



Yale GRADUATE SCHOOL
OF ARTS AND SCIENCES



Learning Reliable and Adaptable Models for Medical Imaging

From Unsupervised Registration to Semantic Segmentation *and Beyond*

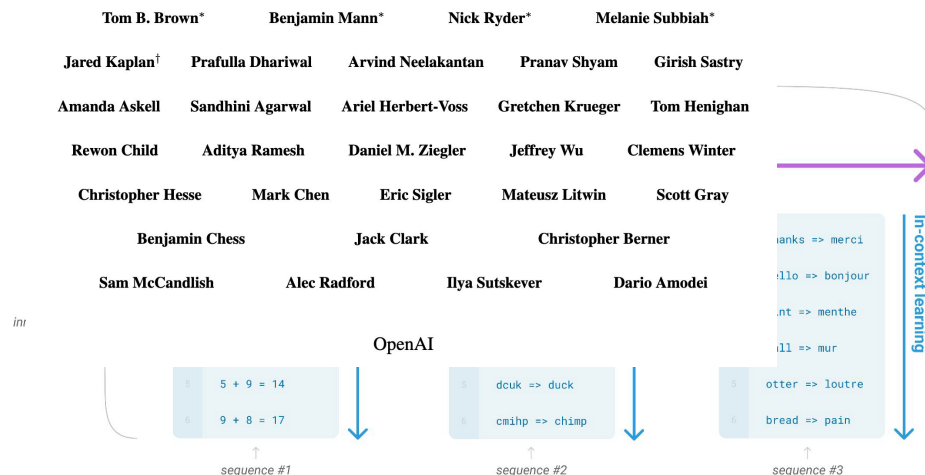
PhD Defense - Xiaoran Zhang

Advisor: Prof. James S. Duncan

Committee: Prof. Alex Wong, Prof. Lawrence H. Staib

Generalist AI models are causing a major paradigm shift

Language Models are Few-Shot Learners



Brown, Tom, et al. "Language models are few-shot learners." *Advances in neural information processing systems* 33 (2020): 1877-1901.

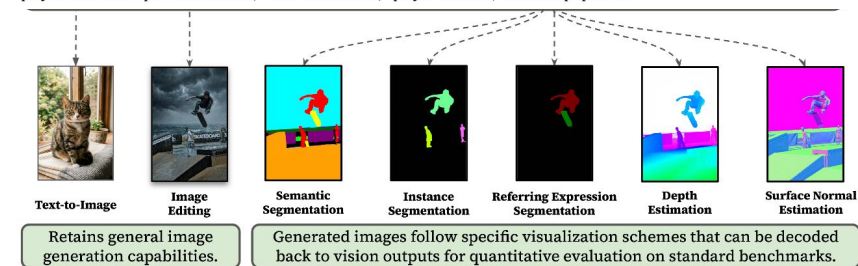
Google DeepMind

2026-05-15

Image Generators are Generalist Vision Learners

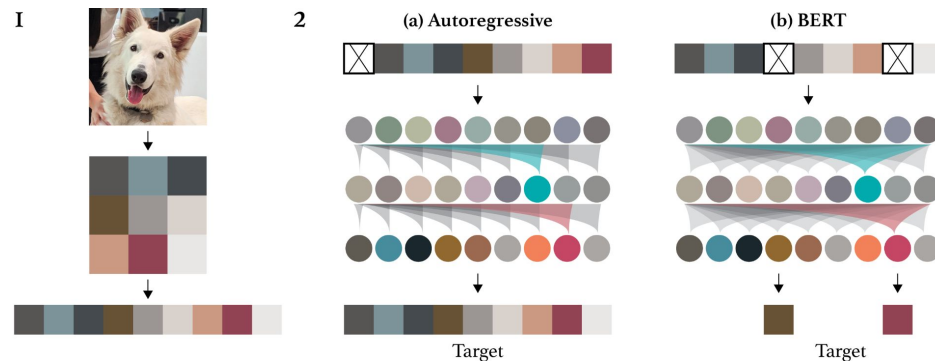
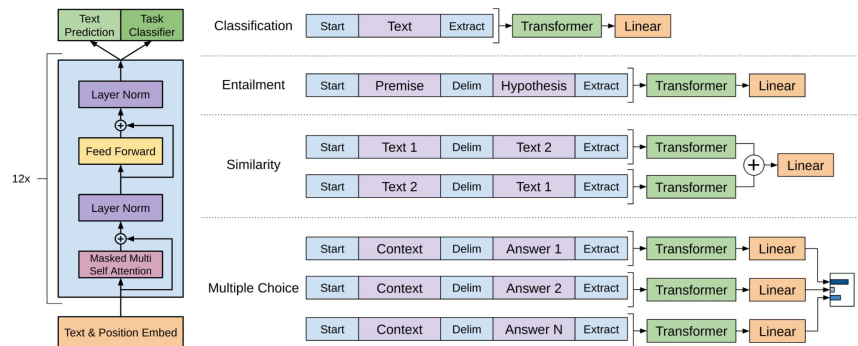
Valentin Gabeur*, Shangbang Long*, Songyou Peng*, Paul Voigtlaender[▽], Shuyang Sun[▽], Yanan Bao[▽], Karen Truong[▽], Zhicheng Wang[▽], Wenlei Zhou[▽], Jonathan T. Barron[▽], Kyle Genova[▽], Nithish Kannan[▽], Sherry Ben[▽], Yandong Li[▽], Mandy Guo[▽], Suhas Yogin[▽], Yiming Gu[†], Huizhong Chen[†], Oliver Wang[‡], Saining Xie[‡], Howard Zhou[‡], Kaiming He[‡], Thomas Funkhouser[‡], Jean-Baptiste Alayrac[‡] and Radu Soricut[‡]

*project leads and equal contributions, [▽]core contributors, [†]project advisors, [‡]leadership sponsors



Gabeur, Valentin, et al. "Image generators are generalist vision learners." *arXiv preprint arXiv:2604.20329* (2026).

The secret sauce: generative pre-training

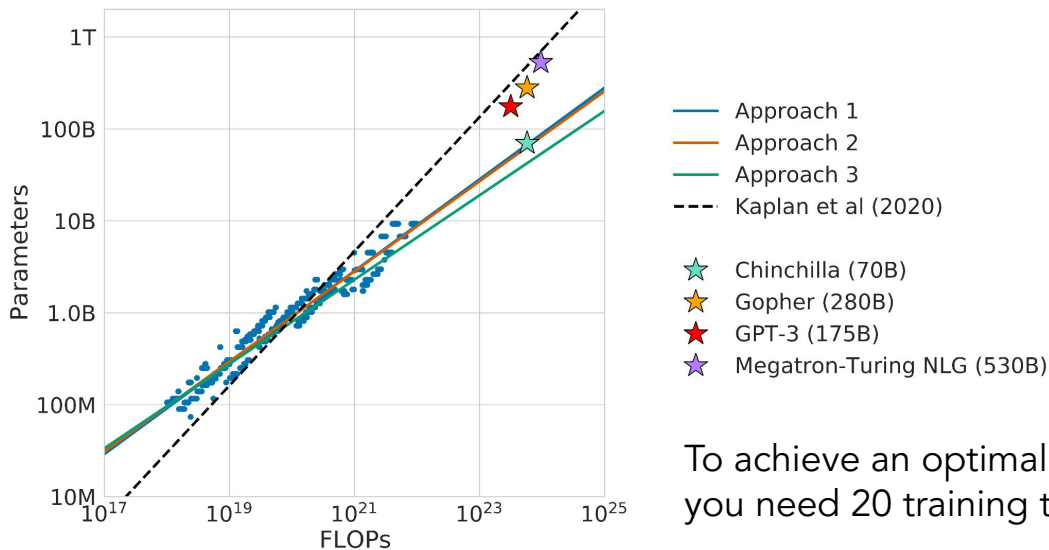


Maximize the likelihood from unsupervised corpus of tokens: $\max \sum_i \log P(u_i | u_{i-k}, \dots, u_{i-1}; \theta)$

Radford, Alec, et al. "Improving language understanding by generative pre-training." (2018).

Chen, Mark, et al. "Generative pretraining from pixels." International conference on machine learning. PMLR, 2020.

But medical imaging can't follow the same recipe



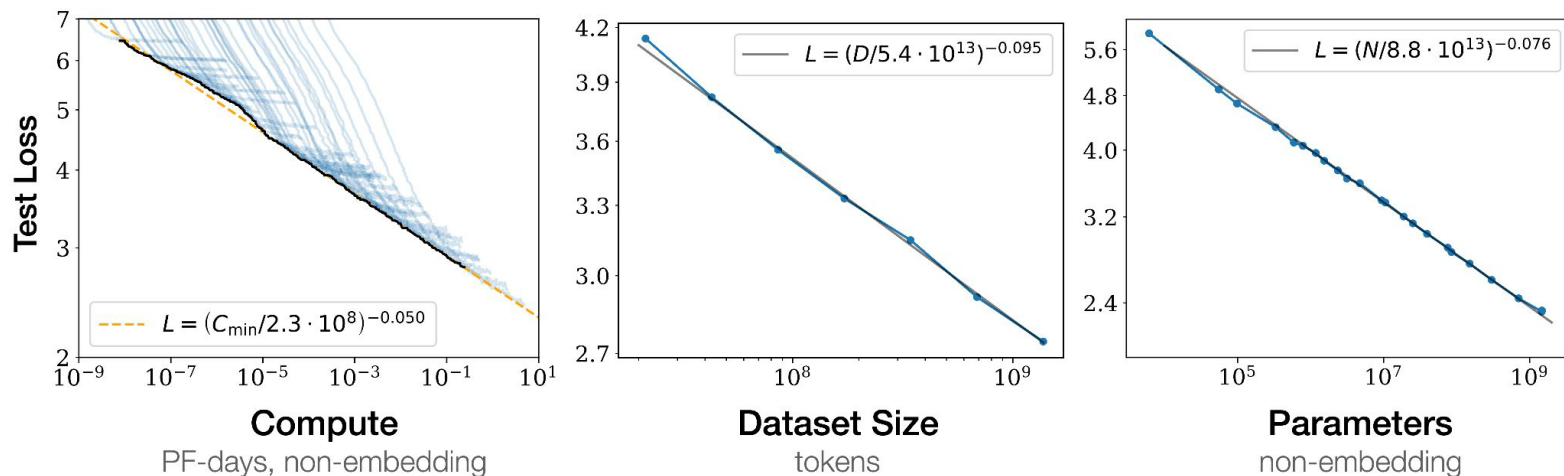
To achieve an optimal compute allocation, you need 20 training tokens per parameter.

