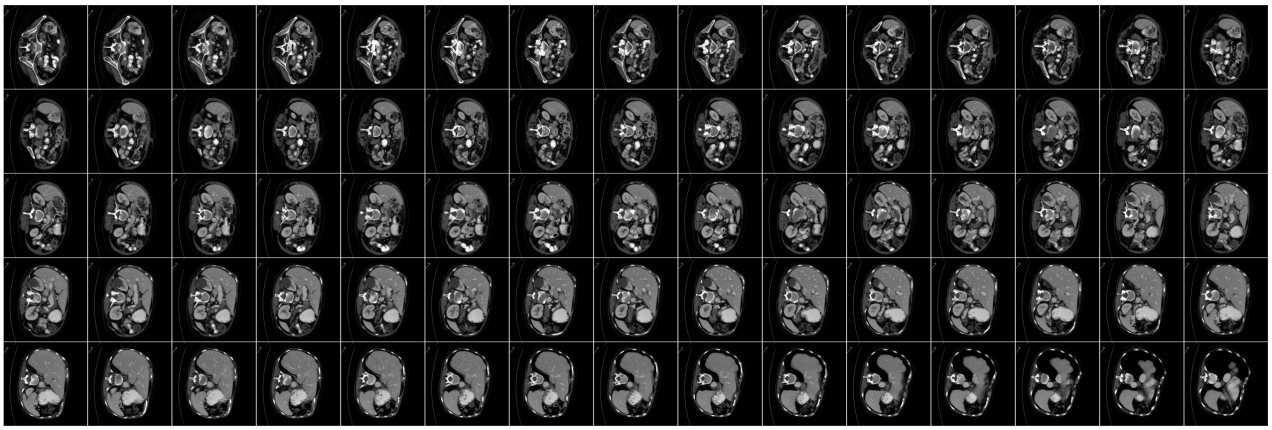

FIGURE 1. Sliced 2D images from a single patient in the dataset.

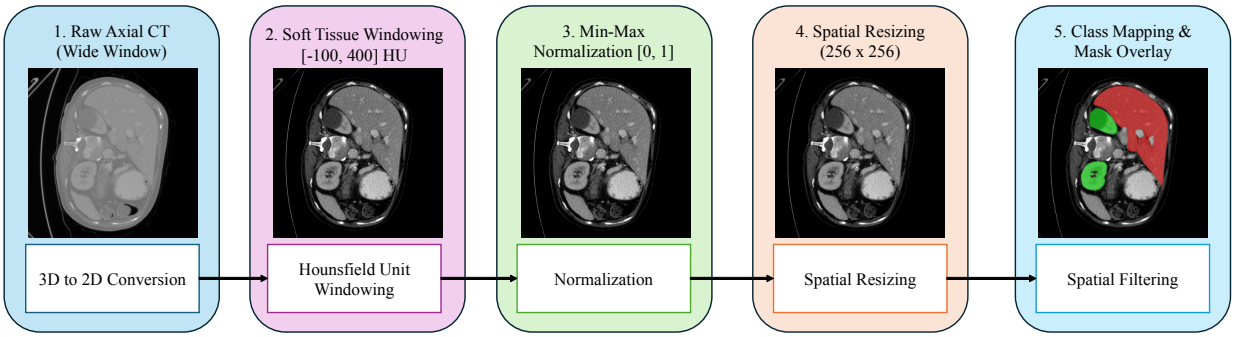

FIGURE 2. A domain-specific preprocessing pipeline optimized for soft tissue visualization.

Sub-heading: Image Level Preprocessing. Initially three-dimensional NIfTI volumes are converted to 2D axial slices. CT intensities were clipped to the range $[-100, 400]$ HU [61], [62], optimized for soft tissue contrast. After windowing, intensities were linearly normalized to [0, 1] to standardize input ranges across different scanners and acquisition protocols. All CT slices were resized to 256×256 pixels using bilinear interpolation, balancing computation efficiency and preserving anatomical details. Only axial slices containing liver or kidney annotations were retained based on ground-truth masks, discarding empty slices to focus computational resources on relevant anatomy. This filtering was performed using the ground-truth masks to ensure consistent slice selection across train/validation/test splits.

Patient Level Data Splitting. Prevention of data leaks is critical to this experimental design process. Patient Level Splitting was the focus of the current study, as opposed to Slice Level or Volume Level Splitting. In Patient Level Splitting, each slice comes from only one patient, in either a training, validation or test set.

This is not the case for the Slice Level and Volume Level Splits where multiple patients' slices are considered in training, validation, and test sets. A large portion of the data is from adjacent slices of the same patient; they contain very similar anatomical features, have minor differences in position and appearance. If Slice Level Splits were done, the model would memorize specific anatomical characteristics associated with training patients, and then would be evaluated against patient-specific samples (both during testing) resulting in inflated performance metrics.

In this study, a patient level split of 70%, 15% and 15% was achieved for the training, validation, and test sets respectively. Each of the patients was logged to ensure reproducibility of the analysis. Random seeds were fixed across each experiment to ensure deterministic results. The test set was never used for selection of the optimal model, hyperparameters or any early stopping decisions.

B. SELF-SUPERVISION USING DIFFUSION

1) Mathematical Formulation

Denoising Diffusion Probabilistic Models (DDPMs) consist of two parameterized Markov chains: a forward diffusion process that systematically corrupts the data, and a learned reverse process that reconstructs it.

The Forward Process. Given an uncorrupted input image x_0 sampled from the real data distribution $q(x)$, the forward pro-

Algorithm 1 Abdominal CT Image Preprocessing Pipeline

Require: Set of 3D CT Volumes I and corresponding Masks M for patients P, Target Labels $L_{target} = \{6(\text{Liver}), 2(\text{R. Kidney}), 3(\text{L. Kidney})\}$

Ensure: Preprocessed 2D slice set S, Patient-level data splits $(P_{train}, P_{val}, P_{test})$

$S \leftarrow \emptyset$, $P_{valid} \leftarrow \emptyset$
for each patient $p \in P$ **do**
  Extract 3D volume I_p and mask M_p
  if $\dim(I_p) = \dim(M_p)$ **then**
    $Z \leftarrow$ number of axial slices in I_p
    has_target $\leftarrow$ False
    for $z = 1$ **to** Z **do**
      $I_{slice} \leftarrow I_p[:, :, z]$
      $M_{slice} \leftarrow M_p[:, :, z]$
      if $M_{slice} \cap L_{target} \neq \emptyset$ **then**
        has_target $\leftarrow$ True
        {1. Hounsfield Unit Windowing}
        $I_{slice} \leftarrow \max(\min(I_{slice}, 400), -100)$
        {2. Min-Max Normalization}
        $I_{slice} \leftarrow \frac{I_{slice} - (-100)}{400 - (-100)}$
        {3. Spatial Resizing}
        $I_{slice} \leftarrow \text{Resize}(I_{slice}, 256 \times 256, \text{method} = \text{Area})$
        $M_{slice} \leftarrow \text{Resize}(M_{slice}, 256 \times 256, \text{method} = \text{Nearest})$
        {4. Class Mapping}
        $M_{final} \leftarrow \text{zeros\_like}(M_{slice})$
        $M_{final}[M_{slice} == 6] \leftarrow 1$ {Map Liver}
        $M_{final}[M_{slice} \in \{2, 3\}] \leftarrow 2$ {Map Kidneys}
        Add (I_{slice}, M_{final}) to S
      end if
    end for
    if *has_target* **then**
      Add p to P_{valid}
    end if
  end if
end for
{5. Strict Patient-Level Splitting}
$P_{train}, P_{temp} \leftarrow \text{Split}(P_{valid}, \text{ratio} = 70 : 30)$
$P_{val}, P_{test} \leftarrow \text{Split}(P_{temp}, \text{ratio} = 50 : 50)$
return $S, P_{train}, P_{val}, P_{test}$

cess q gradually adds Gaussian noise over T timesteps. The amount of noise at each step is controlled by a fixed variance schedule $\beta_1, \ldots, \beta_T \in (0, 1)$. The transition probability for a single step is defined as:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t;\ \sqrt{1-\beta_t}\, x_{t-1},\ \beta_t I) \tag{1}$$

The complete forward process is the product of these conditional probabilities. A critical mathematical property of this process is that it allows us to sample x_t at any arbitrary timestep directly from x_0 without iterating through all previous steps. By defining $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$, the marginal distribution becomes:

$$q(x_t|x_0) = \mathcal{N}(x_t;\ \sqrt{\bar{\alpha}_t}\, x_0,\ (1 - \bar{\alpha}_t) I) \tag{2}$$

Using the reparameterization trick, we can express this as $x_t = \sqrt{\bar{\alpha}_t}\, x_0 + \sqrt{1 - \bar{\alpha}_t}\, \epsilon$, where $\epsilon \sim \mathcal{N}(0, I)$. As $t \to \infty$, the parameter $\bar{\alpha}_t \to 0$, forcing the data distribution to converge to an isotropic standard normal distribution $\mathcal{N}(0, I)$.

