Adapting Vision Foundation Models for Real-time Ultrasound Image Segmentation

Xiaoran Zhang*[1], Eric Z. Chen[2], Lin Zhao[2], Xiao Chen[2], Yikang Liu[2], Boris Maihe[2], James Duncan[1], Terrence Chen[2], Shanhui Sun[2]

1. Yale University, 2. United Imaging Intelligence, *Work done during internship

Project page

Summary

We propose a novel approach that adapts hierarchical vision foundation models for real-time ultrasound image segmentation.

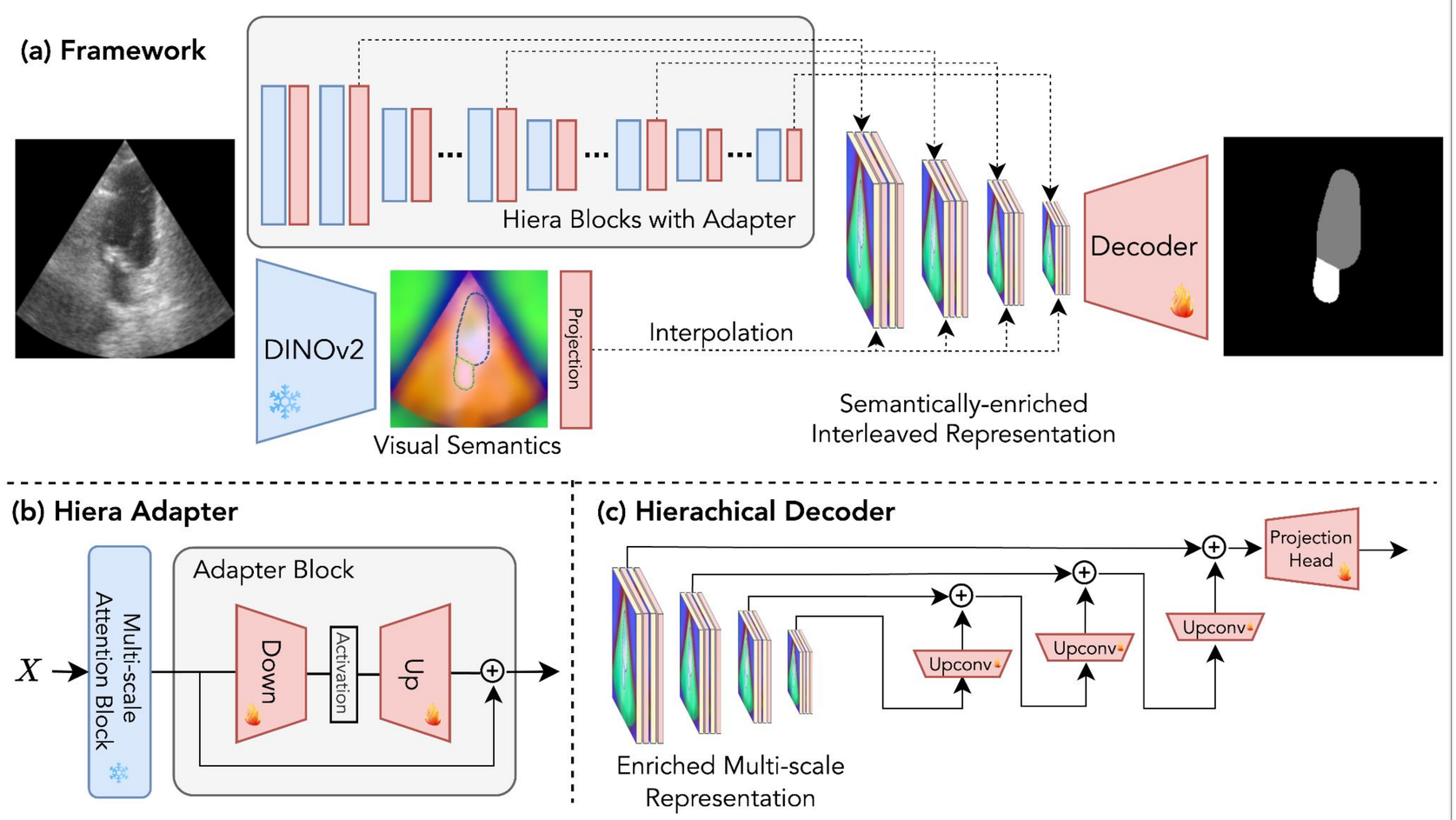
Motivation