For models with 1b parameters, 20 billion tokens are needed (~78 million images*)

Hoffmann, Jordan, et al. "Training compute-optimal large language models." *Advances in neural information processing systems* (2022).

* calculation based on patch size of 16 and image size 256x256

But medical imaging can't follow the same recipe

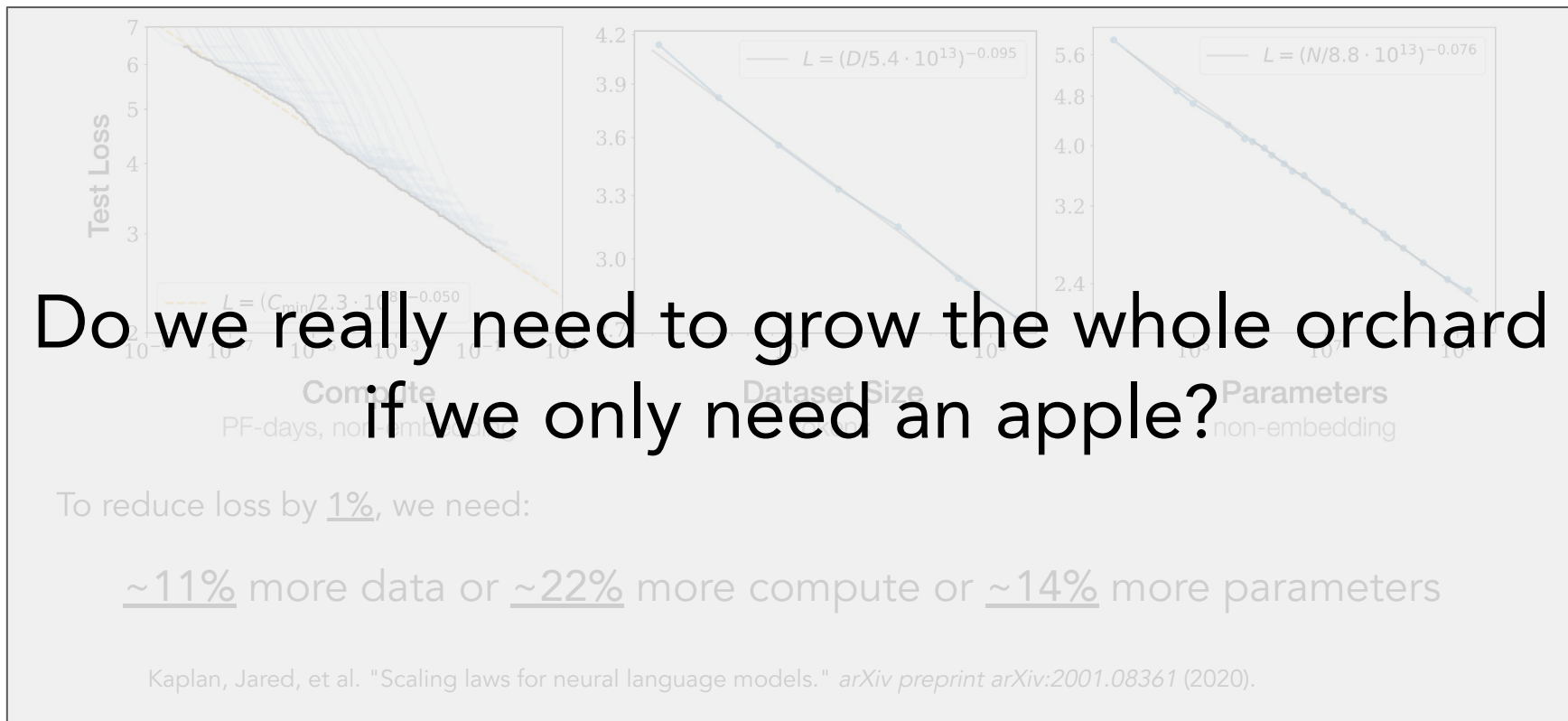


To reduce loss by 1%, we need:

~11% more data or ~22% more compute or ~14% more parameters

Kaplan, Jared, et al. "Scaling laws for neural language models." *arXiv preprint arXiv:2001.08361* (2020).

But medical imaging can't follow the same recipe



Revisit data scaling assumptions in generalist models

Assumption 1: Clean supervision is abundant

Assumption 2: Test distribution is covered by training distribution

Revisit data scaling assumptions in generalist models

Assumption 1: Clean supervision is abundant → however Medical imaging data are often noisy

Build reliable models that identify trustworthy supervision

Assumption 2: Test distribution is covered by training distribution → however Medical imaging data are often heterogeneous

Build adaptable models that are versatile to distribution shifts

"*Reliable*" and "*Adaptable*"

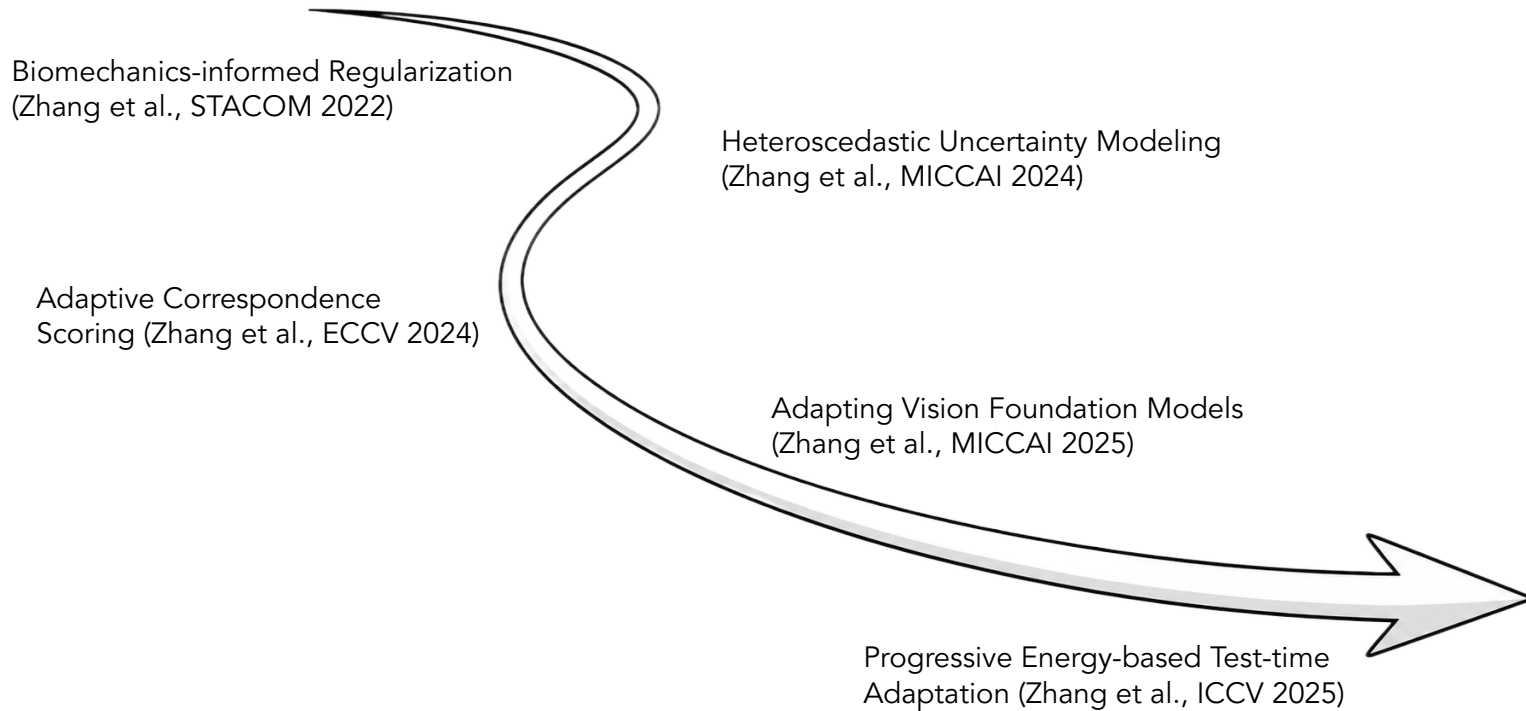
Reliable

Ability of a model to learn from imperfect supervision by identifying, weighting, or constraining the signals that are more trustworthy.

Adaptable

Ability of a model to maintain robust performance under changing data distributions.

Thesis roadmap



Thesis roadmap

Biomechanics-informed Regularization
(Zhang et al., STACOM 2022)

Adaptive Correspondence
Scoring (Zhang et al., ECCV 2024)

Heteroscedastic Uncertainty Modeling
(Zhang et al., MICCAI 2024)

Adapting Vision Foundation Models
(Zhang et al., MICCAI 2025)

Progressive Energy-based Test-time
Adaptation (Zhang et al., ICCV 2025)

Part I:
Identify reliable supervision
signals for unsupervised
registration.

Thesis roadmap

Biomechanics-informed Regularization
(Zhang et al., STACOM 2022)

Adaptive Correspondence
Scoring (Zhang et al., ECCV 2024)

Heteroscedastic Uncertainty Modeling
(Zhang et al., MICCAI 2024)

Part II:
Build adaptable models for
segmentation.