The Reverse Process. The objective of the network is to reverse this noise addition to maximize the log-likelihood of generating a real sample, $\log p_\theta(x_0)$. Since the true reversal step $q(x_{t-1}|x_t)$ is computationally intractable, it is approximated using a neural network p_θ:

$$p_\theta(x_{0:T}) = p(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t) \tag{3}$$

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1};\ \mu_\theta(x_t, t),\ \Sigma_\theta(x_t, t)) \tag{4}$$

KL Divergence and the True Posterior. Training the network requires minimizing the negative log-likelihood $-\log p_\theta(x_0)$. Using Bayesian variational inference, this is bounded by minimizing a combination of terms, the most critical being the Kullback-Leibler (KL) divergence. The KL divergence measures the "distance in probability space" between the network's prediction and the true posterior conditioned on x_0:

$$L_t = \mathbb{E}_q \left[D_{\text{KL}}(q(x_{t-1}|x_t, x_0) \,\|\, p_\theta(x_{t-1}|x_t)) \right] \tag{5}$$

Here, $q(x_{t-1}|x_t, x_0)$ represents the forward process utilizing the known original image x_0 to perfectly calculate the previous step distribution. While using x_0 during generation is impossible, it provides a perfect target during training to approximate the parameters θ.

Reparameterization to Simplified Loss. Both the true posterior and the predicted distribution are Gaussian distributions: $q(x_{t-1}|x_t, x_0) = \mathcal{N}(\tilde{\mu}_t, \tilde{\beta}_t I)$ and $p_\theta(x_{t-1}|x_t) = \mathcal{N}(\mu_\theta, \sigma_\theta^2 I)$. Consequently, minimizing their KL divergence simplifies to minimizing the sum of L2 distances between their means:

$$L_t = \mathbb{E}_q \left[\frac{1}{2\sigma_t^2} \|\tilde{\mu}_t - \mu_\theta(x_t, t)\|^2 \right] \tag{6}$$

Through further algebraic reparameterization, rather than predicting the mean μ, the neural network $\epsilon_\theta(x_t, t)$ is trained to directly predict the noise ϵ that was added at timestep t. Removing the variance-weighting factors yields the final, simplified loss function:

$$L_{\text{simple}} = \mathbb{E}_{t,\, x_0,\, \epsilon} \left[\left\| \epsilon - \epsilon_\theta\left(\sqrt{\bar{\alpha}_t}\, x_0 + \sqrt{1 - \bar{\alpha}_t}\, \epsilon,\ t\right) \right\|^2 \right] \tag{7}$$

This robust, unweighted objective improves training stability and forces the encoder to learn the underlying global anatomical structures required to successfully denoise the image across all possible noise scales.

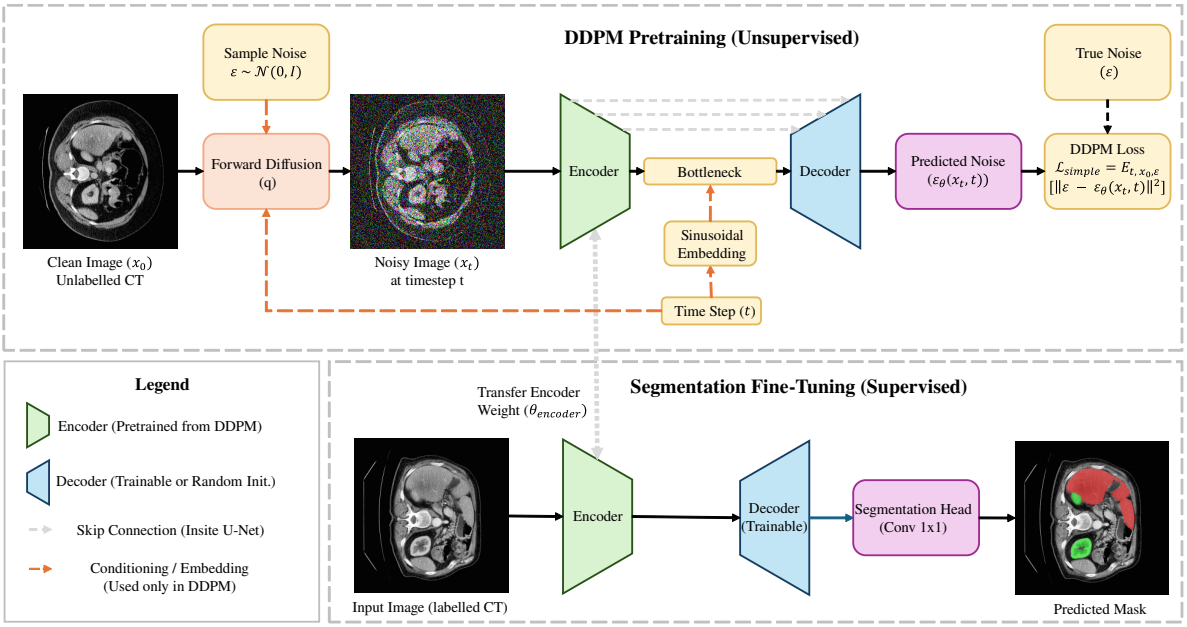

FIGURE 3. DDPM architecture employing U-Net backbone.

2) Architecture and Implementation

The DDPM architecture employs a U-Net backbone with several key design choices optimized for transfer to downstream segmentation tasks:

1) **Encoder-decoder structure:** The network follows the standard U-Net architecture with skip connections, ensuring that pretrained encoder can later integrate seamlessly with a segmentation decoder.
2) **Timestep conditioning:** The timestep t is embedded using sinusoidal position encodings and inject exclusively at the bottleneck layer. This maintains -1 to 1 structural compatibility with standard U-Nets providing the temporal conditioning necessary for diffusion training.
3) **Reduced timesteps:** T = 500 timesteps are used rather than the standard 1,000. This reduces computational cost by 2 fold while maintaining sufficient granularity for learning meaningful denoising features.
4) **Training data:** The DDPM is trained on all available abdominal CT slices without using any annotations. This unsupervised phase sees the full data distribution regardless of which subset will later be used for supervised fine-tuning.

Training proceeded for 500 epochs using the Adam optimizer with learning rate 10^{-4}. Reconstruction quality was monitored by visualizing denoised samples at every t = 50 to verify that the model learned meaningful structure at different noise levels.

C. SUPERVISED FINE-TUNING AND TRANSFER LEARNING

After performing unsupervised DDPM pretraining, the weights learned in the encoder are transferred to downstream segmentation tasks. This study performs three distinct transfer strategies to understand how diffusion contributes to segmentation performance. These include:

1) Transfer Study 1: Random Initialization

Models M1–M3 mentioned in Table 1 represent baseline for the study. The entire U-Net is initialized with random weights (He initialization [63]) and trained end-to-end using only labeled segmentation data. This represents the standard supervised learning approach and provides the performance reference for measuring pretraining benefits.

2) Transfer Study 2: Fine-Tune DDPM Encoder

Models M4–M6 as mentioned in Table 1, initialize the encoder with DDPM-pretrained weights while randomly initializing the decoder and segmentation head. During training, all parameters, encoder, decoder, and head; are fine-tuned using the segmentation loss (Equation 8). This strategy allows the network to adapt pretrained anatomical features to the specific requirements of pixel-wise segmentation while preserving learned structural knowledge.

3) Transfer Study 3: Frozen DDPM Encoder

Models M7–M9 as mentioned in Table 1 use DDPM-pretrained encoders with frozen weights encoder parameters are not updated during segmentation training. Only the decoder and segmentation head learn from labeled data. This experimental condition directly tests whether unsupervised diffusion pretraining creates useful anatomical features without any task-specific adaptation. Strong performance from frozen encoders would provide compelling evidence for learned anatomical priors.

TABLE 1. Model Names and Descriptions.

