StructDiffSeg: A Structure-Aware Diffusion Framework for Anatomically Consistent Medical Image Segmentation

Xin Yu, Yehang Li, Hui Zhao*
School of Computer Science and Technology
Xinjiang University
Urumqi, China
1684541675@qq.com, 107556525382@stu.xju.edu.cn, zhaohui@xju.edu.cn

***Abstract*—Diffusion-based segmentation models have shown strong capability in refining object boundaries through stochastic denoising. However, they often produce fragmented or structurally inconsistent predictions in anatomically constrained scenarios, particularly for multi-organ segmentation. In this work, we propose StructDiffSeg, a structure-aware diffusion framework that incorporates learnable structural priors into the diffusion refinement process to enhance anatomical consistency.**

StructDiffSeg comprises a deterministic base segmentation branch that maintains class-wise structural priors, and a conditional diffusion branch that iteratively refines noisy segmentation masks under both image guidance and structural conditioning. A Class-wise Structural Prior Interaction (CSPI) module captures category-specific structural information across encoder scales, while a Structure-Aware Affinity Fusion (SAAF) mechanism aligns diffusion features with structural representations at the bottleneck stage to stabilize denoising trajectories.

Unlike purely appearance-driven or boundary-focused refinement strategies, StructDiffSeg emphasizes global structural coherence while preserving fine-grained segmentation accuracy. Extensive experiments on the Synapse multi-organ and ACDC cardiac datasets demonstrate that StructDiffSeg consistently outperforms diffusion-based and shape-prior-based baselines in terms of Dice score and Hausdorff distance (HD95). Qualitative results further show that StructDiffSeg produces anatomically plausible segmentations with reduced fragmentation and structural artifacts.

***Index Terms*—Diffusion models, medical image segmentation, structural priors, structure-aware fusion, deep learning**

I. Introduction

Medical image segmentation is a critical component of computer-aided diagnosis and treatment planning, where accurate delineation of anatomical structures directly impacts clinical decision-making. Over the past decade, discriminative deep learning models have rapidly evolved, from fully convolutional architectures such as U-Net [1] to Transformer-based models [2] and recent state-space formulations [3]. While these architectures improve feature representation and global context modeling, they typically formulate segmentation as a deterministic pixel-wise classification task. Consequently, they lack explicit mechanisms for modeling aleatoric uncertainty and may produce overconfident predictions in regions with low contrast or ambiguous boundaries.

To address these limitations, denoising diffusion probabilistic models [4], [5] have recently been applied to medical image segmentation. Representative works such as MedSegDiff [6] and Diff-UNet [7] reformulate segmentation as a conditional generative process, where segmentation masks are progressively recovered from Gaussian noise through iterative denoising. This approach naturally captures pixel-wise uncertainty and enables refinement in challenging scenarios. However, without sufficient structural guidance, diffusion-based segmentation can produce fragmented or anatomically inconsistent predictions, a phenomenon often referred to as structural hallucination.

Existing strategies to mitigate this issue typically rely on implicit guidance, such as image-conditioned features, auxiliary boundary branches, or uncertainty-driven fusion [8]. While these methods improve local boundary accuracy, they provide only soft constraints and are insufficient to maintain global anatomical consistency. In contrast, prior studies on shape priors and structural constraints—such as anatomically constrained neural networks [9], structure-aware loss functions [10], and prototype-based shape modeling exemplified by SPM [11]—demonstrate that explicitly modeling anatomical structure is essential for reliable segmentation. However, these approaches are mostly designed for deterministic frameworks and cannot be directly integrated into iterative diffusion-based refinement.

Motivated by the need to reconcile stochastic refinement with anatomical consistency, we propose **StructDiffSeg**, a structure-aware diffusion framework for medical image segmentation. StructDiffSeg incorporates learnable class-wise structural priors into the conditional diffusion process, guiding the denoising trajectory toward anatomically plausible predictions. A Class-wise Structural Prior Interaction (CSPI) module captures category-specific structural information across encoder scales, while a Structure-Aware Affinity Fusion (SAAF)

*Corresponding author. This article is supported by the Key Research and Development Program of Xinjiang Uygur Autonomous Region (Project No. 2023B01032)

mechanism aligns diffusion features with structural representations at the bottleneck stage, stabilizing iterative refinement and reducing structural artifacts.

The main contributions of this work are summarized as follows:

- We propose a novel structure-aware diffusion framework, **StructDiffSeg**, that integrates learnable anatomical priors into diffusion-based segmentation for improved structural consistency.
- We design the CSPI and SAAF modules to provide multi-scale structural guidance and feature alignment, enhancing boundary refinement while preserving global anatomical coherence.
- Extensive experiments on cardiac and multi-organ segmentation benchmarks demonstrate that StructDiffSeg outperforms state-of-the-art discriminative and diffusion-based methods in terms of Dice score and Hausdorff distance (HD95), producing anatomically plausible segmentations with reduced fragmentation.

II. Related Work

A. Discriminative Medical Image Segmentation Networks

Deep learning-based medical image segmentation has been predominantly driven by discriminative models formulated as pixel-wise classification tasks. Early architectures such as U-Net [1] and its variants established the encoder–decoder paradigm with skip connections, achieving strong performance by effectively capturing local texture and boundary information. Subsequent extensions, including ResUNet [12] and nnU-Net [13], further improved robustness through architectural optimization and data-driven training strategies.

To address the limited receptive field of convolutional networks, Transformer-based models have been introduced into medical image segmentation. Representative works such as TransUNet [14] and Swin-UNet [15] leverage self-attention mechanisms to capture long-range dependencies and global contextual information. More recently, hybrid architectures and state-space models, exemplified by UNETR++ [16] and VM-UNet [17], aim to balance computational efficiency with global modeling capability.