Adapting Vision Foundation Models
(Zhang et al., MICCAI 2025)

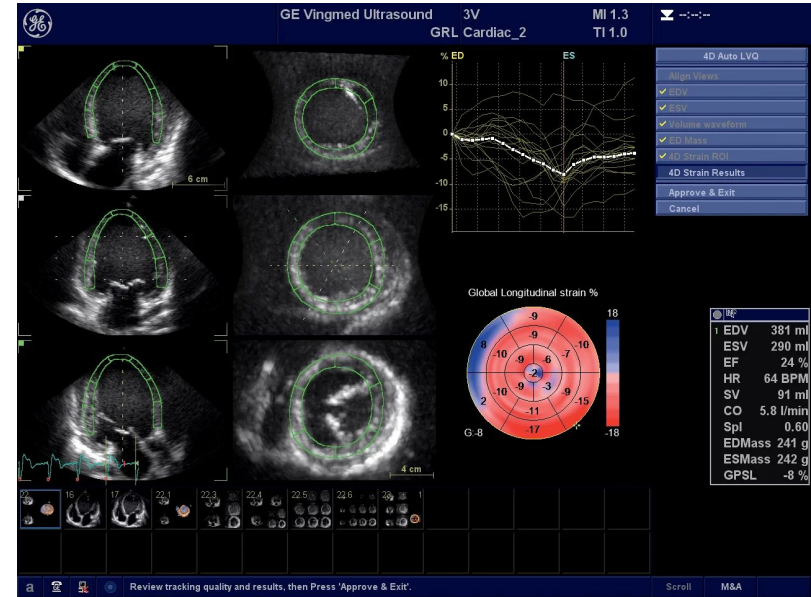
Progressive Energy-based Test-time
Adaptation (Zhang et al., ICCV 2025)

Clinical anchor - cardiac strain analysis

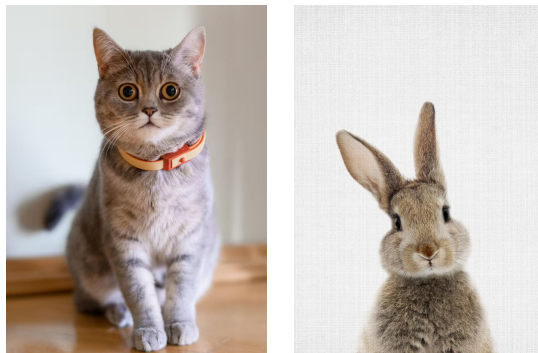
Unsupervised
registration



Semantic
segmentation

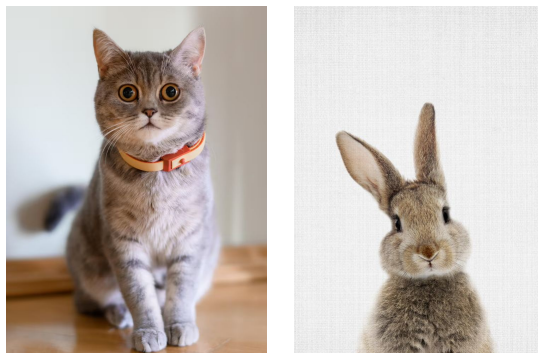


Challenge: data heterogeneity



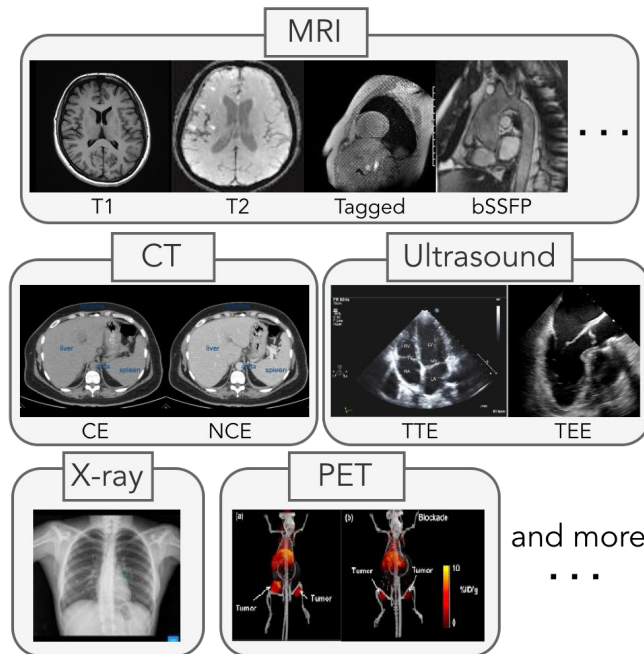
Natural image: RGB

Challenge: data heterogeneity



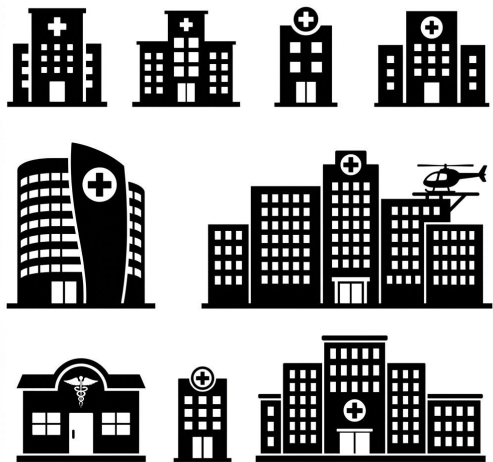
Natural image: RGB

whereas

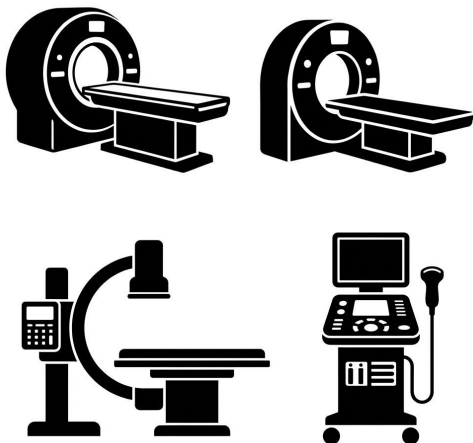


Medical images: MRI/CT/Ultrasound ...

Challenge: the domain gap



Different hospitals

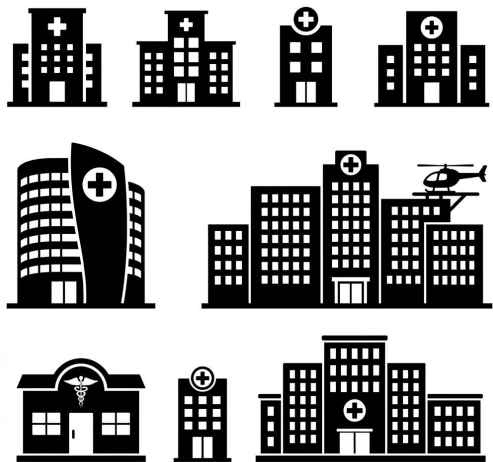


Different scanners

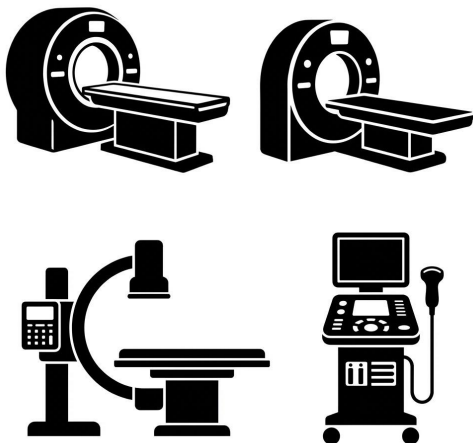


Build a *generalist* model that works for everything is challenging.

Challenge: the domain gap



Different hospitals



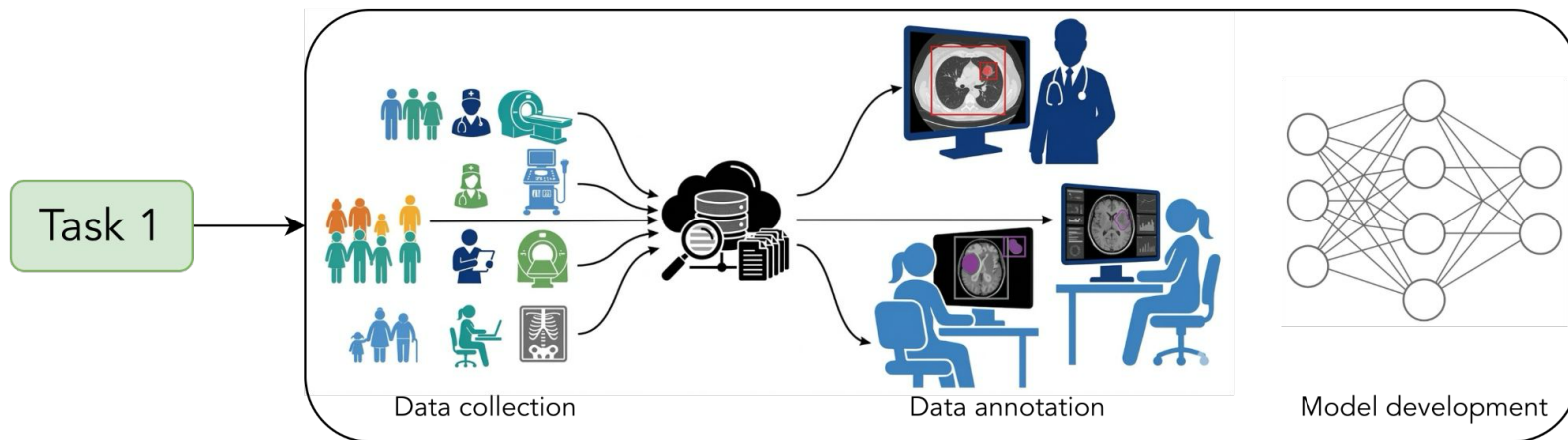
Different scanners



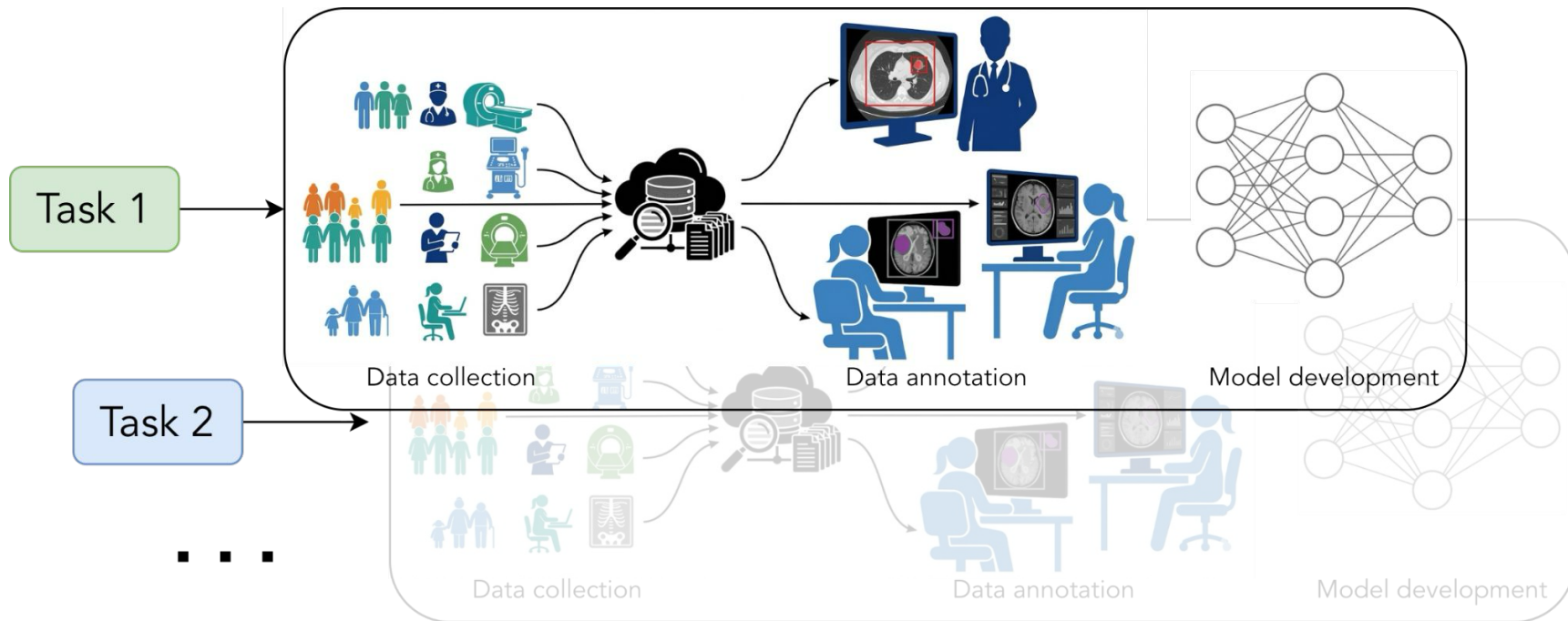
Build a *specialist* model?

