MedSegDiff-V2: Diffusion-Based Medical Image Segmentation with Transformer AAAI 2024

Junde Wu1,6,7,8, Wei Ji2 , Huazhu Fu3 , Min Xu*, 4,6 Yueming Jin1 , Yanwu Xu*5 ,1National University of Singapore, 2University of Alberta, 3A*STAR, 4Carnegie Mellon University, 5Singapore Eye Research Institute, 6Mohamed bin Zayed University of Artificial Intelligence, 7University of Oxford, 8Kids with Tokens

Speaker: Po-Jui Su

26, December 2024

Outline

01

Introduction

02

Method

03

Experiments

04

Conclusion

01 Introduction

Introduction: Medical Image Segmentation

Definition:

- Divide medical images into regions through pixel-wise segmentation to precisely classify and delineate specific organs and lesions.

Challenges about Medical Image Segmentation:

- **Data Scarcity and Imbalance**: Medical datasets often lack sufficient labeled samples, and the imbalance between small foreground regions and dominant background areas leads models to focus on the background.

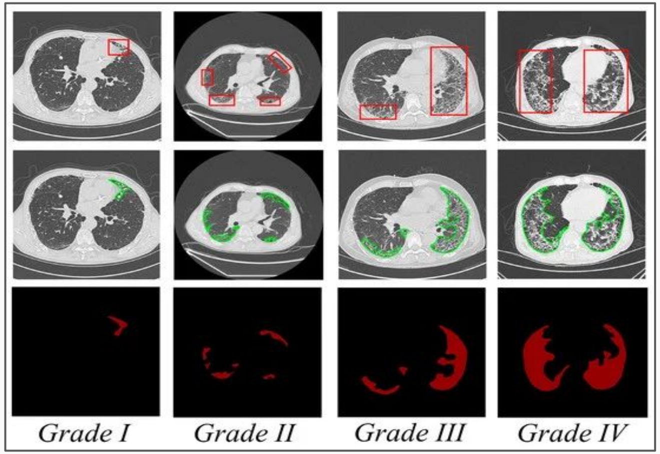


Fig 1. CT [1] manifestations of honeycomb lungs with different grades.

[1] Dong, Y., Yang, B., & Feng, X. (2024). Multitask Learning for Concurrent Grading Diagnosis and Semi-Supervised Segmentation of Honeycomb Lung in CT Images. Electronics, 13(11), 2115.

Introduction: Current Approaches

Current Approaches:

- **U-Net:**
  Utilizes an encoder-decoder architecture to capture multi-level features, making it highly effective for medical image segmentation tasks.

- **Transformer**:
  Using self-attention to generate global feature representations, and variants like TransUNet combine CNNs to integrate local and global features.

- **Diffusion Probabilistic Model:**
  Converts images into noise via random perturbations, then reconstructs the segmentation through reverse diffusion, generating diverse results to capture uncertainty, with most methods using U-Net for feature extraction.

Introduction: Problem Definition

Challenges and Limitations in Existing Methods:

- **U-Net:** Struggles to effectively capture more complex image structures or multi-level features, failing to integrate features from different levels as efficiently as ViTs.

- **Transformer:** Has strong feature extraction capabilities, but **its high sensitivity to noise causes instability[2]** when processing noisy data.

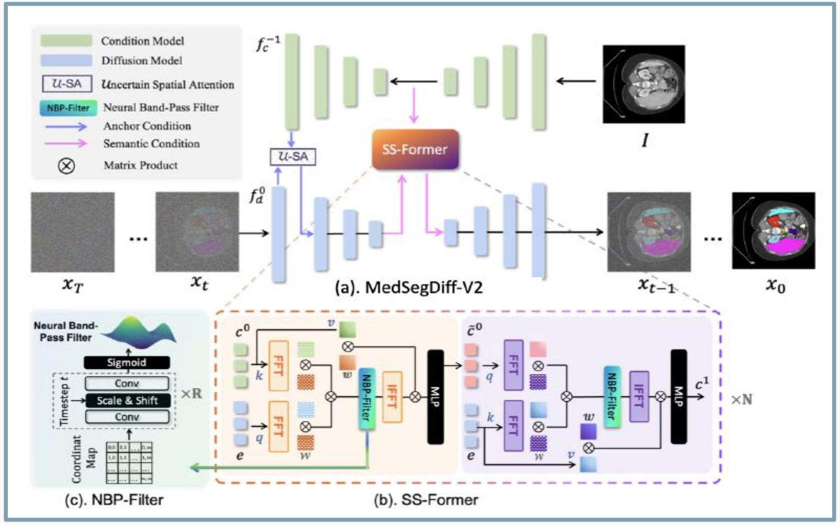
- **DPM:** Leads to **unstable segmentation performance(high variances)** **due to noise in the sampling process**.