- Existing ultrasound segmentation methods often struggle with adaptability to new tasks, relying on costly manual annotations.
- Current real-time approaches fail to match state-of-the-art performance.

Contribution

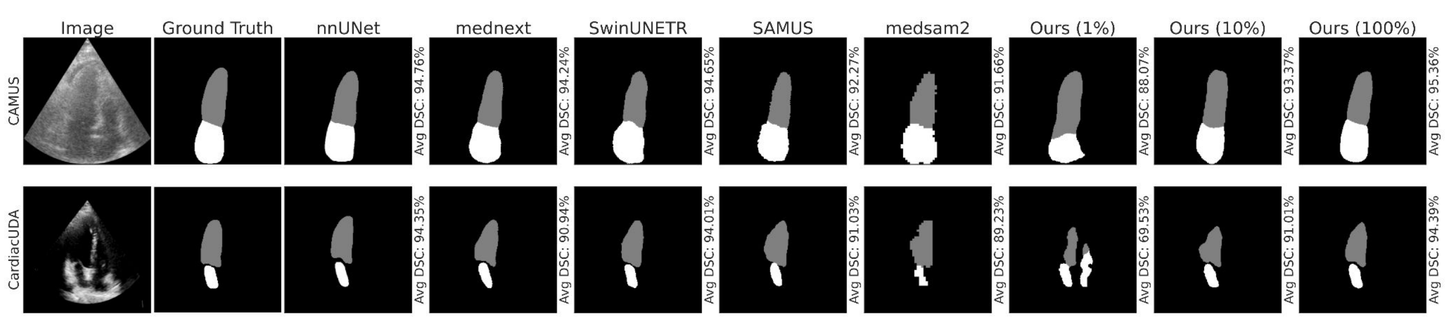
- We propose an approach that adapts Hiera encoders and integrates DINOv2 feature for enhanced feature representation.
- Our method excels under limited supervision and achieves state-of-the-art performance on CAMUS and TN3K with real-time inference.