(modality, hospital, scanner, ...)

Challenge: the annotation bottleneck



Challenge: the annotation bottleneck

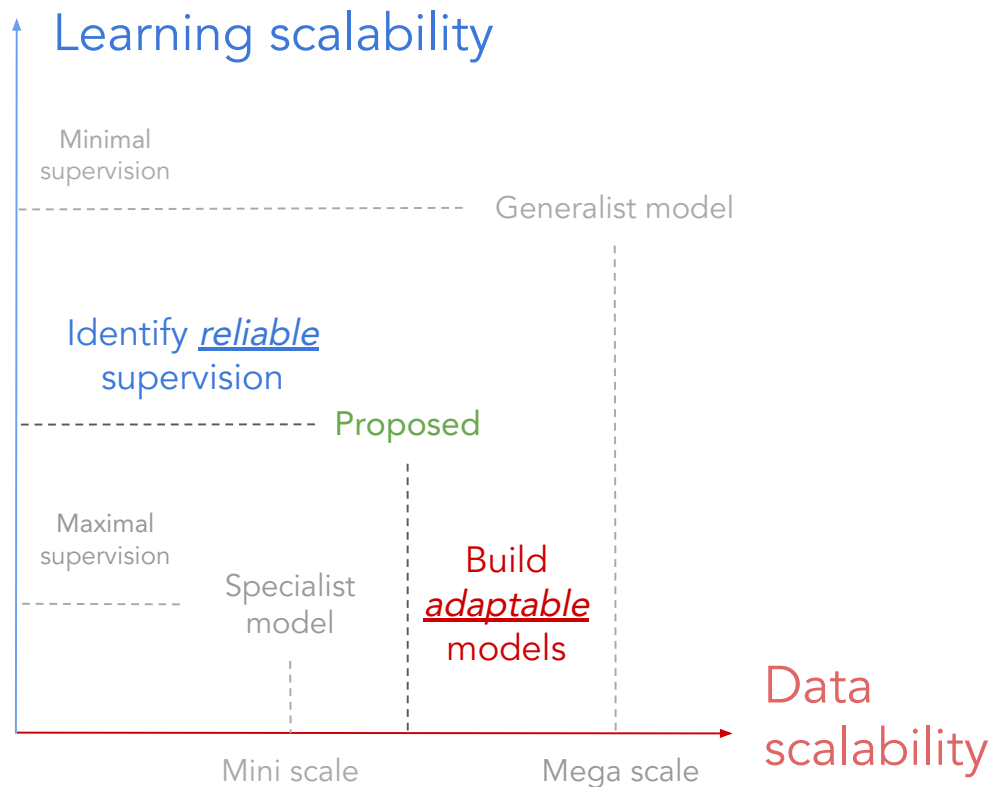


Motivation

Heterogeneous data

Distribution shift

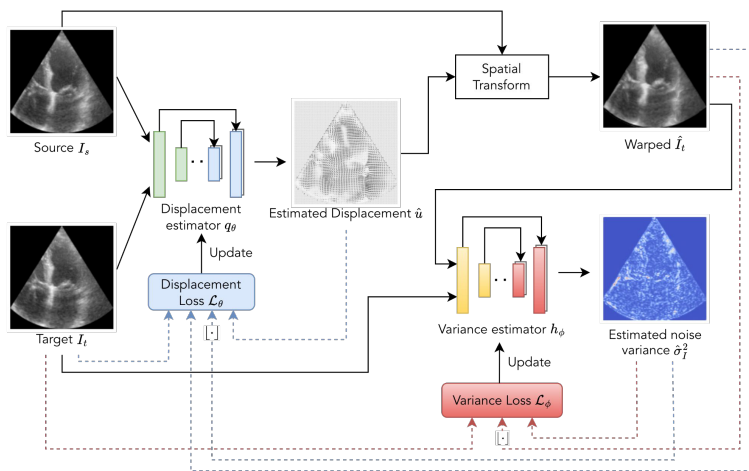
Laborious annotation



Develop *reliable* and *adaptable* models that generalize across domains without requiring extensive annotation.

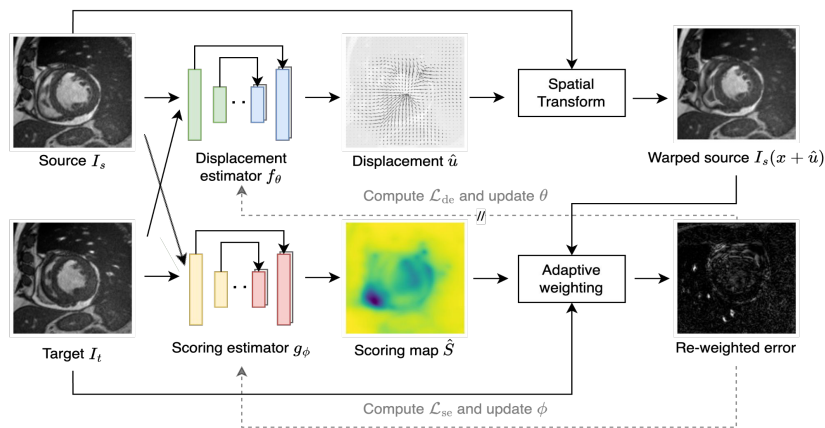
Identify reliable supervision for unsupervised registration

Heteroscedastic Uncertainty Estimation (MICCAI 2024)



Keywords: uncertainty quantification, aleatoric uncertainty, heteroscedastic noise

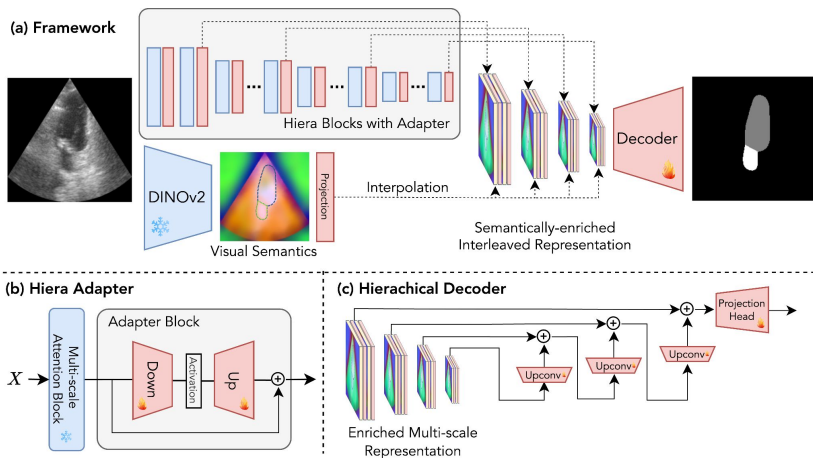
Adaptive Correspondence Scoring (ECCV 2024 *oral*)



Keywords: adaptive weighting, correspondence scoring, collaborative learning, training dynamics

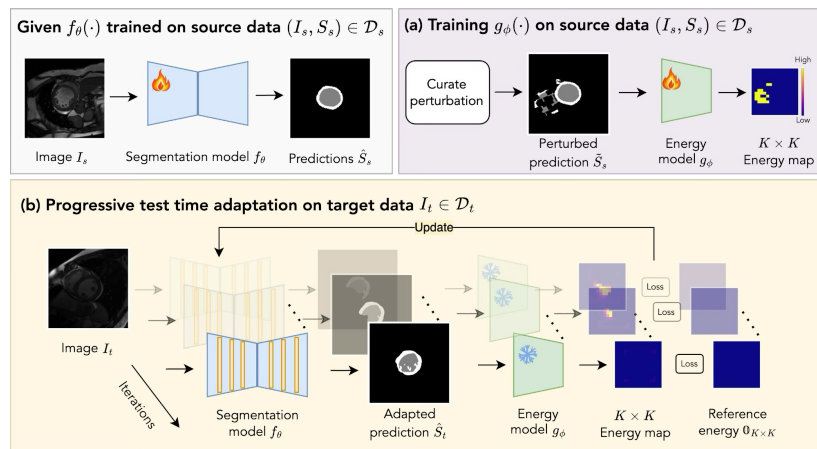
Build adaptable model for segmentation

Adaptation Vision Foundation Model for real-time Segmentation (MICCAI 2025 early accept)



Keywords: parameter-efficient finetuning, vision foundation models, adapters

Progressive Test Time Energy Adaptation for Segmentation (ICCV 2025 highlight)