Despite these advances, discriminative segmentation models remain inherently deterministic and are optimized with pixel-wise supervision. Such formulations lack explicit modeling of predictive uncertainty and often rely heavily on local appearance cues. As a result, they may produce overconfident predictions and exhibit structural inconsistencies in challenging regions, such as low-contrast boundaries or heterogeneous tissues.

B. Diffusion-Based Generative Segmentation Models

Denoising diffusion probabilistic models have recently emerged as a powerful generative framework and have been extended to medical image segmentation to address the limitations of deterministic approaches. MedSegDiff [6] pioneered this direction by reformulating segmentation as a conditional diffusion process, enabling iterative refinement of segmentation masks from noise and providing a natural mechanism for uncertainty modeling. Subsequent works, such as MedSegDiff-V2 [18], further improved stability by incorporating anchor-based conditioning and frequency-domain constraints.

Diff-UNet [7] proposed a hybrid architecture that integrates diffusion models with U-Net-style backbones and introduces uncertainty-guided feature fusion, demonstrating improved robustness in volumetric segmentation tasks. Other studies explored alternative formulations tailored to segmentation, including Bernoulli diffusion for binary masks [19] and hybrid discriminative–generative frameworks such as HiDiff [20].

Although diffusion-based approaches provide a flexible framework for uncertainty-aware refinement and improved boundary delineation, their generative nature also introduces new challenges. In particular, the stochastic sampling process lacks explicit structural constraints, making it difficult to maintain anatomically consistent predictions throughout iterative denoising. Existing methods predominantly rely on appearance-driven conditioning or implicit guidance, which is insufficient to enforce global structural coherence. Consequently, diffusion-based segmentation models may generate fragmented regions or anatomically implausible structures, limiting their reliability in clinical applications.

C. Shape Priors and Structural Constraints

Incorporating anatomical priors has long been recognized as an effective strategy for improving the reliability of medical image segmentation. Classical approaches, including atlas-based methods [21] and statistical shape models [22], impose global constraints on segmentation outcomes but are often computationally expensive and sensitive to inter-patient variability.

With the advancement of deep learning, structural priors have been integrated into neural networks through structure-aware loss functions, constraint-based optimization, and auxiliary regularization strategies. Anatomically Constrained Neural Networks [9] introduced explicit anatomical constraints into segmentation pipelines, improving the plausibility of predicted structures. Constraint-based formulations [23] further enabled weakly supervised learning with structural regularization. More recently, prototype-based approaches, such as Shape Prior Modules [11], learn global anatomical representations through adaptive prototype modeling, demonstrating strong capability in enforcing structural consistency.

Despite their effectiveness, these approaches are primarily designed for deterministic, single-pass inference. As such, they lack mechanisms for iterative refinement and cannot naturally accommodate stochastic generative processes. This limitation makes it challenging to directly integrate these structural priors into diffusion-based segmentation frameworks.

D. Diffusion Models with Structural Guidance

Recent studies [24]–[26] have begun exploring the incorporation of structural cues into diffusion-based segmentation.