- **Transformer + DPM:** Directly combining **Transformer** and **DPM** causes a **Feature Fusion Problem**, leading to performance decline due to the incompatibility between **the semantic features from the Transformer** and **the noisy mask features from DPM.**

[2] Naseer, M. M., Ranasinghe, K., Khan, S. H., Hayat, M., Shahbaz Khan, F., & Yang, M. H. (2021). Intriguing properties of vision transformers. Advances in Neural Information Processing Systems, 34, 23296-23308.

Introduction: Purpose and Contribution

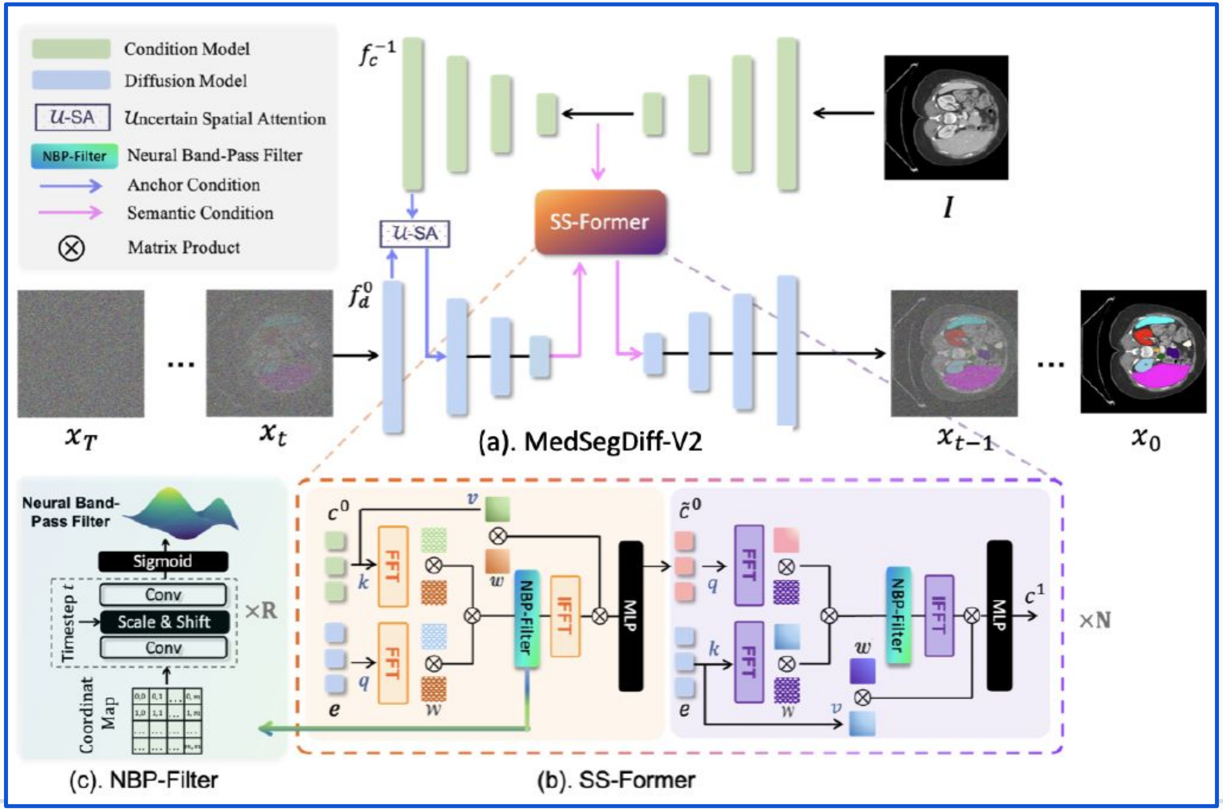
To address the challenges of **feature fusion** and **instability** in segmentation results, this paper proposes the **MedSegDiff-V2 method**, which **integrates Anchor Condition with U-SA and Semantic Condition with SS-Former techniques**, combining Transformer with a diffusion-based framework for medical image segmentation.