Methods



Results

Data efficiency and adaptability on cardiac ultrasound



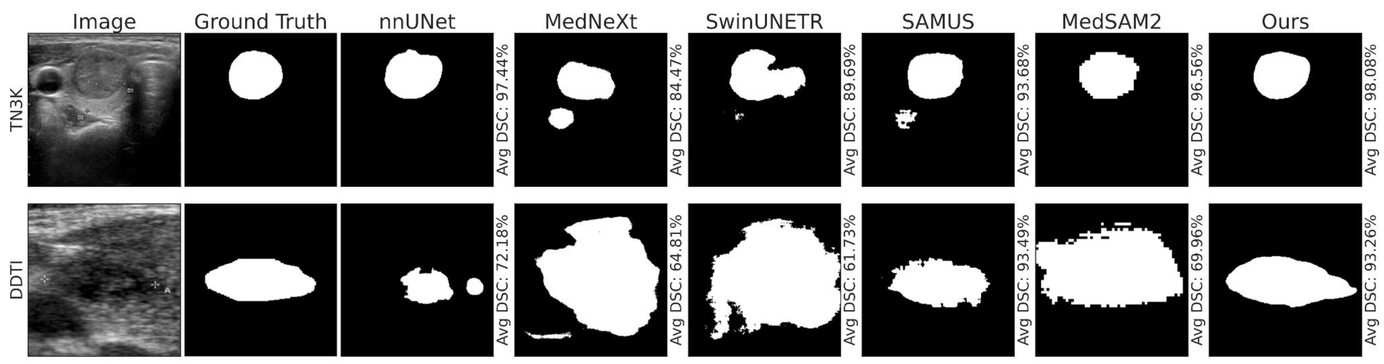
	CAMUS [20]						CardiacUDA [46]					
	1%		10%		100%		1%		10%		100%	
	DSC ↑	HD ↓	DSC ↑	HD ↓	DSC ↑	HD ↓	DSC ↑	HD ↓	DSC ↑	HD ↓	DSC ↑	HD ↓
UNet [35]	59.49	51.51	80.14	26.36	87.70	15.05	34.65	41.04	72.83	26.24	81.69	20.78
nnUNet [16]	67.02	54.22	89.36	8.83	91.69	**6.35**	35.48	58.94	49.49	34.12	**90.39**	8.32
MedNeXt [36]	63.30	48.23	84.73	26.16	89.49	9.83	19.05	158.71	70.97	32.59	85.54	8.34
SwinUNETR [13]	71.77	48.80	84.99	23.81	89.36	12.24	51.54	90.38	81.44	24.11	85.81	9.94
SAMUS [23]	75.00	26.29	87.43	16.24	91.11	8.79	63.65	23.91	82.10	**10.15**	86.14	7.70
MedSAM2 [52]	9.75	137.29	44.24	42.21	84.76	12.96	3.18	154.61	3.82	151.34	75.78	12.95
Ours	**81.96**	**17.16**	**90.30**	**8.44**	**92.01**	6.75	**65.84**	**23.57**	**82.38**	11.07	87.44	**6.45**

More on CAMUS

Method	DSC ↑	IoU ↑	HD95 ↓	ASD ↓
SwinUNet	88.84	80.33	6.10	2.60
H2Former	91.31	84.30	5.27	2.05
MedSAM	85.42	75.14	8.42	3.34
MSA	88.03	78.98	7.53	2.85
SAMed	87.45	78.14	9.17	3.10
SonoSAM	89.80	81.79	6.60	2.45
MemSAM	93.31	87.61	**3.82**	**1.57**
Ours	**93.80**	**88.49**	4.80	1.90

- Our approach remains highly effective under limited supervision, significantly outperforming baselines when trained with only 1% and 10% of the training data.
- We outperform existing state-of-the-art methods on CAMUS.

Cross-dataset generalization on thyroid ultrasound



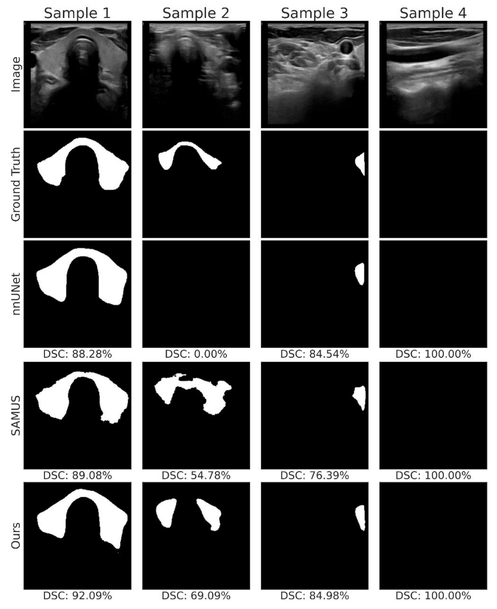
	TN3K [12]		DDTI [30]		Stanford [1]		TG3K [12]		In-house	
	DSC ↑	HD95 ↓	DSC ↑	HD95 ↓	DSC ↑	HD95 ↓	DSC ↑	HD95 ↓	DSC ↑	HD95 ↓
UNet [35]	67.93	41.26	48.43	52.60	89.09	18.96	70.27	65.88	68.50	60.72
nnUNet [16]	85.13	17.24	73.26	40.43	96.92	2.72	77.50	22.52	79.38	18.59
MedNeXt [36]	70.77	32.58	71.59	40.02	97.33	2.59	77.13	40.42	79.75	20.11
SwinUNETR [13]	71.84	37.89	70.59	41.76	**97.63**	2.52	69.72	43.37	76.21	39.52
SAMUS [23]	82.60	18.18	77.48	33.53	96.36	2.70	33.18	58.15	78.31	32.61
MedSAM2 [52]	69.02	31.43	69.69	39.02	87.20	9.58	75.86	21.10	79.47	18.70
Ours	**86.01**	**15.43**	**81.52**	**26.68**	97.33	**2.23**	**83.04**	**12.55**	**82.59**	**17.11**