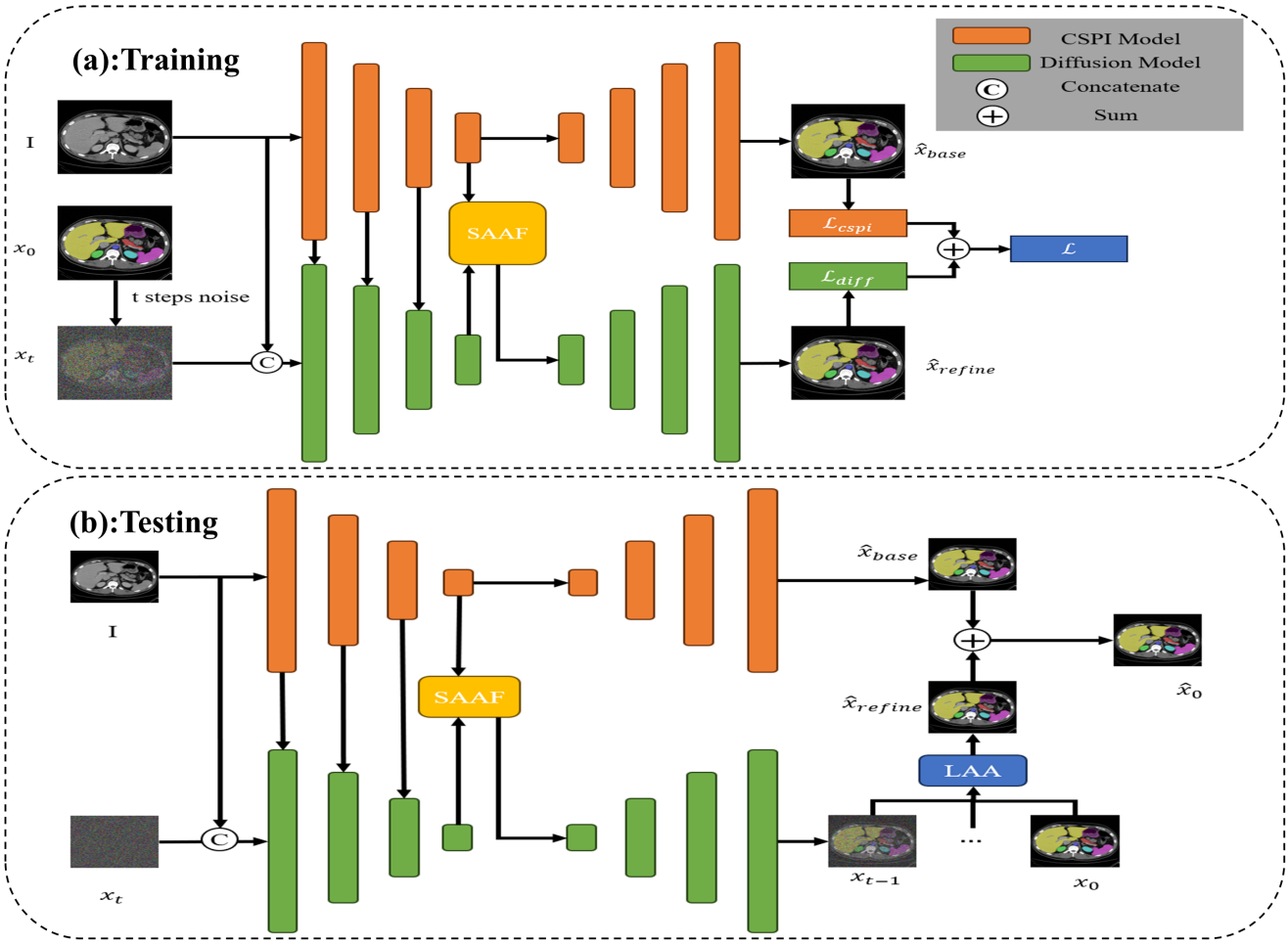

Fig. 1. Overview of the proposed **StructDiffSeg** framework. (a) Training phase: the CSPI branch produces a base segmentation and multi-scale structural embeddings, while the diffusion branch takes noisy masks as input and performs conditional denoising under image and structural guidance. The two branches interact in the encoder via shallow concatenation and bottleneck alignment using the Structure-Aware Affinity Fusion (SAAF) module. (b) Inference phase: the diffusion model iteratively refines predictions through DDIM sampling, and the Logit Accumulation and Averaging (LAA) module aggregates multi-step outputs. The final segmentation is obtained by combining the CSPI prediction and the refined diffusion output.

Existing approaches typically introduce boundary-aware constraints, feature conditioning, or uncertainty-guided refinement to stabilize the denoising process. While these strategies improve local accuracy and reduce noise-induced artifacts, they predominantly rely on implicit or appearance-driven guidance.

Such designs lack explicit modeling of global anatomical structure and do not provide a persistent structural representation throughout the diffusion trajectory. Consequently, structural information cannot be consistently propagated across denoising steps, limiting their ability to enforce anatomically coherent predictions.

To address these limitations, we propose **StructDiffSeg**, which explicitly integrates learnable class-wise structural priors into the diffusion process. By modeling structural representations across multiple feature scales and aligning them with diffusion features during denoising, the proposed framework enables consistent structural guidance throughout iterative refinement, thereby improving both segmentation accuracy and anatomical coherence.

III. Method

A. Overview

As shown in Fig. 1, StructDiffSeg is a structure-aware conditional diffusion framework for medical image segmentation. The model consists of two tightly coupled branches:

(i) a *base segmentation branch* (CSPI branch) that generates structurally stable multi-scale embeddings and an auxiliary segmentation prediction; and

(ii) a *conditional denoising refinement branch* (diffusion branch) that iteratively refines noisy segmentation masks under both image guidance and structural prior conditioning.

Structural priors produced by the CSPI branch provide class-wise anatomical information that guides the diffusion refinement process, enabling anatomically consistent updates throughout the iterative denoising. The two branches interact in the encoder: shallow stages adopt lightweight

concatenation-based fusion, while the bottleneck stage aligns features via a *Structure-Aware Affinity Fusion (SAAF)* module.

The final segmentation output is obtained by summing the refined diffusion logits from the *Logit Accumulation and Averaging (LAA)* module—which aggregates predictions across all S DDIM reverse sampling steps—with the auxiliary CSPI prediction. This design ensures both boundary refinement and global structural consistency.

B. Conditional Denoising Refinement Branch

We model the diffusion refinement as a conditional iterative denoising process in the segmentation space. Let the input image be $I \in \mathbb{R}^{C\times H\times W}$ and the ground-truth segmentation be $x_0 \in \{0,\dots,C-1\}^{H\times W}$. At forward diffusion step t, a noisy mask x_t is generated as

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, \quad \epsilon \sim \mathcal{N}(0, I), \tag{1}$$

where $\bar{\alpha}_t = \prod_{i=1}^{t}(1-\beta_i)$ and β_i is the noise schedule [4].

The conditional denoiser f_θ predicts the refined segmentation

$$\hat{x}_{t-1} = f_\theta(x_t, I, \mathcal{E}, t), \tag{2}$$

where $\mathcal{E}$ denotes the multi-scale structural embeddings from the CSPI branch, providing class-wise anatomical guidance. This corresponds to a structure-conditioned reverse process

$$p_\theta(x_{t-1}|x_t, I, \mathcal{E}). \tag{3}$$

To accelerate inference, we adopt DDIM-style sampling with S reverse steps. Let $\{\hat{x}^{(s)}\}_{s=1}^{S}$ denote the sequence of intermediate predictions across sampling steps s. The Logit Accumulation and Averaging (LAA) aggregates these predictions:

$$\hat{x}_{\text{refine}} = \frac{1}{S}\sum_{s=1}^{S}\hat{x}^{(s)}. \tag{4}$$

Finally, the refined diffusion output is fused with the CSPI prediction via residual summation:

$$\hat{x}_o = \hat{x}_{\text{base}} + \hat{x}_{\text{refine}}, \tag{5}$$

where $\hat{x}_{\text{base}}$ is the auxiliary CSPI prediction. The final segmentation map is obtained by taking $\arg\max$ over classes.