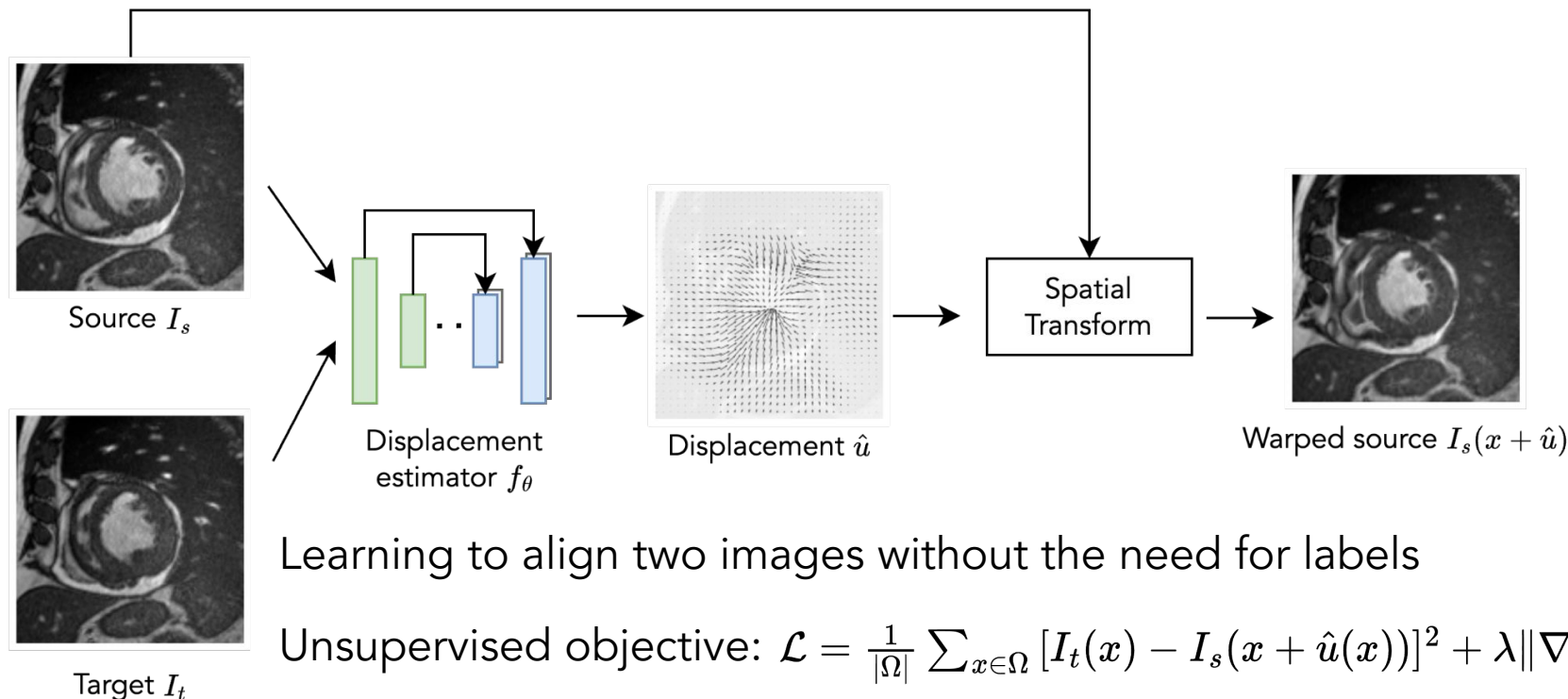
Keywords: test-time adaptation, region-based shape energy model

Overview

Develop *reliable* and *adaptable* models that generalize across domains without requiring extensive annotation.

- 
- 1) Identify *reliable* supervision signals for unsupervised registration.

Unsupervised Image Registration



Learning to align two images without the need for labels

Unsupervised objective: $\mathcal{L} = \frac{1}{|\Omega|} \sum_{x \in \Omega} [I_t(x) - I_s(x + \hat{u}(x))]^2 + \lambda \|\nabla \hat{u}(x)\|^2$

A probabilistic framework for unsupervised registration

Displacement (r.v.) Target image (r.v.) Source image (const.)

↓ ↓ ↓

Objective: $u^* = \arg \max_u p(u|I_t; I_s) = \arg \max_u p(I_t|u; I_s)p(u)$

$$u^* = \arg \min_u \underbrace{-\log p(I_t|u; I_s)}_{\text{Likelihood}} - \underbrace{\log p(u)}_{\text{Prior}}$$

A probabilistic framework for unsupervised registration

$$\text{Objective: } u^* = \arg \max_u p(u|I_t; I_s) = \arg \max_u p(I_t|u; I_s)p(u)$$

$$u^* = \arg \min_u \underbrace{-\log p(I_t|u; I_s)}_{\text{Likelihood}} - \underbrace{\log p(u)}_{\text{Prior}}$$

More expressive likelihood

- Heteroscedastic Uncertainty Modeling (*MICCAI 2024*)
- Adaptive Correspondence Scoring (*ECCV 2024 Oral*)

More plausible prior

- Biomechanics-informed Regularization (*STACOM 2022 Ora*)

A probabilistic framework for unsupervised registration

$$\text{Objective: } u^* = \arg \max_u p(u | I_t; I_s) = \arg \max_u p(I_t | u; I_s) p(u)$$

$$u^* = \arg \min_u \underbrace{-\log p(I_t|u; I_s)}_{\text{Likelihood}} - \underbrace{\log p(u)}_{\text{Prior}}$$

More expressive likelihood

- Heteroscedastic Uncertainty Modeling (MICCAI 2024)
- Adaptive Correspondence Scoring (ECCV 2024 Oral)

Prior



Project page

Heteroscedastic Uncertainty Estimation Framework for Unsupervised Registration

Xiaoran Zhang*, Daniel H. Pak*, Shawn S. Ahn, Xiaoxiao Li, Chenyu You, Lawrence H. Staib,
Albert J. Sinusas, Alex Wong, James S. Duncan

Yale University University of British Columbia

*Equal contribution

Noise is heteroscedastic in medical images

SPACE VARIANT ULTRASOUND FREQUENCY COMPOUNDING BASED ON NOISE CHARACTERISTICS

Medical Image Analysis 20 (2015) 184–197

*IBM Haifa Research
Haifa, Israel; and

(Re

Abstract—Ultrasound phenomenon caused images essentially bl the image and handl are acquired in diffel have been based on statistical and physic acoustic frequencies. understanding of th various noise sourc yael@il.ibm.com)



Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media



Spatially variant noise estimation in MRI: A homomorphic approach

Santiago Aja-Fernández^{a,*}, Tomasz Pięciak^b, Gonzalo Vegas-Sánchez-Ferrero^c

^aLPI, ETSI Telecomunicación, Universidad de Valladolid, Spain

^bAGH University of Science and Technology, Krakow, Poland

^cACIL, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA



Key Words: Frequen

ARTICLE INFO

Article history:

Received 9 July 2014

Received in revised form 11 November 2014

Accepted 15 November 2014

Available online 24 November 2014

Keywords:

Noise estimation

Non-stationarity noise

SENSE

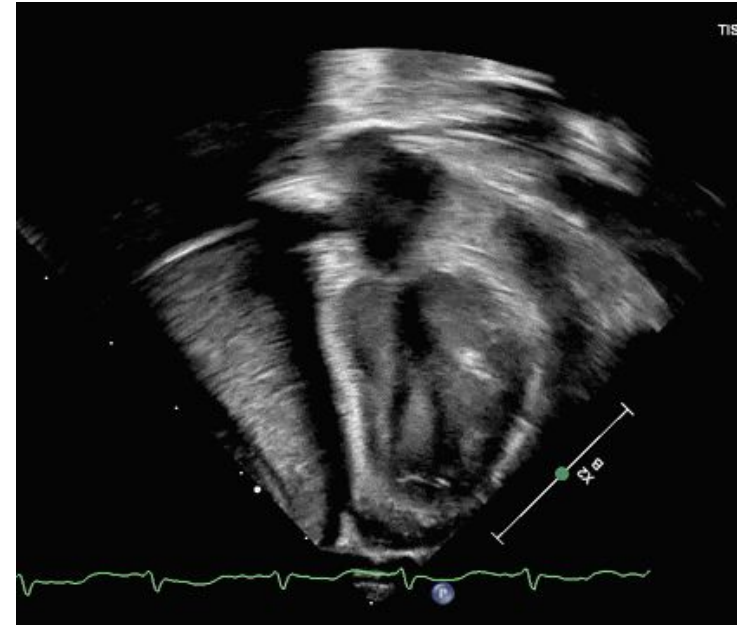
Parallel imaging

Homomorphic filtering

ABSTRACT

The reliable estimation of noise characteristics in MRI is a task of great importance due to the influence of noise features in extensively used post-processing algorithms. Many methods have been proposed in the literature to retrieve noise features from the magnitude signal. However, most of them assume a stationary noise model, i.e., the features of noise do not vary with the position inside the image. This assumption does not hold when modern scanning techniques are considered, e.g., in the case of parallel reconstruction and intensity correction. Therefore, new noise estimators must be found to cope with non-stationary noise. Some methods have been recently proposed in the literature. However, they require multiple acquisitions or extra information which is usually not available (biophysical models, sensitivity of coils). In this work we overcome this drawback by proposing a new method that can accurately estimate the non-stationary parameters of noise from just a single magnitude image. In the derivation, we considered the noise to follow a non-stationary Rician distribution, since it is the most common model in real acquisitions (e.g., SENSE reconstruction), though it can be easily generalized to other models. The proposed approach makes use of a homomorphic separation of the spatially variant noise in two terms: a stationary noise term and one low frequency signal that correspond to the x -dependent variance of noise. The non-stationary variance of noise is then estimated by a low pass filtering with a Rician bias correction. Results in real and synthetic experiments evidence the better performance and the lowest error variance of the proposed methodology when compared to the state-of-the-art methods.

© 2014 Elsevier B.V. All rights reserved.



[1] Erez, Yael, Yoav Y. Schechner, and Dan Adam. "Space variant ultrasound frequency compounding based on noise characteristics." *Ultrasound in medicine & biology* 34.6 (2008): 981-1000.

[2] Aja-Fernandez, Santiago, Tomasz Pie, and Gonzalo Vegas-Sanchez-Ferrero. "Spatially variant noise estimation in MRI: A homomorphic approach." *Medical image analysis* 20.1 (2015): 184-197.