Architecture of MedSegDiff-V2

02 Method

Method Overview and Key Concepts



1. Diffusion+Condition

- Noisy Mask (Xt)
- Feature Extraction and Fusion

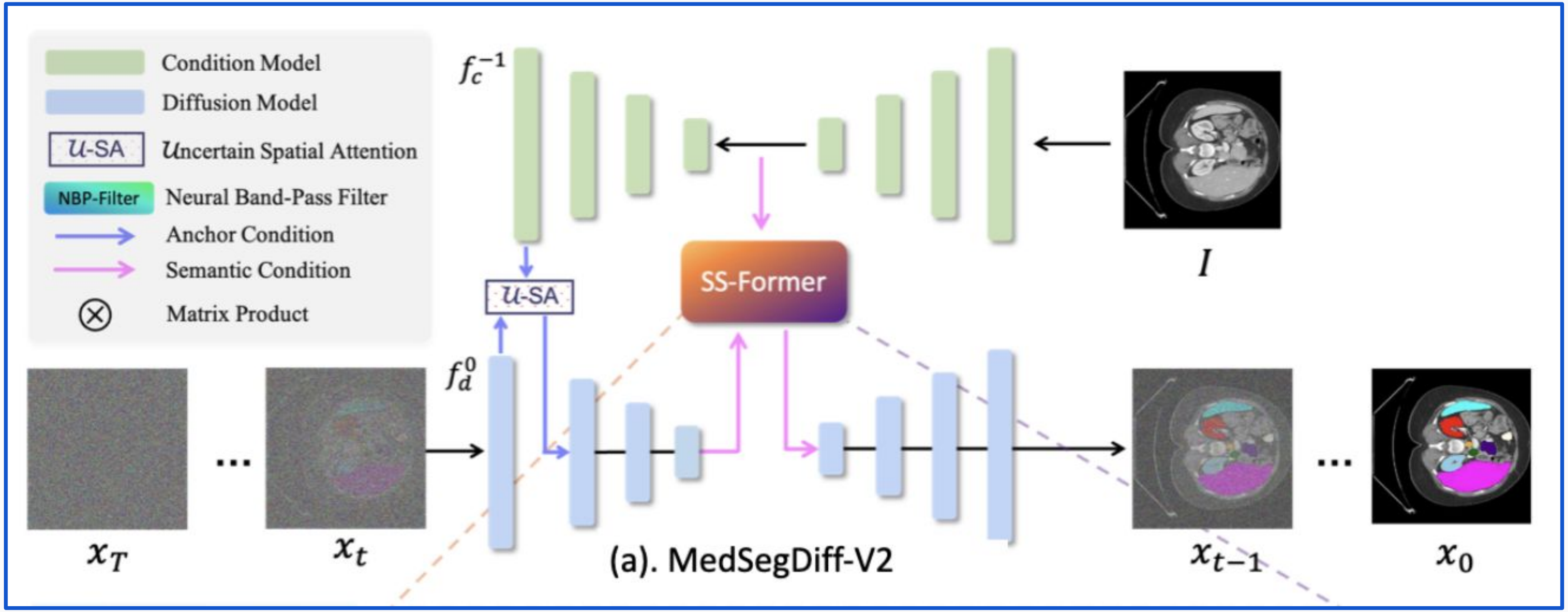
2. Anchor Condition with U-SA

- Smoothed Anchor Feature
- Prior Knowledge Integration
- Feature Fusion and Adjustment

3. Semantic Condition with SS-Former

- Fourier Transform
- Frequency Alignment
- NBP-Filter

Concepts of Diffusion and Condition Model:



Concepts of Diffusion and Condition Model:

Basic Concepts of Diffusion Model:

- Gradually add Gaussian noise to disrupt the training data, and reverse this noise-adding process to progressively denoise and recover the original information.
- **Eq1:** $$\boxed{p_\theta(x_{0:T-1}|x_T) = \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t),}$$
- **$p_\theta(x_{0:T-1} | x_T)$:** Represents the conditional probability from the final noisy state x_T back to the original image x_0 through the reverse process.
- **$p_\theta(x_{t-1} | x_t)$:** Each step of the reverse process predicts the previous state x_{t-1} from the current noisy state x_t
- **T:** Total number of steps in the reverse process.

Concepts of Diffusion and Condition Model:

Basic Concepts of Condition Model:

- **Feature Extraction:** The condition model typically uses U-Net to extract features (includes semantic condition、anchor condition from images.