C. CSPI Branch and Multi-scale Conditioning with Encoder Interaction

The CSPI branch is a deterministic encoder-decoder segmentation network that generates two key outputs: (i) an auxiliary segmentation prediction $\hat{x}_{\text{base}}$, and (ii) a set of multi-scale structural embeddings $\mathcal{E} = \{e^{(i)}\}$, which serve as class-wise anatomical conditioning for the diffusion branch.

Its core design maintains a hierarchical class-wise structural prior $\mathcal{S} = \{S^{(i)}\}$ across encoder scales, refined through inter-class interaction and spatial feature feedback, providing stable structural guidance for iterative denoising.

At each encoder scale i, the CSPI branch performs:

1) **Class-wise Structural Context.** A class-specific latent prior $S^{(i)} \in \mathbb{R}^{C\times d}$ is maintained, where C is the number of classes and d is a scale-invariant embedding dimension. $S^{(i)}$ is either a learnable parameter or initialized from the previous scale.
2) **Class-wise Interaction via Self-Attention.** Multi-head Self-Attention (MSA) operates along the class dimension:
$$\tilde{S}^{(i)} = \text{MSA}(\text{LN}(S^{(i)})) + S^{(i)}, \tag{6}$$
where LN denotes Layer Normalization. This allows the prior to adaptively exchange information across classes.
3) **Spatial Injection.** Encoder feature $F^{(i)} \in \mathbb{R}^{C_i\times H_i\times W_i}$ is flattened along spatial dimensions: $F^{(i)}_{\text{flat}} \in \mathbb{R}^{C_i\times N}$, with $N = H_iW_i$. The refined prior $\tilde{S}^{(i)}$ is projected and interpolated to the spatial resolution, yielding $\tilde{S}^{(i)}_{\text{up}} \in \mathbb{R}^{C\times H_i\times W_i}$, then flattened to $\tilde{S}^{(i)}_{\text{flat}} \in \mathbb{R}^{C\times N}$.
Channel-to-class affinities are computed:
$$A^{(i)} = \text{Softmax}\Big(\frac{F^{(i)}_{\text{flat}}\cdot(\tilde{S}^{(i)}_{\text{flat}})^\top}{\sqrt{C}}\Big) \in \mathbb{R}^{C_i\times C}, \tag{7}$$
and structural context is injected:
$$\tilde{F}^{(i)} = F^{(i)} + A^{(i)}\cdot\tilde{S}^{(i)}_{\text{up}}. \tag{8}$$
4) **Residual Prior Refinement.** Spatially enhanced features $\tilde{F}^{(i)}$ are projected back to update the prior via a lightweight residual block $\psi^{(i)}_{\text{update}}(\cdot)$:
$$\hat{S}^{(i)} = \psi^{(i)}_{\text{update}}(\tilde{F}^{(i)}), \tag{9}$$
$$\Delta S^{(i)} = \text{Reshape}(\hat{S}^{(i)}), \tag{10}$$
$$S^{(i+1)} = \tilde{S}^{(i)} + \Delta S^{(i)}. \tag{11}$$

This bidirectional loop ensures that the prior guides spatial features while spatial evidence refines the prior.

D. Encoder Interaction: Shallow Fusion and Bottleneck SAAF

To incorporate multi-scale structural priors into the diffusion branch, we introduce two fusion mechanisms:

1) **Shallow Fusion.** At early encoder stages, the diffusion feature $x^{(i)}$ is concatenated with the CSPI embedding $e^{(i)}$ and projected via 1×1 convolution:
$$x^{(i)} \leftarrow \phi_i(\text{Concat}(x^{(i)}, e^{(i)})), \tag{12}$$
where $\phi_i(\cdot)$ denotes channel projection. This provides lightweight conditioning for shallow layers.
2) **Bottleneck Fusion via Structure-Aware Affinity Fusion (SAAF).** At the low-resolution bottleneck, simple concatenation may cause misalignment between the diffusion feature $x \in \mathbb{R}^{C\times h\times w}$ and the CSPI bottleneck embedding $e \in \mathbb{R}^{C\times h\times w}$. To explicitly propagate structural cues, we introduce the **SAAF module**, which aligns the diffusion feature with the CSPI feature through an affinity-based mechanism.Flattening the spatial dimensions ($N = h\cdot w$) gives $X, E \in \mathbb{R}^{C\times N}$. We compute channel-wise affinities and normalize via softmax:
$$\alpha = \|E\|_2^2 \in \mathbb{R}^{N\times 1}, \tag{13}$$

$$M = E^\top X \in \mathbb{R}^{N\times N}, \tag{14}$$

$$\mathrm{Aff} = \frac{2M - \alpha}{\sqrt{C}}, \tag{15}$$

$$A = \mathrm{Softmax}(\mathrm{Aff}) \in \mathbb{R}^{N\times N}, \tag{16}$$

$$\tilde{X} = EA \in \mathbb{R}^{C\times N}. \tag{17}$$

Finally, $\tilde{X}$ is reshaped back to $\tilde{x} \in \mathbb{R}^{C\times h\times w}$ and used as the bottleneck feature for the diffusion decoder. SAAF explicitly propagates class-wise structural information from CSPI to the diffusion bottleneck, stabilizing the iterative denoising process while maintaining practical computational cost.