Model Name	Initialization	Organ
M1	Random	Liver
M2	Random	Kidney
M3	Random	Multi Organ
M4	Fine-Tuned Diffusion Encoder	Liver
M5	Fine-Tuned Diffusion Encoder	Kidney
M6	Fine-Tuned Diffusion Encoder	Multi Organ
M7	Frozen Diffusion Encoder	Liver
M8	Frozen Diffusion Encoder	Kidney
M9	Frozen Diffusion Encoder	Multi Organ

4) Segmentation Training Details

All segmentation models were trained using the Dice loss function, which directly optimizes the evaluation metric:

$$L_{\text{Dice}} = 1 - \frac{2\sum_i p_i g_i + \varepsilon}{\sum_i p_i^2 + \sum_i g_i^2 + \varepsilon} \tag{8}$$

where p_i is the predicted probability for pixel i, g_i is the ground truth binary label, and $\varepsilon = 10^{-5}$ prevents division by zero. Dice loss is particularly effective for medical segmentation because it handles class imbalance naturally and directly optimizes region overlap.

The study trained three task variants for each of the transfer strategy, (1) single-organ liver segmentation, (2) single-organ kidney segmentation and (3) multi-organ joint segmentation with three output classes. The respective correspondence for each model has been detailed in Table 1. All models were trained for 50 epochs using Adam optimizer with learning rate 10^{-4}, batch size 8, and standard data augmentation comprising random horizontal flips, random rotations ($\pm 15°$), and random scaling (0.9–1.1×).

D. LOW-DATA SIMULATION

To evaluate the robustness of diffusion pretraining under data scarcity, this study conducts semantic experiments with restricted fine-tuning data. The DDPM pretraining phase was conducted with the complete data, only the supervised-finetuning was performed on the restricted label set.

TABLE 2. Model names and percentage of labeled data for Low Data Simulation Study.

Model Name	Percentage of labeled data	Organ
M10	50%	Liver
M11	50%	Kidney
M12	25%	Liver
M13	25%	Kidney
M14	10%	Liver
M15	10%	Kidney

E. EVALUATION METRICS AND STATISTICAL FRAMEWORK

1) Segmentation Quality Metrics

Four metrics were used to systematically evaluate segmentation performance, capturing both region overlap and boundary accuracy. These metrics include:

1) **Dice Coefficient** [22]: The primary metric for measuring volumetric overlap, ranging from 0 (no overlap) to 1 (perfect agreement).
2) **Intersection over Union (IoU)** [23]: Also known as the Jaccard index, IoU is more sensitive to errors than Dice, penalizing false positives and false negatives more heavily.
3) **95th Percentile Hausdorff Distance (HD95)** [24]: Measures boundary accuracy using the 95th percentile of point-to-surface distances, particularly important for detecting huge boundary errors that Dice might miss.
4) **Average Surface Distance (ASD)** [25]: Calculates the mean distance between corresponding boundary points, offering a smooth and reliable assessment of the overall quality of the boundary.

2) Statistical Significance Testing

The rigorous hypothesis testing employed a non-parametric Wilcoxon signed-rank test for paired comparisons [26], which is suitable for bounded metrics such as Dice scores. To account for multiple comparisons, a Bonferroni correction was applied. Cohen's d effect sizes [27] quantified practical significance, interpreted as small ($0.2 \leq |d| < 0.5$), medium ($0.5 \leq |d| < 0.8$), or large ($|d| \geq 0.8$).

IV. RESULTS

A. IMPLEMENTATION DETAILS

All experiments were carried out in accordance with the methodology described in Section III. The DDPM pretraining phase employed the entire unlabeled training set of 21 patients from the BTCV dataset with 1000 diffusion timesteps and trained for 500 epochs on NVIDIA A2000 GPUs. For the downstream segmentation tasks, models M1==M9 were trained on the complete labeled training set, which consisted of a 70% patient-level split involving 21 patients. In contrast, models M10–M15 were designed to simulate low-data regimes, being trained on 50%, 25% and 10% of the labeled data, respectively. All segmentation models implemented Dice loss optimization (Equation 8) alongside Adam optimizer [64] (learning rate 10^{-4}), a batch size of 8, and were trained for 50 epochs while applying standard augmentation techniques, including horizontal flips, $\pm 15°$ rotation, 0.9–1.1× scaling. The assessment of statistical significance was conducted using Wilcoxon signed-rank tests, applying Bonferroni correction for multiple comparisons (α = 0.05), and effect sizes were measured using Cohen's d.

B. QUANTITATIVE PERFORMANCE ANALYSIS

1) Comprehensive Segmentation Metrics

Table 3 presents the complete segmentation performance of the 15 models. The models were evaluated on four metrics, being, Dice Coefficient and IoU for volumetric overlap, and HD95 and ASD for boundary precision.

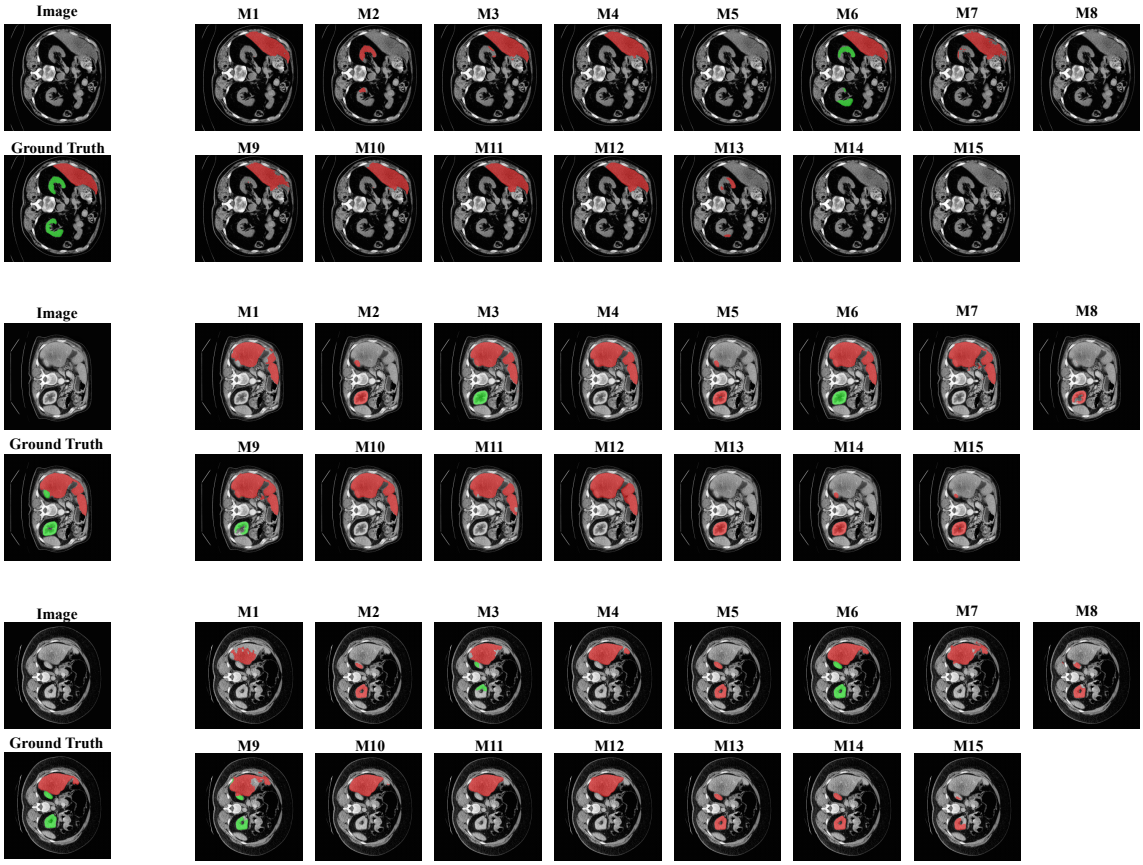

FIGURE 4. **Qualitative Segmentation Predictions Across All 15 Models on Three Representative Test Cases. Axial CT slices with ground truth masks (red: single class representing kidney/liver; red and green for multi class, where red represents liver and green represents kidney) overlaid with predictions from models M1–M15. Columns represent the three test cases; rows show predictions from baseline (M1–M3), fine-tuned diffusion (M4–M6), frozen diffusion (M7–M9), and low-data models (M10–M15). Visual inspection reveals systematic differences in boundary precision, false positive rates, and anatomical plausibility.**