- **Semantic Condition:** Integrates semantic features from the raw image and inputs them into the **SS-Former** to guide the diffusion process for better segmentation.

- **Anchor Condition:** Integrates **rough anchor features** from the Condition Model into the Diffusion Model to provide a stable prediction range, and uses Uncertain Spatial Attention (U-SA) to refine the feature fusion by weighting uncertain spatial locations.

Anchor Condition with Uncertain Spatial Attention

As noted by Naseer et al. (2021)[2], excessive noise sensitivity causes instability. Thus, this paper introduces **Anchor Condition and U-SA** to reduce noise impact and enhance segmentation accuracy.

- Smoothed Anchor Feature:

Applying Gaussian convolution to smooth the anchor features, and compare the smoothed feature map with the original feature map, selecting the most relevant part as the final anchor feature.

- Feature Fusion and Adjustment:

U-SA combines smoothed anchor features with current diffusion features and **integrates semantic prior knowledge from the Condition Model** to improve feature fusion and segmentation accuracy.

Anchor Condition with Uncertain Spatial Attention

- **Eq 2、3:**

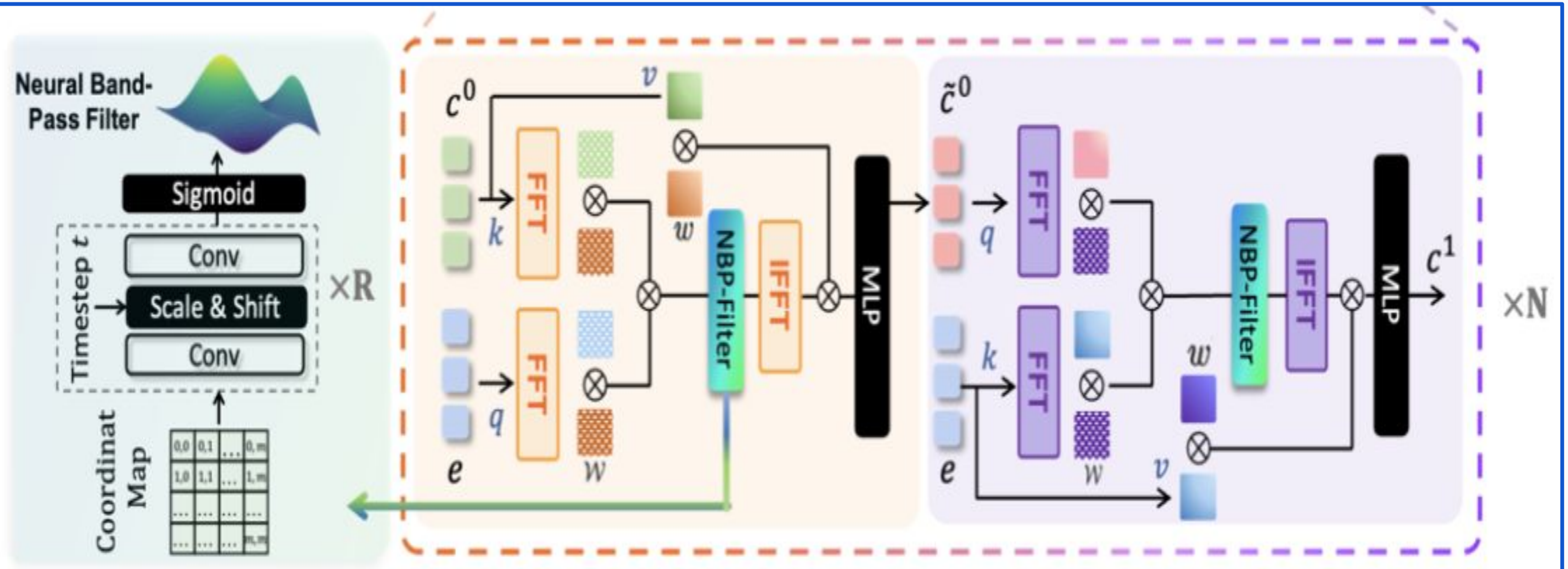
$$\boxed{\begin{aligned} f_{anc} &= Max(f_c^{-1} * k_{Gauss}, f_c^{-1}), \\ f_d'^0 &= Sigmoid(f_{anc} * k_{Conv_{1\times1}}) \cdot f_d^0 + f_d^0, \end{aligned}}$$

Explanation:

- **Eq 2:** The anchor feature $\mathbf{f_{c-1}}$ is smoothed using a **learnable Gaussian convolution kernel** $\mathbf{k_{Gauss}}$, and the maximum value between the smoothed feature map and the original feature map is selected to retain the most important information.