But past works assume noise is homoscedastic (uniform)

$$u^* = \arg \min_u -\log p(I_t|u; I_s) - \log p(u)$$

Homoscedastic (uniform) Gaussian

$$p(I_t|u; I_s) = \mathcal{N}(I_t; I_s \circ u, \sigma_I^2 \mathbb{I})$$

Smoothness prior

$$p(u) \sim \exp(-\lambda \|\nabla u\|^2)$$

But past works assume noise is homoscedastic (uniform)

$$u^* = \arg \min_u -\log p(I_t|u; I_s) - \log p(u)$$

Homoscedastic (uniform) Gaussian

Smoothness prior

$$p(I_t|u; I_s) = \mathcal{N}(I_t; I_s \circ u, \sigma_I^2 \mathbb{I})$$

$$p(u) \sim \exp(-\lambda \|\nabla u\|^2)$$

σ_I^2 is assumed to be uniform
(const. w.r.t. image space)

$$\mathcal{L} = \frac{1}{|\Omega|} \sum_{x \in \Omega} \underbrace{[I_t(x) - I_s(x + \hat{u}(x))]^2}_{\text{Data term}} + \underbrace{\lambda \|\nabla \hat{u}(x)\|^2}_{\text{Regularization term}}$$

Data term

Regularization term

But past works assume noise is homoscedastic (uniform)

$$u^* = \arg \min_u -\log p(I_t|u; I_s) - \log p(u)$$

Homoscedastic (uniform) Gaussian $\swarrow$ Smoothness prior $\searrow$

$$p(I_t|u; I_s) = \mathcal{N}(I_t; I_s \circ u, \sigma_I^2 \mathbb{I}) \quad p(u) \sim \exp(-\lambda \|\nabla u\|^2)$$

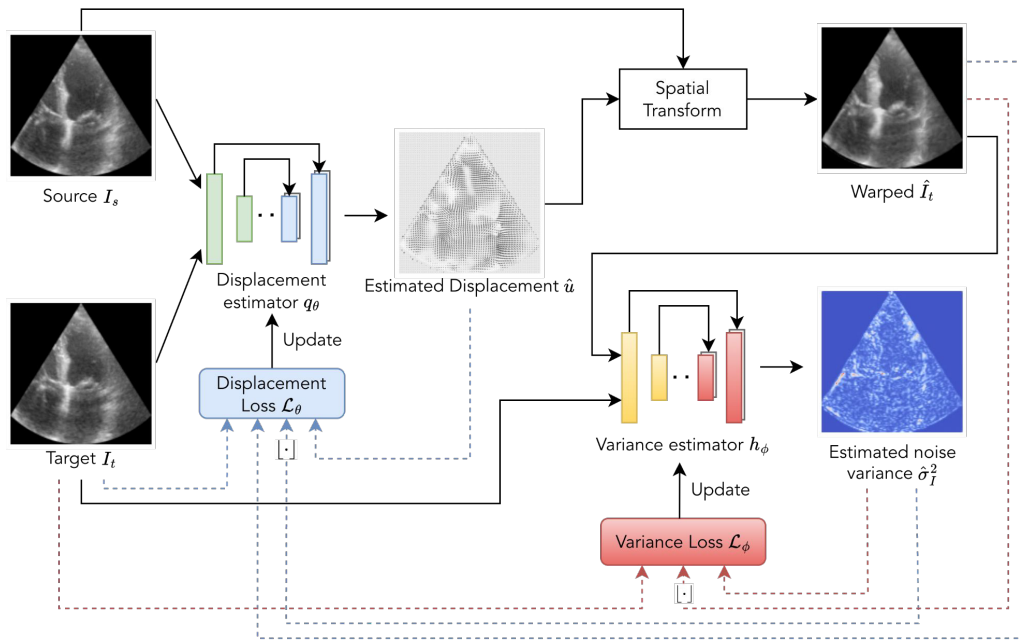
σ_I^2 is assumed to be uniform
(const. w.r.t. image space) $\swarrow$ $\searrow$

$$\mathcal{L} = \frac{1}{|\Omega|} \sum_{x \in \Omega} \underbrace{[I_t(x) - I_s(x + \hat{u}(x))]^2}_{\text{Data term}} + \underbrace{\lambda \|\nabla \hat{u}(x)\|^2}_{\text{Regularization term}}$$

We extend it to be heteroscedastic (non-uniform) and input-dependent:

$$\sigma_I^2 \rightarrow \sigma_I^2(x)$$

Method



Collaborative learning with displacement and heteroscedastic noise variance estimation

Displacement estimator loss:

$$\mathcal{L}_\theta = \frac{1}{|\Omega|} \sum_{x \in \Omega} \left[\tau \left[\left(\frac{I_t(x)}{[\hat{\sigma}_I(x)]} \right)^{2\gamma} \right] [I_t(x) - \hat{I}_t(x)]_2^2 + \lambda \|\nabla \hat{u}(x)\|^2 \right]$$

Variance estimator loss:

$$\mathcal{L}_\phi = \frac{1}{|\Omega|} \sum_{x \in \Omega} \left[[\hat{\sigma}_I(x)^{2\beta}] \left(\frac{1}{\hat{\sigma}_I(x)^2} [I_t(x) - \hat{I}_t(x)]_2^2 + \log \hat{\sigma}_I(x)^2 \right) \right]$$

Method

$$\mathcal{L}_\theta = \frac{1}{|\Omega|} \sum_{x \in \Omega} \left[\mathcal{T} \left[\left(\frac{I_t(x)}{[\hat{\sigma}_I(x)]} \right)^{2\gamma} \right] [I_t(x) - \hat{I}_t(x)]_2^2 + \lambda \|\nabla \hat{u}(x)\|^2 \right]$$

γ – exponentiated signal-to-noise (SNR) weighting

A relative weighting instead of an absolute weighting

γ reflects the influence of current noise variance estimate

$$\text{Weighting} \sim \frac{1}{\hat{\sigma}_I(x)} \quad \rightarrow \quad \text{Weighting} \sim \frac{I_f(x)}{\hat{\sigma}_I(x)}$$

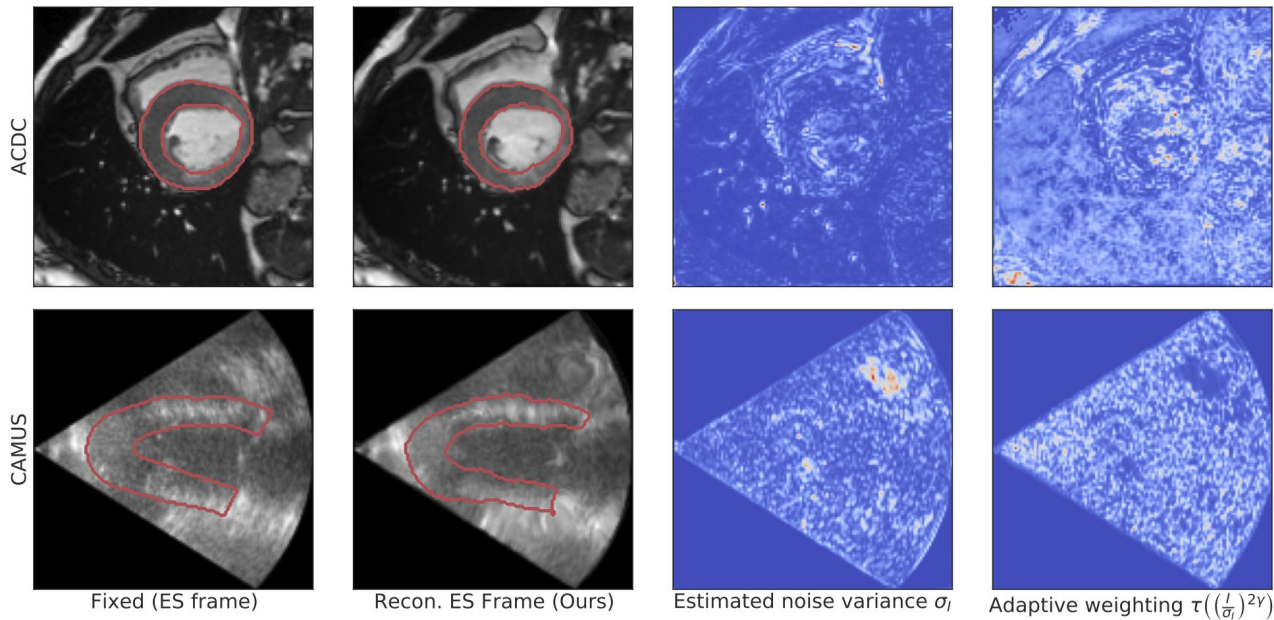
beta-NLL^[1] (absolute)

Ours (relative)

- $\gamma = 0$, loss reduces to MSE
- $\gamma = 1$, SNR weighting
- We select $\gamma = 0.5$ from experiments

[1] Seitzer, Maximilian, et al. "On the Pitfalls of Heteroscedastic Uncertainty Estimation with Probabilistic Neural Networks." *International Conference on Learning Representations*. 2022.

Result



		ACDC [4]			CAMUS [14]		
		DSC ↑	HD ↓	ASD ↓	DSC ↑	HD ↓	ASD ↓
CNN	Undeformed	47.98	7.91	2.32	66.77	10.87	2.61
	Elastix [13]	77.26	4.95	1.28	80.18	10.02	1.81
	vxm (NCC) [3]	78.55	4.94	1.29	77.01	10.23	1.89
	vxm (MI) [3]	78.04	5.25	1.35	78.18	9.83	1.99
	vxm (MSE) [3] †	80.20	4.64	1.24	81.76	8.93	1.70
	NLL [12]	76.49	5.46	1.45	75.24	11.05	2.20
	β -NLL [22]	78.74	5.07	1.33	79.75	9.39	1.93
	AdaFrame [27]	66.38	5.80	1.67	77.88	10.54	1.93
	AdaReg [28]	78.75	5.13	1.33	79.31	9.78	1.88
	Ours	80.73	4.57	1.21	81.96	8.80	1.66
Transformer	tsm (NCC) [5]	73.77	6.64	1.12	73.03	11.87	1.70
	tsm (MI) [5]	73.57	6.57	1.11	74.83	11.94	1.83
	tsm (MSE) [5] †	76.94	5.51	1.30	79.24	10.30	1.79
	NLL [12]	73.12	7.22	1.27	75.08	11.60	1.79
	β -NLL [22]	75.74	6.12	1.29	77.39	10.99	1.86
	AdaFrame [27]	67.95	5.72	1.59	78.06	9.86	1.91
	AdaReg [28]	76.22	5.68	1.29	78.12	10.62	1.84
	Ours	78.12	5.04	1.26	80.38	9.86	1.72

Our designed adaptive weighting map which reflects reliable correspondence is handcrafted from heteroscedastic noise variance estimate.

Result



Can we learn the adaptive weighting that reflects reliable correspondence directly from data?

	ACDC [4]			CAMUS [14]		
	DSC ↑	HD ↓	ASD ↓	DSC ↑	HD ↓	ASD ↓
Undeformed	47.98	7.91	2.32	66.77	10.87	2.61
Elastix [13]	77.26	4.95	1.28	80.18	10.02	1.81
vxm (NCC) [3]	78.55	4.94	1.29	77.01	10.23	1.89
vxm (MI) [3]	78.04	5.25	1.35	78.18	9.83	1.99
vxm (MSE) [3] †	80.20	4.64	1.24	81.76	8.93	1.70
NLL [12]	76.49	5.46	1.45	75.24	11.05	2.20
β-NLL [22]	78.74	5.07	1.35	79.75	9.39	1.93
AdaFrame [27]	76.95	5.72	1.59	78.06	9.86	1.91
AdaReg [28]	76.22	5.68	1.29	78.12	10.62	1.84
Ours	78.12	5.04	1.26	80.38	9.86	1.72

Our designed adaptive weighting map which reflects reliable correspondence is handcrafted from heteroscedastic noise variance estimate.