TABLE 3. **Comprehensive Segmentation Performance Across All Models.**

Model	Dice	IoU	HD95 (mm)	ASD (mm)
M1	0.751±0.363	0.703±0.354	11.76±16.44	4.38±11.11
M2	0.899±0.192	0.853±0.216	14.00±27.32	3.75±8.13
M3	0.783±0.225	0.724±0.236	15.49±18.26	4.26±5.69
M4	0.928±0.160	0.891±0.178	6.50±11.13	1.51±2.08
M5	0.950±0.101	0.916±0.129	8.89±25.25	2.47±7.55
M6	0.950±0.071	0.916±0.081	6.35±12.92	1.42±2.55
M7	0.806±0.284	0.740±0.278	29.04±28.75	6.83±9.07
M8	0.654±0.409	0.604±0.393	22.05±28.84	6.32±12.89
M9	0.739±0.249	0.681±0.244	23.96±24.28	6.15±7.97
M10	0.922±0.173	0.884±0.182	6.44±9.80	1.47±2.01
M11	0.904±0.199	0.860±0.203	9.09±14.40	2.12±3.70
M12	0.886±0.193	0.830±0.204	14.19±18.35	3.29±4.91
M13	0.941±0.124	0.905±0.143	8.08±19.07	1.72±3.60
M14	0.814±0.341	0.783±0.338	9.89±23.61	3.24±10.20
M15	0.712±0.363	0.653±0.364	27.05±36.32	8.60±15.04

2) Statistical Significance Analysis

Three primary statistical comparisons were conducted. The results from these statistical comparisons have been tabulated in Table 4. The first comparison axis was the baseline versus diffusion-pretrained models. The second comparison was between fine-tuned versus frozen encoder configurations, and the third comparison was for data-efficiency in degraded patterns. All comparisons achieved statistical significance ($p < 0.05$ after Bonferroni correction), with effect sizes ranging from small to large.

3) Performance Metric Analysis

For liver segmentation, the Dice Coefficient improved from 0.751±0.363 to 0.928±0.160, which represents a net increase of 23.6% relative to the base mode. Furthermore, variance reduced by over 55% (0.363 to 0.160 standard deviation). These results reveal improvements that exceed statistical significance and the reduction in variance indicates significantly improved reliability and clinical relevance. For multi-organ model (M3 vs M6) the Dice increased from 0.783±0.225 to 0.950±0.071, which represents a mean improvement of 21.3%. Variance reduced by over 68% and thus provides conclusive evidence for model performance.

For liver, diffusion pretraining reduces HD95 by 44.7% (11.76 to 6.50 mm) and ASD by 65.5% (4.38 to 1.51 mm). The multi-organ model had greater improvements with respect to boundary metrics, where, HD95 reduced by over 59.0% from (15.49 to 6.35 mm) and ASD reduced by over 66.6% (4.26 to 1.42 mm). The sub-2 millimeter ADS values

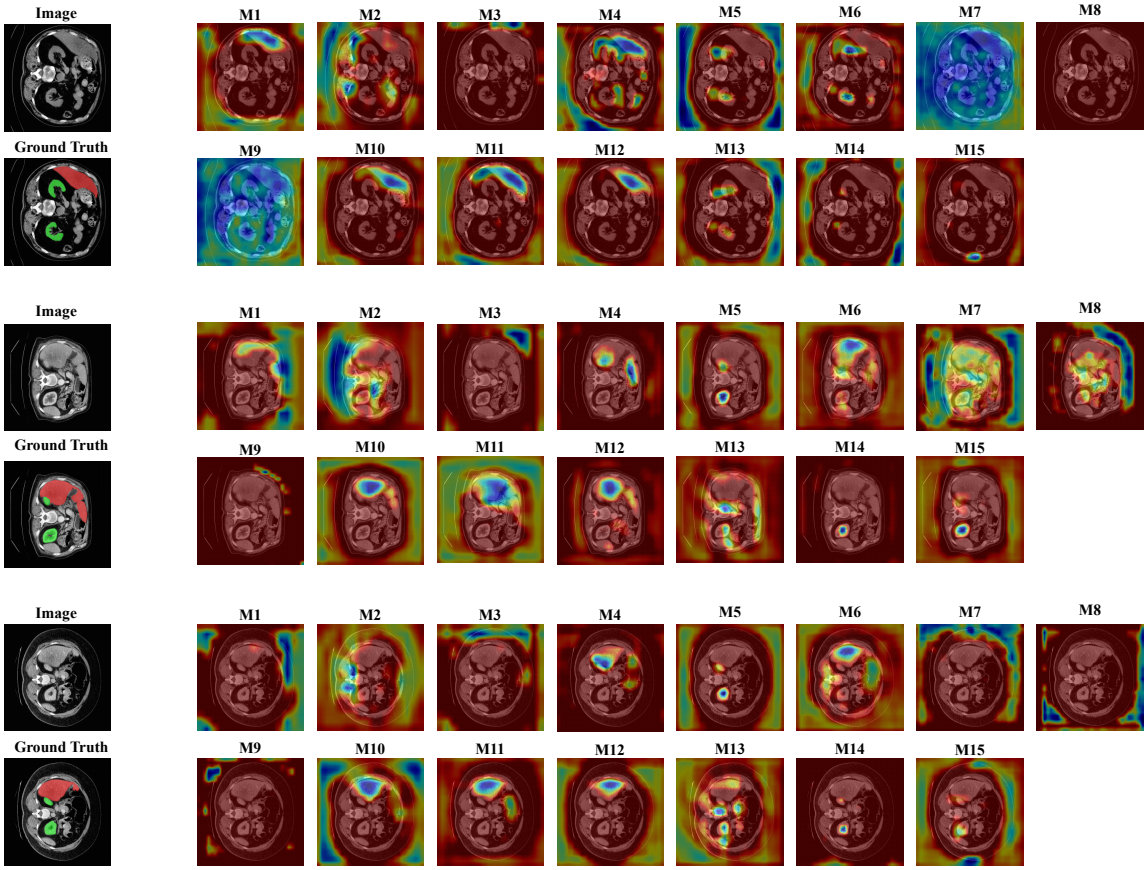

FIGURE 5. **Grad-CAM Attention Heatmaps from Bottleneck Layer Across All 15 Models. Heatmaps overlaid on input CT slices for models M1–M15 on three representative cases. Warm colors (red/yellow) indicate high feature activation; cool colors (blue/green) show low activation. The heatmaps reveal where the network's deepest convolutional features focus when making segmentation predictions.**

TABLE 4. **Statistical Significance Analysis (Wilcoxon Signed-Rank Test with Bonferroni Correction).**

Comparison	Models	Raw p-value	Adjusted p-value	Cohen's d	Effect Size
Baseline vs Fine-Tuned Diffusion					
Liver	M1 vs M4	2.66×10^{-26}	3.20×10^{-25}	0.529	Medium
Kidney	M2 vs M5	2.00×10^{-11}	2.41×10^{-10}	0.271	Small
Multi-Organ	M3 vs M6	9.40×10^{-31}	1.13×10^{-29}	0.766	Medium
Fine-Tuned vs Frozen Encoder					
Liver	M4 vs M7	6.34×10^{-28}	7.61×10^{-27}	−0.475	Small
Kidney	M5 vs M8	4.22×10^{-27}	5.06×10^{-26}	−0.737	Medium
Multi-Organ	M6 vs M9	4.74×10^{-34}	5.69×10^{-33}	−0.897	Large
Data Efficiency (Liver)					
100% vs 50%	M4 vs M10	4.79×10^{-8}	5.75×10^{-7}	−0.071	Small
100% vs 25%	M4 vs M11	5.78×10^{-15}	6.94×10^{-14}	−0.118	Small
100% vs 10%	M4 vs M12	7.38×10^{-22}	8.86×10^{-21}	−0.226	Small
Data Efficiency (Kidney)					
100% vs 50%	M5 vs M13	2.05×10^{-4}	2.46×10^{-3}	−0.071	Small
100% vs 25%	M5 vs M14	4.62×10^{-14}	5.54×10^{-13}	−0.436	Small
100% vs 10%	M5 vs M15	2.20×10^{-25}	2.64×10^{-24}	−0.686	Medium

are in the range usually considered as inter-rater variability in manual annotations. This suggests that diffusion-pretrained models achieve close to human-level boundary precision.