E. Uncertainty-Rectified Learning

StructDiffSeg leverages diffusion stochasticity to estimate pixel-wise uncertainty and reweight the training loss. Given the diffusion prediction logits, we compute the pixel-wise uncertainty map $U \in [0,1]^{H\times W}$ as the Shannon entropy of the predicted class probabilities and apply a schedule-controlled reweighting:

$$w = 1 + s(e)\cdot U, \tag{18}$$

$$s(e) = 1 - \exp\Big(-10\big(\tfrac{e}{E}\big)^2\Big), \tag{19}$$

where e and E denote the current and total training epochs, respectively. The diffusion-branch loss is defined as:

$$\mathcal{L}_{\mathrm{diff}} = \lambda_{\mathrm{ce}}\cdot\mathbb{E}[w\odot\ell_{\mathrm{CE}}(\hat{x}_{\mathrm{refine}}, x_o)] + \lambda_{\mathrm{dice}}\cdot\ell_{\mathrm{Dice}}(\hat{x}_{\mathrm{refine}}, x_o) \tag{20}$$

The CSPI branch is supervised by the same CE+Dice loss without uncertainty weighting:

$$\mathcal{L}_{\mathrm{cspi}} = \lambda_{\mathrm{ce}}\cdot\ell_{\mathrm{CE}}(\hat{x}_{\mathrm{base}}, x_o) + \lambda_{\mathrm{dice}}\cdot\ell_{\mathrm{Dice}}(\hat{x}_{\mathrm{base}}, x_o) \tag{21}$$

The overall objective is:

$$\mathcal{L} = \mathcal{L}_{\mathrm{diff}} + \mathcal{L}_{\mathrm{cspi}} \tag{22}$$

IV. Experiments

A. Experimental Setup

Datasets. We evaluate the proposed **StructDiffSeg** framework on two widely used medical image segmentation benchmarks: ACDC and Synapse.

The *ACDC (Automated Cardiac Diagnosis Challenge)* dataset contains 100 cardiac MRI scans with annotations for three anatomically important structures: the left ventricle (LV), right ventricle (RV), and myocardium (Myo). The dataset exhibits large inter-subject anatomical variability and inter-slice inconsistency, making it suitable for evaluating robustness and anatomical consistency.

The *Synapse Multi-organ Segmentation* dataset consists of 30 abdominal CT scans with pixel-wise annotations for eight organs. Due to low soft-tissue contrast and complex spatial relationships among organs, Synapse presents significant challenges for maintaining anatomical coherence and accurate boundary delineation.

Evaluation metrics. Following common practice, segmentation performance is evaluated using the Dice similarity coefficient (DSC) and the 95% Hausdorff Distance (HD95). DSC measures the volumetric overlap between predictions and ground truth, while HD95 evaluates the boundary distance between segmentation results and reference annotations. Higher DSC and lower HD95 indicate better segmentation performance.

Implementation details. All experiments are implemented in PyTorch and conducted on a NVIDIA GeForce RTX4090 GPU. Input images are resized to 256×256.

The diffusion model is trained with $T = 1000$ forward diffusion steps under a standard Gaussian noise schedule. During inference, DDIM sampling is adopted to accelerate reverse diffusion, reducing the number of sampling steps to 10 while maintaining comparable performance.

For uncertainty-rectified learning (URL), the weighting factor is progressively increased during training, allowing the model to focus more on uncertain and hard-to-segment regions in later stages, thereby stabilizing the diffusion refinement process.

B. Quantitative Results

Results on Synapse. As reported in Table I, **StructDiffSeg** consistently outperforms both discriminative and diffusion-based baselines across most organs on the Synapse dataset. In particular, it achieves the highest average Dice score and the lowest average HD95 among all compared methods, demonstrating its effectiveness in improving both volumetric accuracy and boundary precision.

Compared with diffusion-based methods such as Diff-UNet and MedSegDiff-V2, StructDiffSeg exhibits more stable segmentation performance, especially on anatomically challenging organs. This improvement can be attributed to the incorporation of class-wise structural priors and structure-aware feature alignment, which guide the denoising process toward anatomically consistent predictions.

Notably, StructDiffSeg shows clear gains on organs with weak boundaries and complex morphology, such as the pancreas (PC) and spleen (SP). These results indicate that introducing explicit structural conditioning helps suppress fragmented regions and improves robustness in ambiguous areas, where purely appearance-driven diffusion models often struggle.

Results on ACDC. Quantitative comparisons on the ACDC dataset are summarized in Table II. **StructDiffSeg** achieves the highest average Dice score across the three cardiac structures (RV, Myo, and LV), outperforming both state-of-the-art discriminative models and recent diffusion-based approaches.

Compared with diffusion-based baselines such as Diff-UNet, StructDiffSeg yields more consistent improvements, particularly on the myocardium (Myo), which is known to exhibit complex geometry and ambiguous boundaries. This improvement demonstrates the effectiveness of incorporating

TABLE I

QUANTITATIVE COMPARISON ON THE SYNAPSE MULTI-ORGAN DATASET. THE BEST RESULTS ARE IN **BOLD**. DICE↑ IS HIGHER BETTER AND HD95↓ IS LOWER BETTER. ABBREVIATIONS: GB, GALLBLADDER; KL/KR, LEFT/RIGHT KIDNEY; PC, PANCREAS; SP, SPLEEN; SM, STOMACH.