- **Eq 3:** The smoothed anchor feature $\mathbf{f_{anc}}$ is fused with the current diffusion feature f_d^0 **using a 1x1 convolution**, followed by **sigmoid activation**. The fused result is then added to the diffusion feature to enhance segmentation accuracy.

Structure of SS-Former



(c). NBP-Filter

(b). SS-Former

Concepts of SS-Former

SS-Former:

- Learns the interaction between semantic features and diffusion noise features in the frequency domain using **Fourier Transform**.

- Aligns the features to improve feature fusion, addressing the domain gap between semantic and noise embeddings.

NBP-Filter:

- Aligns features in the frequency domain using a neural network and **Fourier Transform**.

- Ensures relevant frequencies are preserved by learning a specific spectrum, adapts based on diffusion steps, and transforms the features back to the spatial domain for further processing.

03 Experiments

Experiment Details

Dataset:

- Multi-organ Segmentation: AMOS2022、BTCV
- Multi-modality Images: REFUGE-2、BraTs-2021、ISIC、TNMIX

Evaluation Metrics:

- **Dice Score↑、Intersection over Union (IOU)↑、Hausdorff Distance (95HD)↓**

Training Configuration:

- **GPU:** 4× NVIDIA A100 GPUs
- **Image size:** 256×256
- **Optimization:** AdamW
- **Learning rate:** 1E-4
- **Batch size:** 32

Experiment Results on Multi-organ Segmentation

The comparison of MedSegDiff-V2 with SOTA segmentation methods over AMOS dataset evaluated by Dice Score.

Methods	Spleen	R.Kid	L.Kid	Gall.	Eso.	Liver	Stom.	Aorta	IVC	Panc.	RAG	LAG	Duo.	Blad.	Pros.	Avg
TransUNet	0.881	0.928	0.919	0.813	0.740	0.973	0.832	0.919	0.841	0.713	0.638	0.565	0.685	0.748	0.692	0.792
UNetr	0.926	0.936	0.918	0.785	0.702	0.969	0.788	0.893	0.828	0.732	0.717	0.554	0.658	0.683	0.722	0.762
Swin-UNetr	0.959	0.960	0.949	0.894	0.827	**0.979**	0.899	0.944	0.899	0.828	0.791	0.745	0.817	**0.875**	0.841	0.880
nnUNet	0.965	0.959	0.951	0.889	0.820	0.980	0.890	0.948	0.901	0.821	0.785	0.739	0.806	0.869	0.839	0.878
EnsDiff	0.905	0.918	0.904	0.732	0.723	0.947	0.838	0.915	0.838	0.704	0.677	0.618	0.715	0.673	0.680	0.786
SegDiff	0.885	0.872	0.891	0.703	0.654	0.852	0.702	0.874	0.819	0.715	0.654	0.632	0.697	0.652	0.695	0.753
MedSegDiff	0.963	0.965	0.953	0.917	0.846	0.971	0.906	0.952	0.918	0.854	0.803	0.751	0.819	0.868	0.855	0.889
MedSegDiff + TransUNet	0.941	0.932	0.921	0.934	0.813	0.946	0.867	0.921	0.880	0.821	0.793	0.528	0.788	0.813	0.837	0.849
Anchor	0.872	0.901	0.892	0.784	0.802	0.910	0.835	0.908	0.810	0.735	0.682	0.651	0.583	0.631	0.728	0.781
MedSegDiff-V2	**0.971**	**0.969**	**0.964**	**0.932**	**0.864**	0.976	**0.934**	**0.968**	**0.925**	**0.871**	**0.815**	**0.762**	**0.827**	0.873	**0.871**	**0.901**

Experiment Results on Multi-organ Segmentation

The comparison of MedSegDiff-V2 with SOTA segmentation methods over BTCV dataset evaluated by Dice Score.