It is important to emphasize that the noted increase in boundary measures compared to Dice indicates an important discovery – namely, the fact that diffusion pre-training not only improves the number of correct pixel classification (and thus results in a linear increase in the value of Dice), but also changes the distribution of the mistakes. While baseline networks show completely random mistakes and non-uniform

boundaries, resulting in a high HD95 value with low Dice values, diffusion networks create spatially consistent predictions, with plausible anatomy.

The frozen encoder experiments (M7–M9) provide the most compelling evidence that DDPM pretraining learns genuine anatomical structure rather than generic image features. The organ-specific results reveal nuanced insights:

1) **Liver (M7):** A Dice of 0.806±0.284 represents 86.9% of the fine-tuned performance (M4: 0.928), with the frozen encoder outperforming the random baseline (M1: 0.751) by 7.3%. This demonstrates that unsupervised diffusion pretraining alone; without ever seeing segmentation labels; learns liver-specific features sufficient for reasonable segmentation. The Cohen's d = −0.475 (M4 vs M7) indicates a small-to-medium performance gap, suggesting that task-specific fine-tuning provides incremental rather than transformative benefit.
2) **Kidney (M8):** The dramatic failure of the frozen kidney encoder (Dice = 0.654, worse than baseline M2: 0.899) with Cohen's d = −0.737 (medium effect) reveals that kidney segmentation requires more precise, task-specific features than liver. We hypothesize this stems from anatomical differences: kidneys are smaller, bilaterally symmetric, and have more complex bean-shaped morphology with the renal hilum requiring fine-grained boundary precision. The diffusion-learned global structural priors are insufficient without supervised adaptation.
3) **Multi-Organ (M9):** The large effect size (Cohen's d = −0.897) between frozen and fine-tuned multi-organ models indicates that organ disambiguation; distinguishing liver from kidney despite similar tissue densities; requires supervised signal that pure generative pretraining cannot provide. However, M9 still outperforms the random baseline M3 on variance (0.249 vs 0.225), suggesting partial anatomical understanding.

In all the analyses conducted, diffusion pretraining invariably leads to variance reduction: liver 55.9%, kidney 47.3%, and multi-organ 68.4%. This variance reduction is significant from a clinical point of view. Consider an analysis with a mean Dice score of 0.95 and a standard deviation of 0.30. Such a system will definitely fail in 5% of the test cases, making it necessary to manually check each prediction done. On the other hand, M6 demonstrates a very narrow distribution (0.950 ± 0.071), and thus enables safe automated screening where only anomalies need human intervention. The reduction in variance indicates that diffusion-based features include organ invariant information.

4) Low-Data Performance

Organ-Specific Data Efficiency Patterns. The low-data experiments (M10–M15) reveal striking differences in how liver and kidney segmentation degrade under label scarcity. Table 5 isolates these models for focused analysis:

TABLE 5. Data Efficiency Analysis (Dice Coefficient Only).

Model	Organ	Labeled Data	Dice	% Performance
M4	Liver	100%	0.928±0.160	100.0%
M10	Liver	50%	0.922±0.173	99.4%
M11	Liver	25%	0.904±0.199	97.4%
M12	Liver	10%	0.886±0.193	95.5%
M5	Kidney	100%	0.950±0.101	100.0%
M13	Kidney	50%	0.941±0.124	99.1%
M14	Kidney	25%	0.814±0.341	85.7%
M15	Kidney	10%	0.712±0.363	75.0%

Liver segmentation shows impressive data efficiency, maintaining 99.4% of its full performance even when it only gets half of the labeled data. It can still do well with just 10% of the labels (approximately 3–4 annotated patients). Looking at the results in Table 4, with Cohen's d between −0.071 to −0.226, we see that performance dips gradually. Hence the features learned through diffusion work great for livers and are very adaptable to similar tasks.

For kidneys, once the dataset drops below 50%, performance takes a nosedive. With 25% of the data, it only keeps 85.7% of its ability, dropping further to 75.0% at 10%. As the data shrink, The Cohen's d progression (−0.071 to −0.436 to −0.686) indicates accelerating degradation. In addition, the larger variance seen at lower data points indicates some serious unpredictability (M15: 0.363 standard deviation vs M5: 0.101), suggesting potential problems with prediction accuracy. This extra vulnerability for kidney segmentation could be due to the kidneys being more complex. Their bilateral symmetry, changing positions, and detailed internal structures probably need far more training samples to obtain reliable results.

Liver segmentation achieve a Dice score over 0.92 with only 50% labeled data, that means you only really need around 17 to 18 annotated CT scans rather than 35. Factoring in annotation costs of $100-$150 per volume and around 45-60 minutes of a radiologist's time, going with less labeled data saves roughly 50% on cash ($3,500 vs $1,750) and time (26 hours versus 13 hours), all while barely giving up any performance accuracy.

C. QUALITATIVE VISUAL ANALYSIS

Fig. 4 presents segmentation predictions from all 15 models on three representative test cases selected to demonstrate: (1) typical anatomy with clear boundaries, (2) challenging anatomy with irregular liver morphology, and (3) a case with adjacent liver-kidney boundaries prone to organ confusion.

D. INTERPRETABILITY ANALYSIS VIA GRAD-CAM

To elucidate the mechanistic basis for diffusion pretraining's performance improvements, we applied Mask-Guided Gradient-weighted Class Activation Mapping (Grad-CAM) to the bottleneck layer of all models. Standard Grad-CAM often fails in dense prediction tasks due to background dominance; therefore, gradients were backpropagated explicitly through ground-truth masks to isolate attention maps specifi-

cally responsible for predicting target organs. Fig. 5 presents these attention visualizations across all 15 models on the same three test cases as Fig. 4.

One interesting phenomenon observed in the attention maps of the diffusion-pretrained network was the presence of negative encoding or a stencil effect at the bottleneck. Unlike baseline models that predominantly attend to internal organ textures, the diffusion pretrained encoder exhibited high activation in the surrounding anatomical context, while suppressing activations within the organ itself. This strongly suggests that the diffusion pretraining shifts the network's delineation strategy from local texture-matching to global contextual boundary-delineation. It identifies an organ by confidently mapping the diverse anatomical structures that surround it.

Grad-CAM analysis supports the main idea that diffusion pretraining adds global anatomical priors to encoder networks, changing how features are learned. Moving from scattered baseline attention to confined diffusion attention means DDPM pretraining makes the network understand organ placement along with their looks. This spatial info acts as a strong regularizer during fine-tuning. Thus, it prevents overfitting to local textures and lets the model make sensible, anatomy-based predictions.

The organ-specific differences in frozen encoder performance (liver: 86.9% retention, kidney: catastrophic failure) suggest that anatomical complexity and organ size influence the quality of diffusion-learned priors. Large, spatially extended organs with relatively simple topology (liver: single irregular polyhedron) may be easier to learn unsupervised than small, bilaterally symmetric organs with complex morphology (kidneys: paired bean-shaped structures with intricate hilum). This insight may guide future work on selective pretraining strategies or hybrid approaches where different organ types receive customized pretraining.

V. DISCUSSION

This study proves that unsupervised pre-training using diffusion has the capacity to change the way medical image segmentation works by introducing anatomical information into encoders before performing any supervised optimization. These results support the core premise of this work: generative models can serve as unsupervised learners of anatomical properties used in later discriminative tasks when they have been trained for image reconstruction.

A. PRINCIPAL FINDINGS AND MECHANISTIC INTERPRETATION

The consistent improvements achieved in Dice score, IoU, HD95, and ASD metrics show that diffusion pre-training helps overcome the main problems faced by randomly initialized U-Nets. In supervised training, the goal is focused on achieving high pixel-wise accuracy, and, therefore, the network tries to match local textures while providing little consideration of the wider anatomical context. This approach is clearly visible from the distribution of the Grad-CAM maps in the baseline models.

But diffusion pretraining shifts this around. The goal here is predicting and eliminating Gaussian noise across a thousand steps. That doesn't work just by matching local textures. To handle a messed-up liver image at step 500, say, the net needs to understand that livers have certain shapes and typical spots where they are anatomically located; and what intensities those spots should have. As a result, the encoder learns various structural bits: simple details like texture and edges; more complex things like liver lobes and vessels; and, finally, the complete anatomical setup.