Method	Avg. Dice↑	Avg. HD95↓	Aorta	GB	KL	KR	Liver	PC	SP	SM
U-Net [1]	76.85	39.70	89.07	**69.72**	77.77	68.60	93.43	53.98	86.67	75.58
Att-UNet [27]	77.77	36.02	**89.55**	68.88	77.98	71.11	93.57	58.04	87.30	**75.75**
ViT [2]	71.29	32.87	73.73	55.13	75.80	72.20	91.51	45.99	81.99	73.95
TransUnet [14]	77.48	31.69	87.23	63.13	81.87	77.02	94.08	55.86	85.08	75.62
DiffEnsemble [8]	74.22	30.11	86.78	51.94	**84.60**	76.32	92.60	46.07	86.74	68.69
MedSegDiff-V2 [18]	78.16	28.06	87.35	68.80	82.3	**79.21**	93.93	56.31	87.87	69.55
SPMNet [11]	76.84	34.47	86.21	67.03	82.00	73.70	92.16	55.66	85.60	72.35
Diff-UNet [7]	76.96	24.57	87.45	67.08	79.55	72.01	93.55	55.09	88.32	72.63
StructDiffSeg(Ours)	**79.08**	**23.50**	89.10	68.87	83.44	76.16	**94.77**	**60.41**	**90.26**	69.63

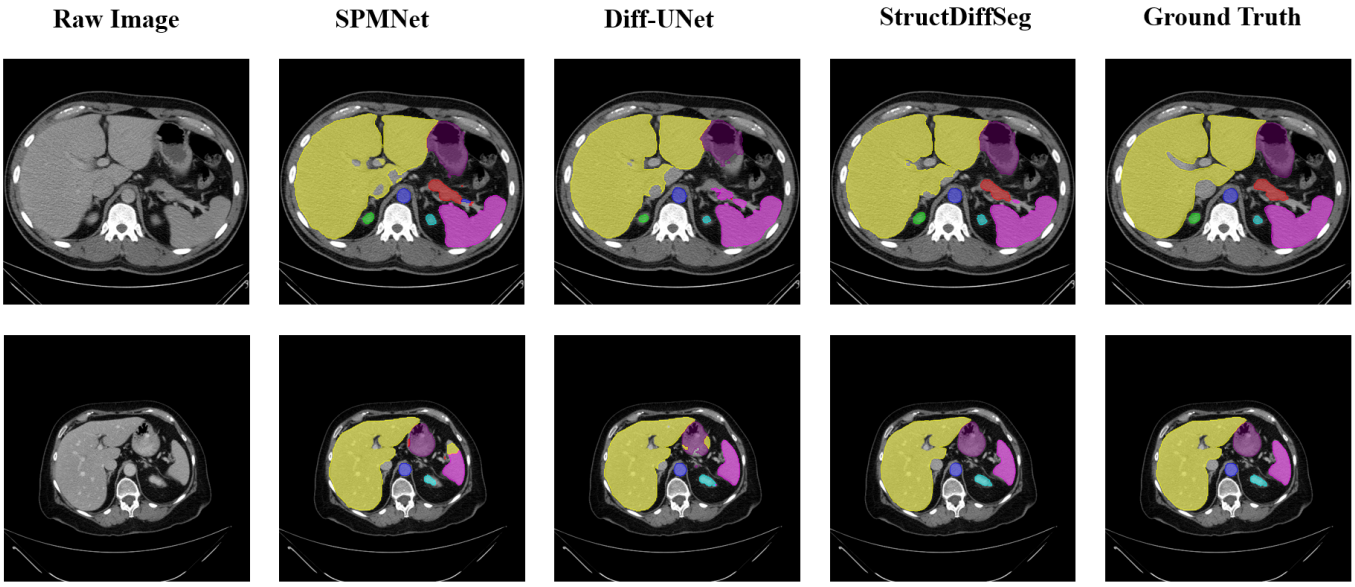

Fig. 2. Qualitative comparison on the Synapse multi-organ dataset. From left to right: Raw Image, SPMNet, Diff-UNet, StructDiffSeg, and Ground Truth. StructDiffSeg better preserves organ boundaries and reduces fragmented and spurious predictions, leading to more anatomically consistent segmentation results.

TABLE II

QUANTITATIVE COMPARISON ON THE ACDC DATASET. THE BEST RESULTS ARE IN **BOLD**.

Method	Avg. Dice↑	Avg. HD95↓	RV	Myo	LV
U-Net [1]	87.55	–	87.10	80.63	94.92
Att-UNet [27]	86.75	–	87.58	79.20	93.47
ViT [2]	87.57	–	86.07	81.88	94.75
TransUNet [14]	89.71	–	88.86	84.53	95.73
Swin-Unet [15]	90.00	–	88.55	85.62	95.83
DTBNet [28]	90.03	–	88.96	84.93	**96.21**
SPMNet [11]	90.71	1.91	**89.61**	88.92	93.61
Diff-UNet [7]	90.06	1.28	87.22	89.33	93.62
StructDiffSeg(Ours)	**90.77**	**1.27**	87.88	**89.92**	94.51

class-wise structural priors and structure-aware feature alignment in guiding the diffusion refinement process.

In terms of boundary accuracy, StructDiffSeg achieves competitive HD95 performance, indicating that the proposed method improves boundary smoothness without degrading anatomical consistency. Furthermore, the results suggest that structural conditioning effectively reduces fragmented regions and stabilizes predictions in challenging cardiac structures.

We further evaluate statistical significance using the Wilcoxon signed-rank test against Diff-UNet. The improvements in Dice are statistically significant on both datasets ($p = 0.0269$ for Synapse and $p = 0.0019$ for ACDC), while the differences in HD95 are not statistically significant.

C. Computational Efficiency

We further evaluate the computational efficiency of StructDiffSeg on the Synapse dataset, as summarized in Table III.