Model	Spleen	R.Kid	L.Kid	Gall.	Eso.	Liver	Stom.	Aorta	IVC	Veins	Panc.	AG	Ave
TransUNet	0.952	0.927	0.929	0.662	0.757	0.969	0.889	0.920	0.833	0.791	0.775	0.637	0.838
UNetr	0.968	0.924	0.941	0.750	0.766	0.971	0.913	0.890	0.847	0.788	0.767	0.741	0.856
Swin-UNetr	0.971	0.936	0.943	0.794	0.773	0.975	0.921	0.892	0.853	0.812	0.794	0.765	$\boxed{0.869}$
nnUNet	0.942	0.894	0.910	0.704	0.723	0.948	0.824	0.877	0.782	0.720	0.680	0.616	0.802
EnsDiff	0.938	0.931	0.924	0.772	0.771	0.967	0.910	0.869	0.851	0.802	0.771	0.745	0.854
SegDiff	0.954	0.932	0.926	0.738	0.763	0.953	0.927	0.846	0.833	0.796	0.782	0.723	0.847
MedSegDiff	0.973	0.930	0.955	0.812	0.815	0.973	0.924	0.907	0.868	0.825	0.788	0.779	$\boxed{0.879}$
MedSegDiff +TransUNet	0.912	0.876	0.846	0.645	0.718	0.947	0.824	0.876	0.715	0.775	0.672	0.618	0.785
Anchor	0.928	0.882	0.873	0.652	0.750	0.951	0.829	0.855	0.731	0.714	0.683	0.602	0.787
MedSegDiff-V2	**0.978**	**0.941**	**0.963**	**0.848**	**0.818**	**0.985**	**0.940**	**0.928**	**0.869**	**0.823**	**0.831**	**0.817**	$\boxed{\mathbf{0.895}}$

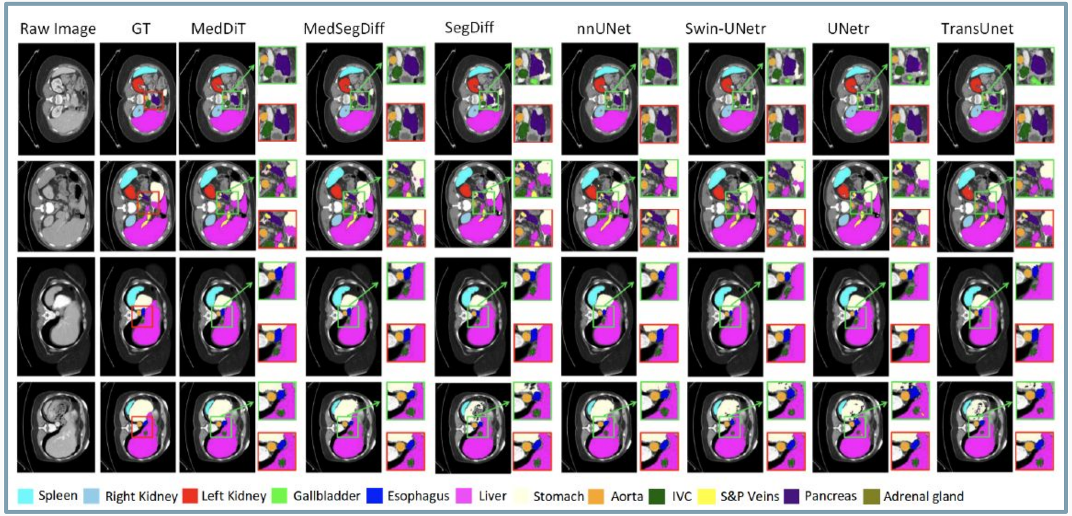
Experiment Results on Multi-modality Images

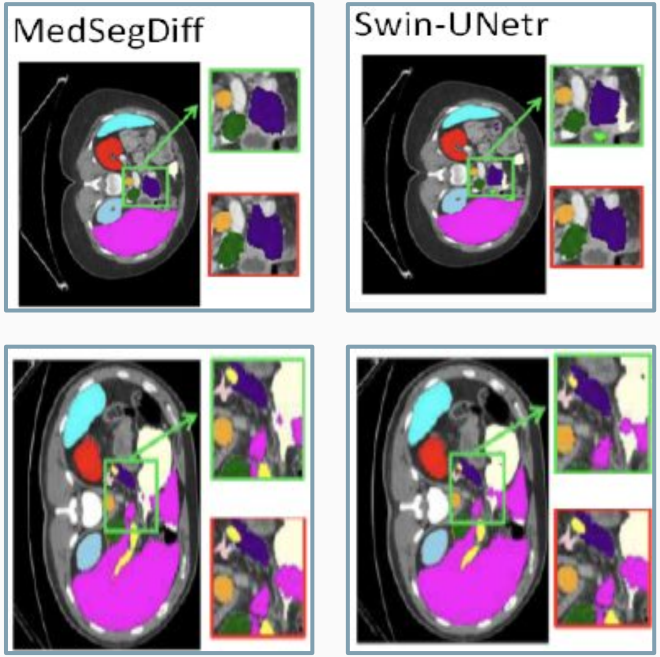
The comparison of MedSegDiff-V2 with SOTA segmentation methods on different image modalities.