When transferred to segmentation, these pretrained features provide powerful priors. The inverse bounded Grad-CAM attention patterns in M4–M6 reveal that diffusion encoders inherently suppress activations in anatomically implausible regions; not because supervision taught them to, but because the denoising objective required learning spatial constraints. This explains the disproportionate improvement in boundary metrics (HD95: -44.7% to -68.3%, ASD: -65.5% to -74.3%) relative to Dice: diffusion features encode smooth, anatomically coherent boundaries as fundamental structural units, whereas baseline models must infer boundaries indirectly from pixel-wise classification gradients.

B. INFERENCES FROM FROZEN ENCODER EXPERIMENT

The empirical results with the frozen encoder represent the most convincing evidence. M7 reaches 86.9% of M4 performance in spite of the absence of segmentation labels which shows that diffusion pre-training captures some real anatomical structures and not just general image features which are applicable to different tasks.

Differences in performance of the frozen encoders designed for particular organs (positive results in case of liver and negative in case of kidney) represent some interesting details. It is hypothesized that the reasons behind the differences have something to do with structural complexities:

Properties of the liver which make it easy to learn in an unsupervised way:

- Large spatial footprint (spans 10–15 axial slices)
- Unique anatomical position (only organ in right upper quadrant)
- Relatively simple topology (single connected component, though irregular)
- High visual distinctiveness (adjacent to low-density lung, fat)

Properties of the kidney which make unsupervised learning difficult:

- Small spatial footprint (5–7 axial slices)
- Bilateral symmetry requiring paired representation learning
- Complex morphology (bean-shaped with concave hilum, intricate vasculature)
- Variable positioning (can shift 2–3 cm with respiration)

- Lower visual contrast (surrounded by psoas muscle, perirenal fat with similar densities)

It shows that pre-training via unsupervised diffusion is particularly beneficial for large organs with unique anatomy, while the smaller organs or organs with different anatomy gain from being subjected to supervised training. Future research should target organ-specific pre-training methods or hybrid methods where diffusion-based features are combined with specific supervised information. Increasing the complexity of the model and training it using labels for the whole abdomen might also bring about some improvement.

C. ANALYSIS OF VARIANCE REDUCTION

While mean performance improvements are scientifically important, the 47–68% variance reduction is clinically more significant. In deployment scenarios compared to peak performance, metrics like consistency across diverse anatomies, imaging protocols, and patient populations often matters more. A model achieving a Dice of 0.95 ± 0.30 requires manual verification of every prediction to catch the 5% catastrophic failures; one achieving 0.95 ± 0.07 enables automated screening where only outliers trigger human review.

The variance collapse in diffusion-pretrained models suggests they encode anatomical invariants robust to patient-specific variations. We hypothesize three mechanisms:

1) **Distributional priors suppress outliers:** Diffusion models learn the data distribution during pretraining, hence when fine-tuning on segmentation the encoders weight represent a degree of distribution knowledge. Hence, predictions deviating from typical anatomy receive lower confidence.
2) **Global structural constraints prevent local overfitting:** As seen from the Grad-CAM analysis, diffusion used a stencil effect, which intern learnt the global anatomical relationships. It prevents diffusion from learning spurious correlations in limited training data, therefore a prediction is penalized not only for pixel-wise loss but also for violating learned spatial priors.
3) **Multi-scale feature hierarchies enable graceful degradation:** Diffusion encoders learn features at multiple spatial scales (due to the multi-timestep denoising objective). When encountering unusual anatomy, the model can fall back to coarse-scale features (approximate organ position) even if fine-scale features (precise boundary texture) prove unreliable. Baseline models lack this hierarchical fallback mechanism.

D. ANALYSIS OF DATA EFFICIENCY

The low-data experiments reveal that diffusion pretraining's value compounds when labeled data is scarce. Maintaining 99.4% liver performance with 50% labels and 95.5% with only 10% labels demonstrates that learned anatomical priors can partially substitute for supervised examples. In a standard learning paradigm a model must learn everything from labels, to what organs look like, where they are located and how they vary across patients. In low data regimes, the performance of traditional models collapse, variance explores and predictions become clinically useless.

Diffusion pretraining breaks this dependency, where the encoder is initialized with rich anatomical priors such as organ shape, location and variability. The limited segmentation labels need only refine these priors for task-specific discrimination, rather than learning anatomical structure from scratch. This explains the small Cohen's d effect sizes for liver data efficiency ($d = -0.071$ to -0.226). Performance degradation is minimal because the hard work of anatomical learning occurred during unsupervised pretraining.

The organ-specific degradation patterns (liver: graceful, kidney: catastrophic) likely reflect anatomical complexity interacting with data efficiency. Liver's large spatial extent and distinctive position may enable robust unsupervised learning, allowing fine-tuning with minimal supervision. Kidney's small size, bilateral symmetry, and morphological complexity may require more extensive supervised signals to anchor the diffusion priors to task-specific features. This suggests data efficiency is not uniform across anatomy. Organs with favorable structural characteristics may enable extreme label reduction, while complex structures still benefit substantially from richer supervision.

VI. LIMITATIONS

Evaluation focused on liver and kidney which are two anatomically favorable organs (large, well-contrasted). Generalization to smaller structures (pancreas, adrenal glands), pathological tissues (tumors with irregular boundaries), or organs with lower contrast (spleen, bowel) requires validation. The failure of the frozen encoder for kidney suggest that organ-specific characteristics do influence diffusion pretraining effectiveness.

Furthermore, all experiments used BTCV dataset sourced from a single institution. Conducting cross-dataset validation, which involves different scanners, protocols and patient demographics, is crucial for evaluating true generalization beyond the training distribution. The observed reduction in variance may partially reflect the memorization of dataset-specific characteristics rather than pure anatomical learning.

Our slice-based approach compromises the volumetric context. Although the segmentation of individual axial slices is performed with precision, the model is unable to enforce 3D anatomical constraints. An extension to 3D diffusion models can capture the volumetric relationships, thereby enhancing robustness and allowing for accurate 3D boundary metrics.

While Grad-CAM visualizations offer interpretable attention maps, they are approximate explanations, illustrating which areas the network considers significant. The bounded attention patterns observed in diffusion models may either represent genuine anatomical priors or suggest the presence of learned shortcuts.

Future research directions include extending to 3D volumetric diffusion, segmentation of pathological tissues, multi-

modal pretraining, and investigations into cross-dataset generalization.

VII. CONCLUSION

This study illustrates that unsupervised diffusion-based pretraining effectively integrates robust and interpretable anatomical features into encoder networks, fundamentally altering the approach to medical image segmentation from a focus on texture-matching to anatomy-aware prediction. The evidence supporting this transformation is both multifaceted and convergent:

Quantitative: There are statistically significant enhancements across all metrics (Dice: +5.7% to +23.6%, $p < 10^{-10}$; boundary precision: -38% to -74%), accompanied by medium-to-large effect sizes and a notable reduction in variance (47-68

Qualitative: Visual assessments indicate that diffusion models produce smooth, anatomically plausible boundaries while preserving global structural coherence. This eliminates the scattered errors typically associated with baseline models.

Mechanistic: Grad-CAM visualizations reveal that diffusion encoders develop bounded, organ-specific attention patterns even when in a frozen state. This thereby demonstrating that unsupervised anatomical learning takes place during the pretraining phase.

Practical: The remarkable data efficiency (achieving 95.5% liver performance with only 10% of labels) addresses significant challenges to clinical implementation in resource-limited environments.

Achieving 87% of fine-tuned liver performance without utilizing labels serves as compelling evidence that the generation-as-pretext-task is effective. The denoising objective forces networks to internalize anatomical structures as a prerequisite for successful reconstruction. When these learned priors are applied to segmentation tasks, they provide substantial regularization, mitigating the risk of overfitting to misleading texture correlations and enables robust generalization across diverse anomalies.

In addition to its direct clinical applications, this works establishes a broader principle: generative models trained on unsupervised reconstruction learn structural representations that are applicable to discriminative dense prediction tasks. As diffusion models progress, their role as unsupervised feature learners for spatial understanding tasks in medical imaging and potentially other domains deserve increasing attention from both research and clinical communities.

ACKNOWLEDGMENT

The authors would like to acknowledge Manipal Institute of Technology Bengaluru, Manipal Academy of Higher Education, Manipal, India, for providing the necessary facilities and computational support to conduct this research. We express our sincere gratitude to the creators of the Beyond the Cranial Vault (BTCV) dataset [59] for making their data publicly accessible. Additionally, we would like to thank the developers of PyTorch, CUDA, and other open-source packages that enabled our model development and experiments.