Compared with diffusion-based baseline Diff-UNet, StructDiffSeg exhibits comparable training and inference time, while maintaining a similar level of computational complexity in terms of FLOPs. Although StructDiffSeg introduces additional parameters due to the structural prior modeling and feature alignment modules, the overall computational overhead

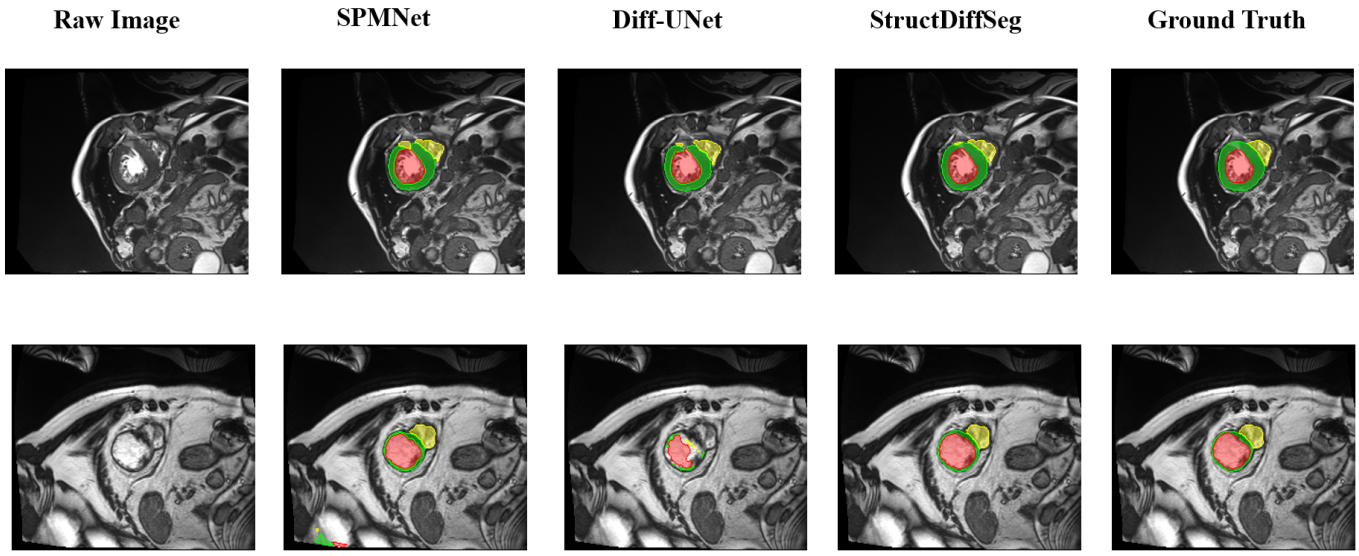

Fig. 3. Qualitative comparison on the ACDC dataset. From left to right: Raw Image, SPMNet, Diff-UNet, StructDiffSeg, and Ground Truth. StructDiffSeg produces more anatomically consistent cardiac structures with reduced fragmented regions and improved boundary delineation.

TABLE III
COMPUTATIONAL EFFICIENCY COMPARISON ON THE SYNAPSE DATASET. THE BEST RESULTS ARE IN **BOLD**.

Method	Training Time (h)	Inference Time (s)	Params (M)	FLOPs (G)
Diff-UNet [7]	3.09	0.90	22.09	145.36
SPMNet [11]	**2.42**	**0.77**	**17.94**	**51.98**
StructDiffSeg (Ours)	3.22	0.83	25.00	143.03

TABLE IV
ABLATION STUDY OF KEY COMPONENTS ON SYNAPSE AND ACDC. THE BEST RESULTS ARE IN **BOLD**.

Model	Synapse Dice↑	Synapse HD95↓	ACDC Dice↑	ACDC HD95↓
Diff-UNet (Baseline)	76.96	24.57	90.06	1.28
+CSPI	78.07	24.04	90.49	1.65
+SAAF	78.69	23.85	90.68	1.36
+URL (Ours)	**79.08**	**23.50**	**90.77**	**1.27**

remains moderate. Compared with SPMNet, StructDiffSeg requires higher computational cost, which is expected since diffusion-based refinement involves iterative denoising steps. However, the improved segmentation accuracy and structural consistency justify the additional computation. These results demonstrate that StructDiffSeg achieves a favorable trade-off between performance and efficiency, making it practical for medical image segmentation tasks.

D. Ablation Study

Component-wise analysis. To validate the contribution of each proposed component, we conduct a comprehensive ablation study on both datasets, as shown in Table IV.

Starting from the diffusion baseline, introducing the Class-wise Structural Prior Interaction module (+CSPI) leads to consistent improvements in both Dice and HD95, indicating that explicit structural priors help constrain the solution space and improve anatomical consistency.

Further incorporating the Structure-Aware Affinity Fusion module (+SAAF) yields additional performance gains, particularly in HD95. This suggests that aligning structural priors with diffusion features at the bottleneck is crucial for stabilizing the denoising process and reducing fragmented predictions.

Finally, with uncertainty-rectified learning (+URL), the full **StructDiffSeg** model achieves the best overall performance on both Synapse and ACDC. The improvement in HD95 is especially pronounced, demonstrating that dynamically emphasizing high-uncertainty regions enables better refinement of ambiguous boundaries.

E. Qualitative Results

Visual comparison. Qualitative comparisons are presented in Fig. 2 and Fig. 3.

As observed, diffusion-based baselines (e.g., Diff-UNet) tend to produce fragmented regions and small isolated artifacts, particularly in areas with low contrast or ambiguous boundaries. In contrast, discriminative methods with shape priors generate more globally coherent structures, but often suffer from over-smoothed boundaries and loss of fine details.