		REFUGE2-Disc		REFUGE2-Cup		BraTs			TNMIX		ISIC	
		Dice	IoU	Dice	IoU	Dice	IoU	HD95	Dice	IoU	Dice	IoU
Optic Disc/Cup	ResUNet	92.9	85.5	80.1	72.3	78.4	71.3	18.71	78.3	70.7	87.1	78.2
	BEAL	93.7	86.1	83.5	74.1	78.8	71.7	18.53	78.6	71.6	86.6	78.0
Brain Tumor	TransBTS	94.1	87.2	85.4	75.7	87.6	78.44	12.44	83.8	75.5	88.1	80.6
	SwinBTS	95.2	87.7	85.7	75.9	88.7	81.2	10.03	84.5	76.1	89.8	82.4
Thyroid Nodule	MTSeg	90.3	83.6	82.3	73.1	82.2	74.5	15.74	82.3	75.2	87.5	79.7
	UltraUNet	91.5	82.8	83.1	73.78	84.5	76.3	14.03	84.5	76.2	89.0	81.8
Skin Lesion	FAT-Net	91.8	84.8	80.9	71.5	79.2	72.8	17.35	80.8	73.4	90.7	83.9
	BAT	92.3	85.8	82.0	73.2	79.6	73.5	15.49	81.7	74.2	91.2	84.3
General Med Seg	nnUNet	94.7	87.3	84.9	75.1	88.5	80.6	11.20	84.2	76.2	90.8	83.6
	TransUNet	95.0	87.7	85.6	75.9	86.6	79.0	13.74	83.5	75.1	89.4	82.2
	UNetr	94.9	87.5	83.2	73.3	87.3	80.6	12.81	81.7	73.5	89.7	82.8
	Swin-UNetr	95.3	87.9	84.3	74.5	88.4	81.8	11.36	83.5	74.8	90.2	83.1
Diffusion Based	EnsemDiff	94.3	87.8	84.2	74.4	88.7	80.9	10.85	83.9	75.3	88.2	80.7
	SegDiff	92.6	85.2	82.5	71.9	85.7	77.0	14.31	81.9	74.8	87.3	79.4
	MedsegDiff	95.1	87.6	85.9	76.2	88.9	81.2	10.41	84.8	76.4	91.3	84.1
	MedsegDiff+TransUNet	91.8	84.5	82.1	72.6	86.1	78.0	13.88	79.2	71.4	84.6	75.5
Proposed	MedSegDiff-V2	**96.7**	**88.9**	**87.9**	**80.3**	**90.8**	**83.4**	**7.53**	**88.7**	**81.5**	**93.2**	**85.3**

Experiment Results - Visual Comparison

The visual comparison with SOTA segmentation models on BTCV.





Ablation Study

An ablation study on Anchor Conditioning and SS-Former.

Anc.Cond.		Sem.Cond.		AMOS	BTCV	OpticCup	BrainTumor	ThyroidNodule
SA	$\mathcal{U}$-SA	SS-Former (w/o Filter)	NBP-Filter	Ave-Dice (%)	Ave-Dice (%)	Dice (%)	Dice (%)	Dice (%)
				78.6	85.4	84.6	88.2	84.1
✓				83.5	85.8	85.2	88.7	84.6
	✓			86.7	86.6	85.7	89.4	86.5
	✓	✓		87.8	87.1	86.5	89.8	86.8
	✓	✓	✓	**90.1**	**89.5**	**87.9**	**90.8**	**88.7**

 # SA denotes Spatial Attention

04 Conclusion

Conclusion

- **Contribution:**

  This paper enhances the diffusion-based medical image segmentation framework, named MedSegDiff-V2, by integrating the novel SS-Former structure, which effectively captures the interaction between noise and semantic features, into the original UNet backbone.

- **Experiment results:**

  The experimental results show that our approach outperforms the SOTA methods across various evaluation metrics in both multi-organ segmentation and multi-modality image datasets.