REFERENCES

[1] Y. Deng, Y. Sun, Y. Zhu, Y. Xu, Q. Yang, S. Zhang, Z. Wang, J. Sun, W. Zhao, X. Zhou, and K. Yuan, "A new framework to reduce doctor's workload for medical image annotation," *IEEE Access*, vol. 7, pp. 107 097–107 104, 2019.

[2] A. Lawley, R. Hampson, K. Worrall, and G. Dobie, "A cost focused framework for optimizing collection and annotation of ultrasound datasets," *Biomedical Signal Processing and Control*, vol. 92, p. 106048, 2024.

[3] C. Qu, T. Zhang, H. Qiao, Y. Tang, A. Yuille, and Z. Zhou, "AbdomenAtlas-8K: Annotating 8,000 CT volumes for multi-organ segmentation in three weeks," Preprint, 2023.

[4] A. Schmidt, P. Morales-Álvarez, and R. Molina, "Probabilistic modeling of inter- and intra-observer variability in medical image segmentation," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 21 040–21 049.

[5] F. Yang, G. Zamzmi, S. Angara, S. Rajaraman, A. Aquilina, Z. Xue, S. Jaeger, E. Papagiannakis, and S. Antani, "Assessing inter-annotator agreement for medical image segmentation," *IEEE Access*, vol. 11, pp. 21 300–21 312, 2023.

[6] P. Roshanzamir, H. Rivaz, J. Ahn, H. Mirza, N. Naghdi, M. Anstruther, M. Battié, M. Fortin, and Y. Xiao, "How inter-rater variability relates to aleatoric and epistemic uncertainty: a case study with deep learning-based paraspinal muscle segmentation," arXiv:2308.06964, pp. 74–83, 2023.

[7] M. Riera-Marín *et al.*, "Calibration and uncertainty for multirater volume assessment in multiorgan segmentation (CURVAS) challenge results," *Computers in Biology and Medicine*, vol. 197, p. 111024, 2025.

[8] 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 (MICCAI)*. Cham: Springer International Publishing, 2015, pp. 234–241.

[9] R. Azad, E. Aghdam, A. Rauland, Y. Jia, A. Avval, A. Bozorgpour, S. Karimijafarbigloo, J. Cohen, E. Adeli, and D. Merhof, "Medical image segmentation review: The success of U-Net," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, pp. 10 076–10 095, 2022.

[10] N. Ibtehaz and M. Rahman, "MultiResUNet: Rethinking the U-Net architecture for multimodal biomedical image segmentation," *Neural Networks*, vol. 121, pp. 74–87, 2019.

[11] T. Hussain and H. Shouno, "MAGRes-UNet: Improved medical image segmentation through a deep learning paradigm of multi-attention gated residual U-Net," *IEEE Access*, vol. 12, pp. 40 290–40 310, 2024.

[12] J. Chen, J. Mei, X. Li, Y. Lu, Q. Yu, Q. Wei, X. Luo, Y. Xie, E. Adeli, Y. Wang, M. Lungren, S. Zhang, L. Xing, L. Lu, A. Yuille, and Y. Zhou, "TransUNet: Rethinking the U-Net architecture design for medical image segmentation through the lens of transformers," *Medical Image Analysis*, vol. 97, p. 103280, 2024.

[13] Z. Zhou, M. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE Transactions on Medical Imaging*, vol. 39, pp. 1856–1867, 2019.

[14] L. Dai, M. Johar, and M. Alkawaz, "Review of semi-supervised medical image segmentation based on the U-Net," *Academic Journal of Science and Technology*, 2024.

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

[16] F.-A. Croitoru, V. Hondru, R. T. Ionescu, and M. Shah, "Diffusion models in vision: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, pp. 10 850–10 869, 2022.

[17] A. Kazerouni, E. K. Aghdam, M. Heidari, R. Azad, I. Hacihaliloglu, and D. Merhof, "Diffusion models in medical imaging: A comprehensive survey," *Medical image analysis*, vol. 88, p. 102846, 2022.

[18] Y. Huang, J. Huang, Y. Liu, M. Yan, J. Lv, J. Liu, W. Xiong, H. Zhang, S. Chen, and L. Cao, "Diffusion model-based image editing: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 47, pp. 4409–4437, 2024.

[19] C. F. Higham, D. J. Higham, and P. Grindrod, "Diffusion models for generative artificial intelligence: An introduction for applied mathematicians," *SIAM Rev.*, vol. 67, pp. 607–623, 2025.

[20] H. Chung and J.-C. Ye, "Score-based diffusion models for accelerated mri," *Medical image analysis*, vol. 80, p. 102479, 2021.

[21] V. Purma, S. Srinath, S. Srirangarajan, A. Kakkar, and A. Prathosh, "Genselfdiff-his: generative self-supervision using diffusion for histopathological image segmentation," *IEEE Transactions on Medical Imaging*, vol. 44, no. 2, pp. 618–631, 2024.

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

[23] P. Jaccard, "Étude comparative de la distribution florale dans une portion des Alpes et des Jura," *Bulletin de la Société Vaudoise des Sciences Naturelles*, vol. 37, pp. 547–579, 1901.

[24] F. Hausdorff, *Grundzüge der Mengenlehre*. Leipzig: Veit & Comp., 1914.

[25] T. Heimann and H.-P. Meinzer, "Statistical shape models for 3D medical image segmentation: A review," *Medical Image Analysis*, vol. 13, no. 4, pp. 543–563, 2009.

[26] R. A. Fisher, *Statistical Methods for Research Workers*. Edinburgh: Oliver and Boyd, 1925.

[27] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. Lawrence Erlbaum Associates, 1988.

[28] Z. Guo, X. Li, H. Huang, N. Guo, and Q. Li, "Deep learning-based image segmentation on multimodal medical imaging," *IEEE Transactions on Radiation and Plasma Medical Sciences*, vol. 3, pp. 162–169, 2019.

[29] R. Gu, G. Wang, T. Song, R. Huang, M. Aertsen, J. Deprest, S. Ourselin, T. Vercauteren, and S. Zhang, "Ca-net: Comprehensive attention convolutional neural networks for explainable medical image segmentation," *IEEE Transactions on Medical Imaging*, vol. 40, pp. 699–711, 2020.

[30] H. Roth, C. Shen, H. Oda, M. Oda, Y. Hayashi, K. Misawa, and K. Mori, "Deep learning and its application to medical image segmentation," *ArXiv*, vol. abs/1803.08691, 2018.

[31] M. A. Abdou, "Literature review: efficient deep neural networks techniques for medical image analysis," *Neural Computing and Applications*, vol. 34, pp. 5791 – 5812, 2022.

[32] Y. Xu, R. Quan, W. Xu, Y. Huang, X. Chen, and F. Liu, "Advances in medical image segmentation: A comprehensive review of traditional, deep learning and hybrid approaches," *Bioengineering*, vol. 11, 2024.

[33] F. Milletari, N. Navab, and S. A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 Fourth International Conference on 3D Vision (3DV)*. IEEE, 2016, pp. 565–571.

[34] N. Siddique, S. Paheding, C. Elkin, and V. Devabhaktuni, "U-Net and its variants for medical image segmentation: A review of theory and applications," *IEEE Access*, vol. 9, pp. 82 031–82 057, 2020.

[35] F. Isensee, P. Jaeger, S. Kohl, J. Petersen, and K. Maier-Hein, "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature Methods*, vol. 18, pp. 203–211, 2020.

[36] A. Lin, B. Chen, J. Xu, Z. Zhang, G. Lu, and D. Zhang, "DS-TransUNet: Dual swin transformer U-Net for medical image segmentation," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–15, 2021.

[37] A. Shaker, M. Maaz, H. Rasheed, S. Khan, M. Yang, and F. Khan, "UNETR++: Delving into efficient and accurate 3D medical image segmentation," *IEEE Transactions on Medical Imaging*, vol. 43, pp. 3377–3390, 2022.

[38] L. Alzubaidi, J. Bai, A. Al-Sabaawi, J. I. Santamaría, A. Albahri, B. S. Al-dabbagh, M. Fadhel, M. Manoufali, J. Zhang, A. H. Al-timemy, Y. Duan, A. Abdullah, L. Farhan, Y. Lu, A. Gupta, F. Albu, A. Abbosh, and Y. Gu, "A survey on deep learning tools dealing with data scarcity: definitions, challenges, solutions, tips, and applications," *Journal of Big Data*, vol. 10, pp. 1–82, 2023.