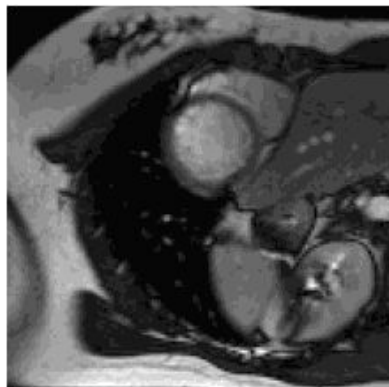
Adaptive Correspondence Scoring for Unsupervised Medical Image Registration

Xiaoran Zhang, John C. Stendahl, Lawrence Staib, Albert J. Sinusas, Alex Wong, James S. Duncan

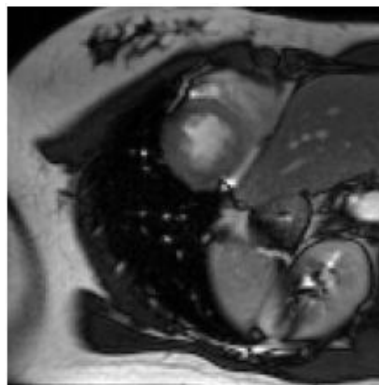
Yale University

Violations of basic assumptions

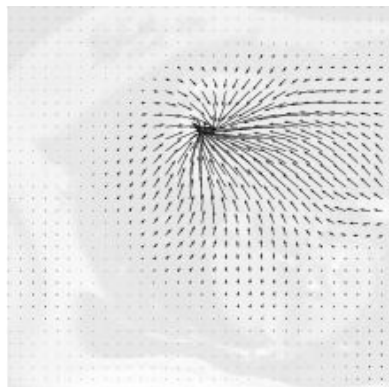
MRI video



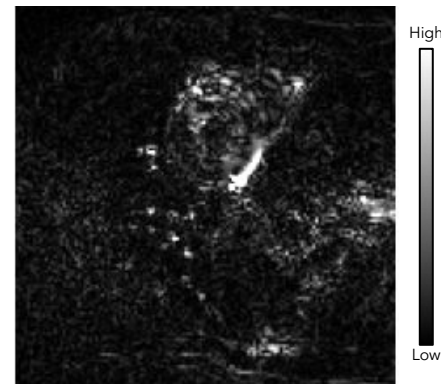
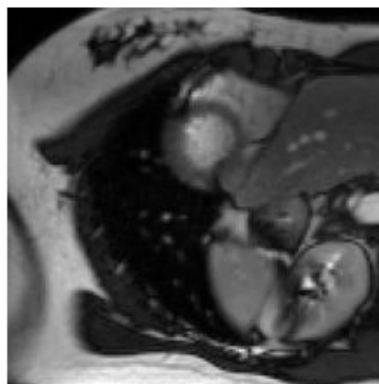
Target



Displacement between
first and last frame



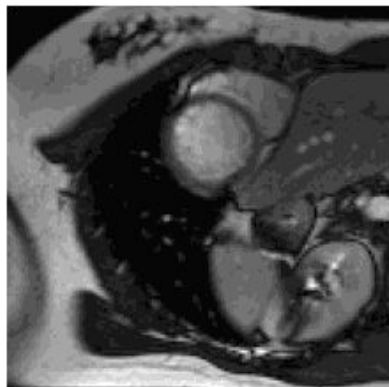
Reconstructed target



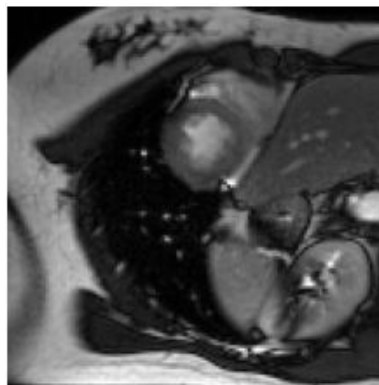
Error map

Violations of basic assumptions

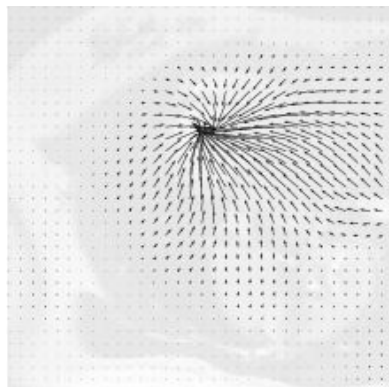
MRI video



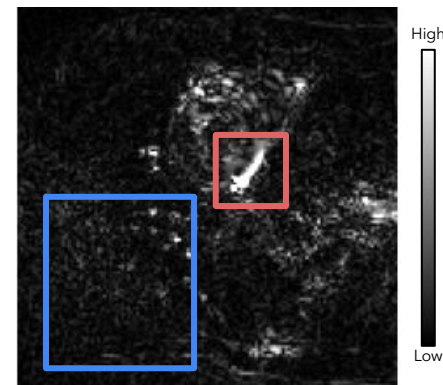
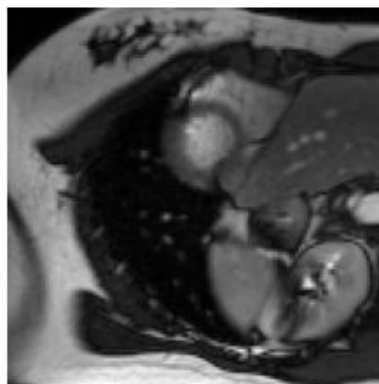
Target