More on TN3K

Method	ACC ↑	IoU ↑	DSC↑	HD95↓
TRFE	96.71	68.33	81.91	17.96
SegNet	96.72	66.54	79.91	17.13
DeepLabv3	97.19	70.60	82.77	13.92
TRFE+	97.04	71.38	83.30	13.23
SHAN	96.73	73.59	84.61	**4.05**
Ours	**97.60**	**78.13**	**86.01**	15.42

- Our method demonstrates strong generalization capability when trained on TN3K and tested on DDTI and outperforms existing state-of-the-art methods on other thyroid ultrasound datasets.

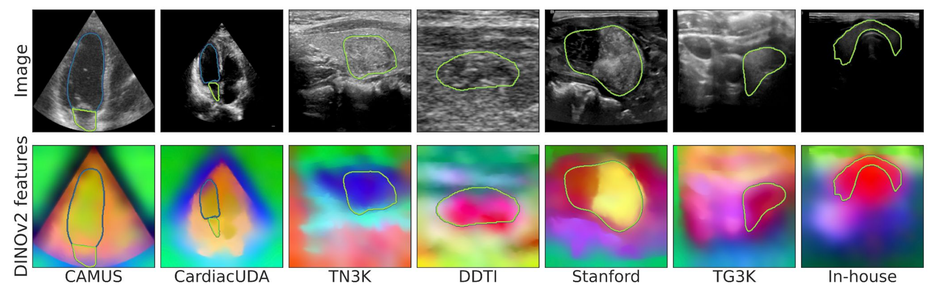
In house evaluation



We evaluate real-world applicability using image-level detection metrics (precision, recall, specificity, F1-score) on an in-house dataset where the thyroid gland may be absent.

Model	Precision ↑	Recall ↑	Specificity ↑	F1-Score ↑
UNet	64.75	93.92	12.41	96.86
nnUNet	91.97	93.48	86.02	96.63
MedNeXt	93.90	86.96	90.32	93.03
SwinUNETR	62.09	95.61	0.00	97.75
SAMUS	**99.71**	98.26	**99.50**	99.12
MedSAM2	96.67	85.61	94.95	92.25
Ours	87.21	**98.41**	75.27	**99.20**

DINOv2 feature visualization



We visualize DINOv2 features via the first three PCA components, showing meaningful semantics for segmenting indistinct ultrasound boundaries.

Ablation study on CAMUS

H-Dec.	H-Adp.	DINOv2	Interleave	DSC ↑	HD ↓
✗	✗	✗	✗	78.23	30.98
✓	✗	✗	✗	90.85	9.42
✓	✓	✗	✗	91.89	6.84
✓	✓	✓	✗	91.90	6.78
✓	✓	✓	✓	**92.01**	**6.75**

(1) Hierarchical decoder (H-Dec.) (2) Hiera adapter (H-Adp.) (3) DINOv2 feature integration (4) Feature interleaving.

Inference speed

Our method runs at ~30 FPS on a single GPU (NVIDIA L40S) and ~77 FPS with TensorRT when tested on 224 × 224 images.

Methods	FPS
nnUNet	10
SAMUS	15
MedSAM2	17
Ours	30

Paper: https://arxiv.org/abs/2503.24368

Project page: https://voldemort108x.github.io/adp_seg/