[39] M. A. Bansal, D. R. Sharma, and D. M. Kathuria, "A systematic review on data scarcity problem in deep learning: Solution and applications," *ACM Computing Surveys (CSUR)*, vol. 54, pp. 1 – 29, 2022.

[40] S. E. Whang and J.-G. Lee, "Data collection and quality challenges for deep learning," *Proceedings of the VLDB Endowment*, vol. 13, pp. 3429 – 3432, 2020.

[41] C. Janiesch, P. Zschech, and K. Heinrich, "Machine learning and deep learning," *Electronic Markets*, vol. 31, pp. 685 – 695, 2021.

[42] J. Gui, T. Chen, J. Zhang, Q. Cao, Z. Sun, H. Luo, and D. Tao, "A survey on self-supervised learning: Algorithms, applications, and future trends," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 12, pp. 9052–9071, 2024.

[43] "Dive into the details of self-supervised learning for medical image analysis," *Medical Image Analysis*, 2023.

[44] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2020. [Online]. Available: https://arxiv.org/abs/2002.05709

[45] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. [Online]. Available: https://arxiv.org/abs/1911.05722

[46] L. Jing and Y. Tian, "Self-supervised visual feature learning with deep neural networks: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. [Online]. Available: https://arxiv.org/abs/1902.06162

[47] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[48] X. Chen, L. Yao, and Y. Zhang, "Residual attention U-Net for automated multi-class segmentation of COVID-19 chest CT images," Preprint, 2019.

[49] 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 (CVPR)*, 2022. [Online]. Available: https://arxiv.org/abs/2111.06377

[50] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of NAACL-HLT*, 2019. [Online]. Available: https://arxiv.org/abs/1810.04805

[51] X. Chen, H. Fan, R. Girshick, and K. He, "Improved baselines with momentum contrastive learning," arXiv:2003.04297, 2020. [Online]. Available: https://arxiv.org/abs/2003.04297

[52] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proceedings of the 25th International Conference on Machine Learning (ICML)*, 2008.

[53] G. Müller-Franzes, J. Niehues, F. Khader, S. T. Arasteh, C. Haarburger, C. Kuhl, T. Wang, T. Han, S. Nebelung, J. N. Kather, and D. Truhn, "A multimodal comparison of latent denoising diffusion probabilistic models and generative adversarial networks for medical image synthesis," *Scientific Reports*, vol. 13, 2022.

[54] K. Gong, K. A. Johnson, G. E. Fakhri, Q. Li, and T. Pan, "Pet image denoising based on denoising diffusion probabilistic model," *European Journal of Nuclear Medicine and Molecular Imaging*, vol. 51, pp. 358–368, 2022.

[55] W. Wang, J. Bao, W. gang Zhou, D. Chen, D. Chen, L. Yuan, and H. Li, "Semantic image synthesis via diffusion models," *ArXiv*, vol. abs/2207.00050, 2022.

[56] D. Baranchuk, I. Rubachev, A. Voynov, V. Khrulkov, and A. Babenko, "Label-efficient semantic segmentation with diffusion models," arXiv:2112.03126, 2021.

[57] T. Amit, T. Shaharbany, and L. Wolf, "SegDiff: Image segmentation with diffusion probabilistic models," arXiv:2112.00390, 2021.

[58] J. Wu, H. Fu, J. Xu, and Y. Zhang, "MedSegDiff: Medical image segmentation with diffusion probabilistic models," arXiv:2301.11798, 2023.

[59] E. Gibson, F. Giganti, Y. Hu, E. Bonmati, S. Bandula, K. Gurusamy, B. Davidson, S. P. Pereira, M. J. Clarkson, and D. C. Barratt, "Multi-organ abdominal ct reference standard segmentations," 2018. [Online]. Available: https://doi.org/10.5281/zenodo.1169361

[60] S. Pan, C.-W. Chang, T. Wang, J. Wynne, M. Hu, Y. Lei, T. Liu, P. Patel, J. Roper, and X. Yang, "Abdomen CT multi-organ segmentation using token-based MLP-Mixer," *Medical Physics*, vol. 50, no. 5, pp. 3027–3038, 2023.

[61] J. T. Bushberg, J. A. Seibert, E. M. Leidholdt, and J. M. Boone, *The Essential Physics of Medical Imaging*, 3rd ed. Lippincott Williams & Wilkins, 2011.

[62] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. W. M. van der Laak, B. van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017.

[63] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.

[64] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

DECLARATIONS

A. AUTHOR CONTRIBUTIONS STATEMENT

A. G. conceived the study, designed the methodology, implemented the DDPM architecture and transfer learning, performed experiments, analyzed the results, prepared the figures, and wrote the initial manuscript draft. D. G. aided in tabulating the conducted experiments results, analyzed the results, reviewed the figures and tables, and aided in the initial manuscript draft. S. B., S. A., and T. K. M. supervised the research, provided conceptual guidance and contributed to the interpretation of results, reviewed and edited the manuscript, and approved the final version.

B. ETHICS STATEMENT

This article does not contain any studies with human participants or animals performed by any of the authors.

ADDITIONAL INFORMATION

- **Funding:** Open access funding is provided by Manipal Academy of Higher Education.
- **Conflicts of Interest:** The authors declare that they have no competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.
- **Data Availability:** The experimental evaluations were performed using the Beyond the Cranial Vault (BTCV) multi-organ abdominal CT dataset, which is publicly accessible via the official Synapse repository at https://www.synapse.org/#!Synapse:syn3193805.

- **ORCID iDs:** Akshat G: https://orcid.org/0009-0000-9244-1513.

AKSHAT G is currently pursuing the B.Tech Honors degree in computer science (minor in A.I. for Healthcare) with Manipal Institute of Technology Bengaluru, India. He has worked as an ML Research Intern, developing hierarchical deep learning architectures for multi label defect classification, and build end to end consumer products for his startup. His research interests include semi-supervised learning, self-supervised learning, data-centric AI, computer vision, deep learning, foundational AI and focus on developing scalable AI applications for real world challenges.

DIVYANSH GUPTA is currently pursuing a third-year B.Tech. degree in Computer Science and Engineering at the Manipal Institute of Technology, Bengaluru, Manipal Academy of Higher Education (MAHE), Manipal, India. He is passionate about artificial intelligence and focuses on creating scalable machine learning solutions for real-world challenges. His areas of interest include machine learning and deep learning, particularly in the development of robust and application-oriented models.

SHALEEN BHATNAGAR is an Assistant Professor (Senior Scale) with the School of Computer Engineering, Manipal Institute of Technology, Bengaluru, Manipal Academy of Higher Education (MAHE), Manipal, India. She has more than 14 years of experience in teaching and research. Her research focuses on biometrics, computer vision, pattern recognition, and image processing, with recent emphasis on machine learning, deep learning, and quantum computing for real-world applications. She has authored and co-authored several research papers in these areas. She is a Senior Member of IEEE and actively contributes to academic and professional activities within the computing research community.

SHILPA ANKALAKI received the Ph.D. degree in computer science and engineering from Visveswaraya Technological University, Belagavi, India. She is currently an Assistant Professor (Senior Scale) with the School of Computer Engineering, Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal, Bengaluru. She has authored several research articles published in various international journals and conferences. Her research interests include machine learning, deep learning, data mining, artificial intelligence, and computer vision.

TUSAR KANTI MISHRA (Senior Member, IEEE) received the B.Tech. and M.Tech. degrees, and the Ph.D. degree from NIT Rourkela, in 2015, along with an MHRD Fellowship. He is currently an Associate Professor with the School of Computer Engineering, Manipal Institute of Technology Bengaluru. He has more than 20 years of teaching experience. His research expertise is of ten years in the domain of artificial intelligence, computer vision, pattern recognition, along with allied application areas. He has availed the Erasmus Fellowship and worked under Prof. L. R. B. Schomaker, the Director of the ALICE Laboratory, The Netherlands, for a period of six months. He feels proud of being a student of renowned researchers, such as Dr. B. Majhi and Dr. Arun K. Pujari. He has been a reviewer for multiple reputed journals. So far, he has filed multiple patents and two copyrights with the IPR, Government of India. He has more than 35 research publications and counting. He has bagged a research funding grant of INR 30 lakhs from SERB, Government of India. He is also a member of the BoS, Sambalpur University Institute of Information Technology