Compared with these approaches, **StructDiffSeg** achieves more stable and visually consistent segmentation results. The predicted masks exhibit coherent anatomical structures while preserving sharper boundaries. Fragmented regions and spurious predictions are effectively reduced, especially in challenging regions such as thin or low-contrast structures.

These improvements can be attributed to the integration of class-wise structural priors and structure-aware feature alignment, which guide the diffusion process toward anatomically consistent solutions. Overall, the qualitative results demonstrate that StructDiffSeg provides a better balance between global structural coherence and local boundary accuracy.

V. Conclusion

This paper presents **StructDiffSeg**, a structure-aware conditional diffusion framework for medical image segmentation, which explicitly incorporates anatomical priors into the diffusion refinement process. By integrating class-wise structural prior interaction and structure-aware affinity fusion, the proposed method effectively stabilizes stochastic denoising and reduces fragmented and spurious predictions commonly observed in diffusion-based segmentation models.

The proposed design introduces only moderate computational overhead, as structural alignment is performed at the bottleneck stage without additional heavy projection modules. Extensive experiments on the Synapse and ACDC datasets demonstrate that StructDiffSeg consistently improves segmentation accuracy and boundary delineation, achieving superior Dice scores and competitive HD95 compared with both discriminative and diffusion-based baselines.

These results suggest that incorporating explicit structural conditioning into diffusion models is a promising direction for achieving anatomically consistent and robust medical image segmentation.

In future work, we plan to extend this framework to more diverse datasets and explore more advanced conditioning strategies for improving efficiency and generalization.

References

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

[2] A. Dosovitskiy, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[3] L. Zhu, B. Liao, Q. Zhang, X. Wang, W. Liu, and X. Wang, "Vision mamba: Efficient visual representation learning with bidirectional state space model," *arXiv preprint arXiv:2401.09417*, 2024.

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

[5] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," *Advances in neural information processing systems*, vol. 34, pp. 8780–8794, 2021.

[6] 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.

[7] Z. Xing, L. Wan, H. Fu, G. Yang, Y. Yang, L. Yu, B. Lei, and L. Zhu, "Diff-unet: A diffusion embedded network for robust 3d medical image segmentation," *Medical Image Analysis*, p. 103654, 2025.

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

[9] O. Oktay, E. Ferrante, K. Kamnitsas, M. Heinrich, W. Bai, J. Caballero, S. A. Cook, A. De Marvao, T. Dawes, D. P. O'Regan *et al.*, "Anatomically constrained neural networks (acnns): application to cardiac image enhancement and segmentation," *IEEE transactions on medical imaging*, vol. 37, no. 2, pp. 384–395, 2017.

[10] A. Bozorgpour, Y. Sadegheih, A. Kazerouni, R. Azad, and D. Merhof, "Dermosegdiff: A boundary-aware segmentation diffusion model for skin lesion delineation," in *International workshop on predictive intelligence in medicine*. Springer, 2023, pp. 146–158.

[11] X. You, J. He, J. Yang, and Y. Gu, "Learning with explicit shape priors for medical image segmentation," *IEEE Transactions on Medical Imaging*, 2024.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[13] 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.

[14] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "Transunet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021.

[15] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, "Swin-unet: Unet-like pure transformer for medical image segmentation," in *European conference on computer vision*. Springer, 2022, pp. 205–218.

[16] A. Shaker, M. Maaz, H. Rasheed, S. Khan, M.-H. Yang, and F. S. Khan, "Unetr++: delving into efficient and accurate 3d medical image segmentation," *IEEE Transactions on Medical Imaging*, vol. 43, no. 9, pp. 3377–3390, 2024.

[17] J. Ruan, J. Li, and S. Xiang, "Vm-unet: Vision mamba unet for medical image segmentation," *ACM Transactions on Multimedia Computing, Communications and Applications*, 2024.

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

[19] T. Chen, C. Wang, and H. Shan, "Berdiff: Conditional bernoulli diffusion model for medical image segmentation," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2023, pp. 491–501.

[20] T. Chen, C. Wang, Z. Chen, Y. Lei, and H. Shan, "Hidiff: Hybrid diffusion framework for medical image segmentation," *IEEE Transactions on Medical Imaging*, 2024.

[21] H. Kalinic, "Atlas-based image segmentation: A survey," *Croatian Scientific Bibliography*, pp. 1–7, 2009.

[22] 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.

[23] H. Kervadec, J. Dolz, M. Tang, E. Granger, Y. Boykov, and I. B. Ayed, "Constrained-cnn losses for weakly supervised segmentation," *Medical image analysis*, vol. 54, pp. 88–99, 2019.

[24] W. Ding, S. Geng, H. Wang, J. Huang, and T. Zhou, "Fdiff-fusion: Denoising diffusion fusion network based on fuzzy learning for 3d medical image segmentation," *Information Fusion*, vol. 112, p. 102540, 2024.

[25] B. Kim, Y. Oh, and J. C. Ye, "Diffusion adversarial representation learning for self-supervised vessel segmentation," *arXiv preprint arXiv:2209.14566*, 2022.

[26] Z. Zhang, G. Fan, T. Liu, N. Li, Y. Liu, Z. Liu, C. Dong, and S. Zhou, "Introducing shape prior module in diffusion model for medical image segmentation," in *2023 6th International Conference on Mechatronics, Robotics and Automation (ICMRA)*. IEEE, 2023, pp. 185–190.

[27] 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.

[28] K. Chen, Y. Wang, and J. Cheng, "Dtbnet: medical image segmentation model based on dual transformer bridge," in *2024 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2024, pp. 1–8.