Displacement between
first and last frame



Reconstructed target



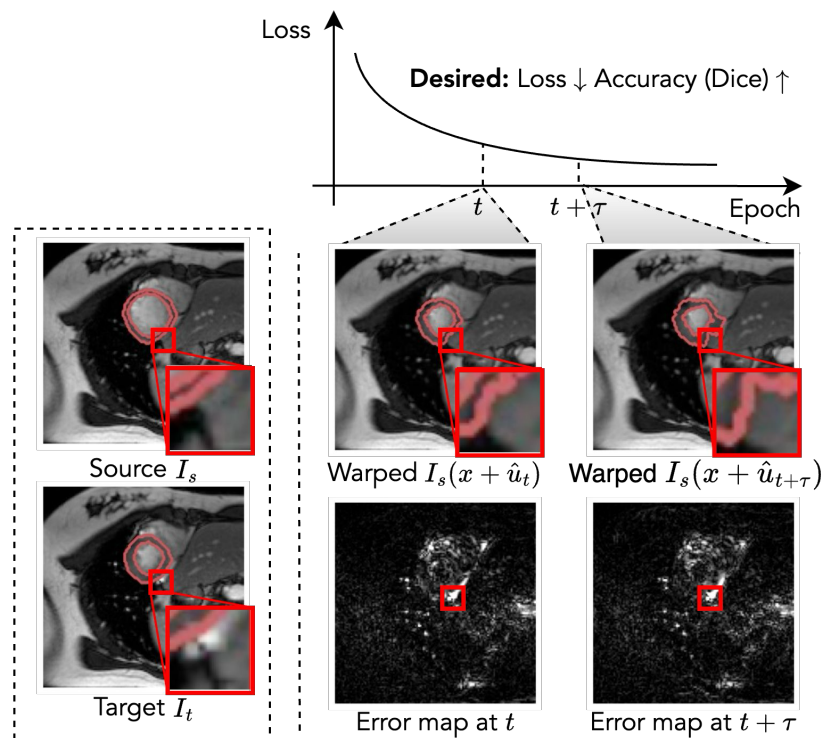
Error map

Compressive structural
visibility issue

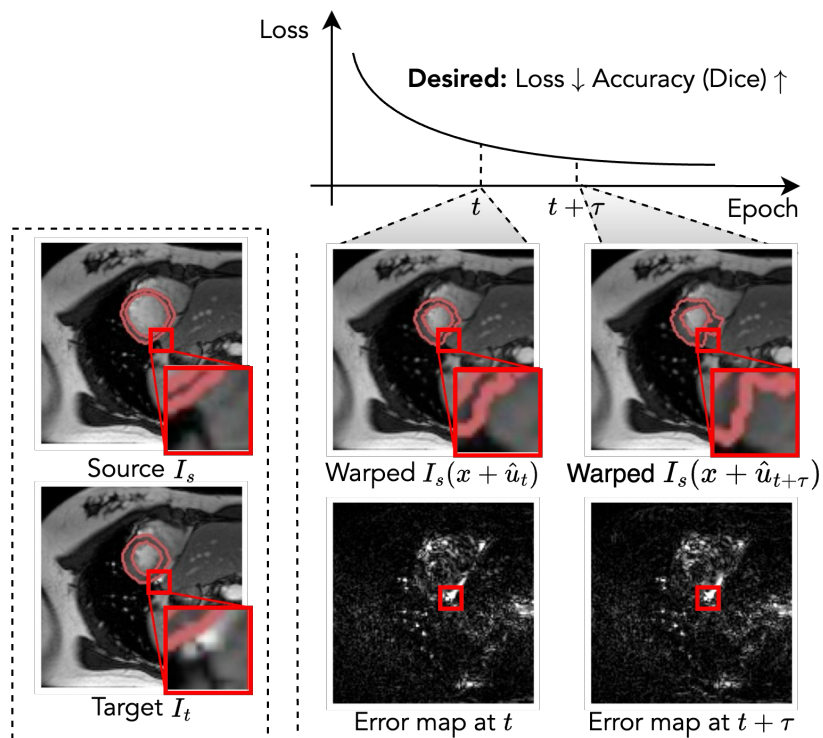
Heteroscedastic noise

... and many others

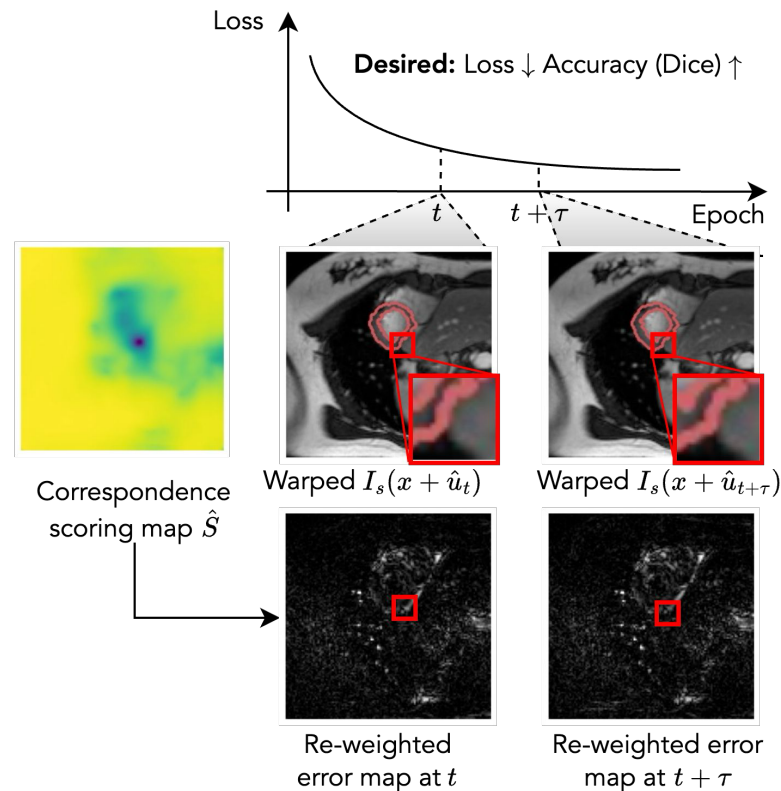
A closer look



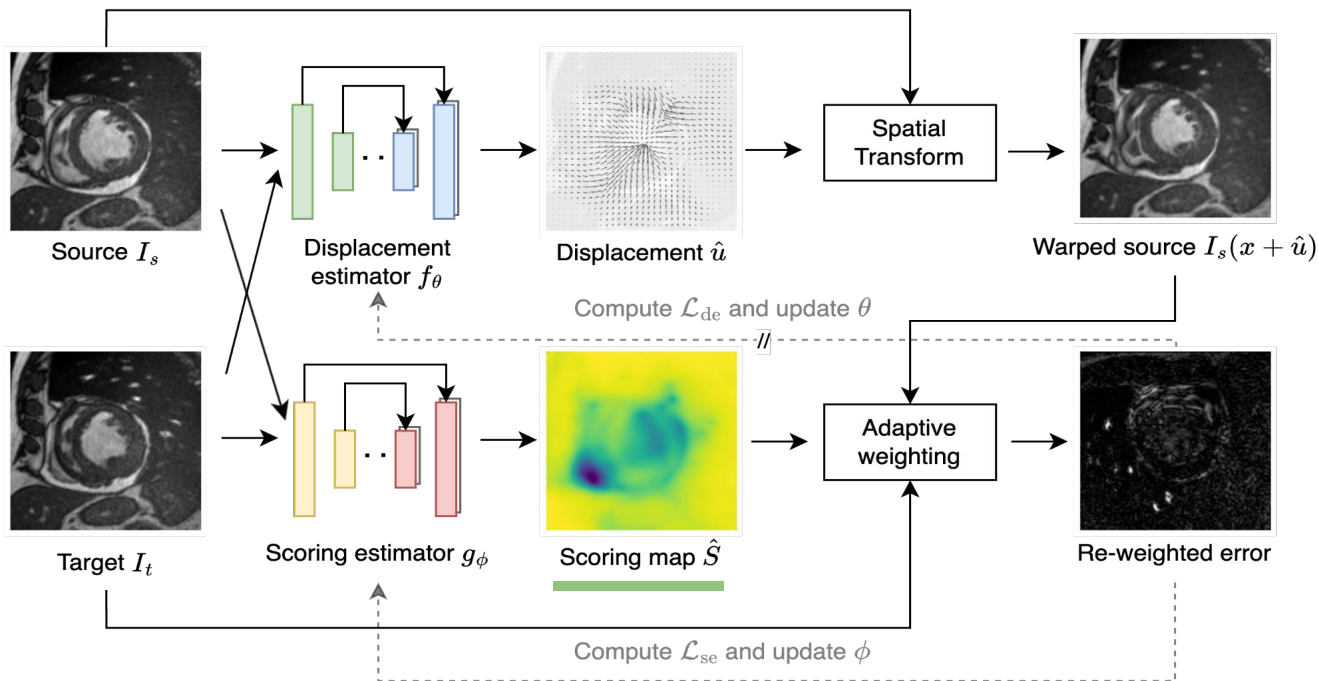
A closer look



Our approach:

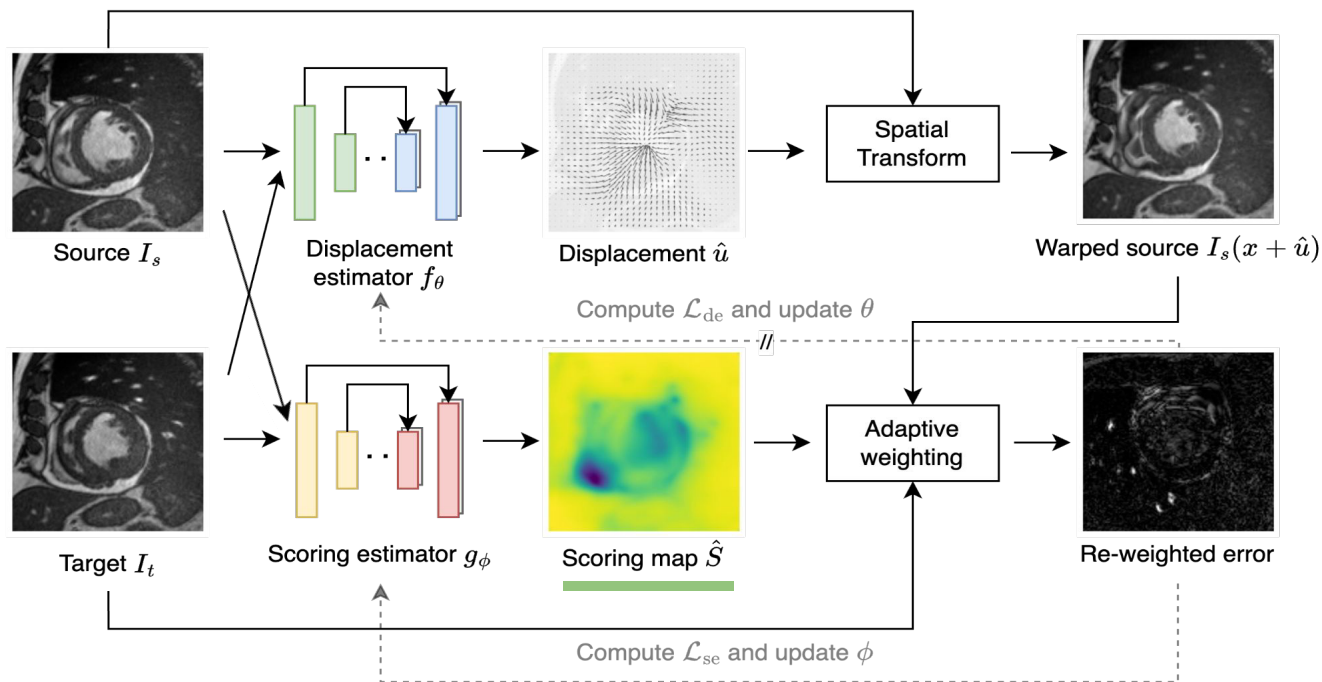


Method



Displacement estimator loss:
$$\frac{1}{|\Omega|} \sum_{x \in \Omega} [\hat{S}(x)] [I_t(x) - I_s(x + \hat{u}(x))]^2 + \lambda \|\nabla \hat{u}(x)\|^2$$

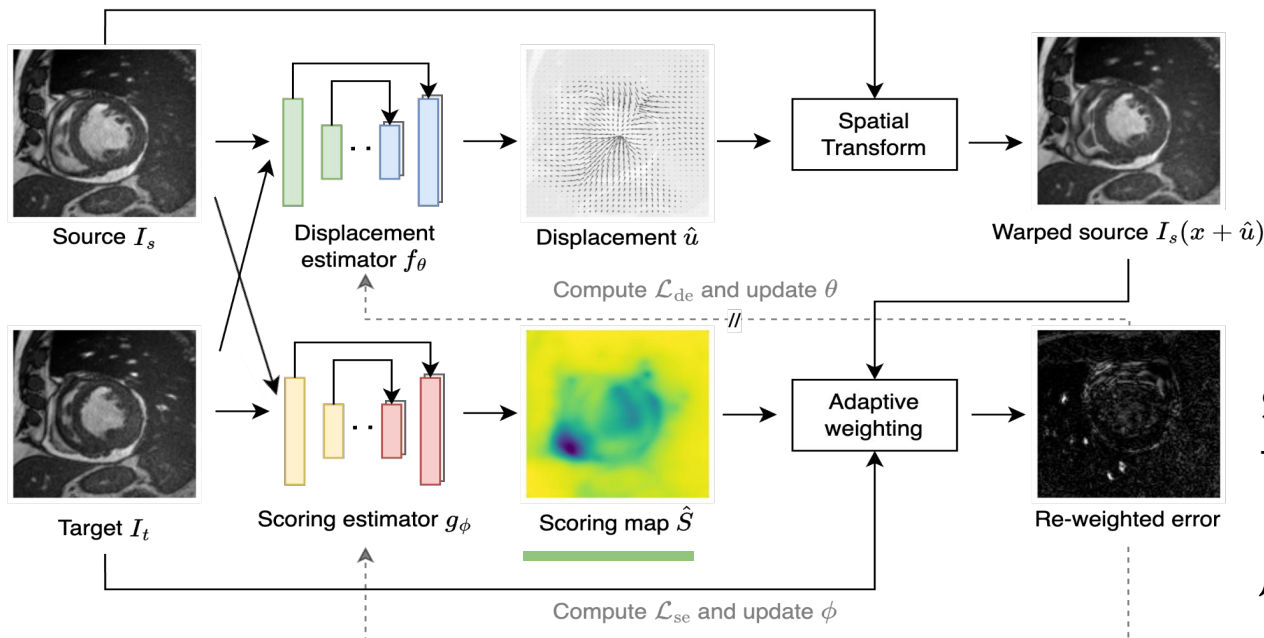
Method



But there is a trivial solution...

Unsupervised correspondence scoring loss: $\mathcal{L}_{ucs} = \frac{1}{|\Omega|} \sum_{x \in \Omega} \hat{S}(x) [I_t(x) - I_s(x + \lfloor \hat{u}(x) \rfloor)]^2$

Method



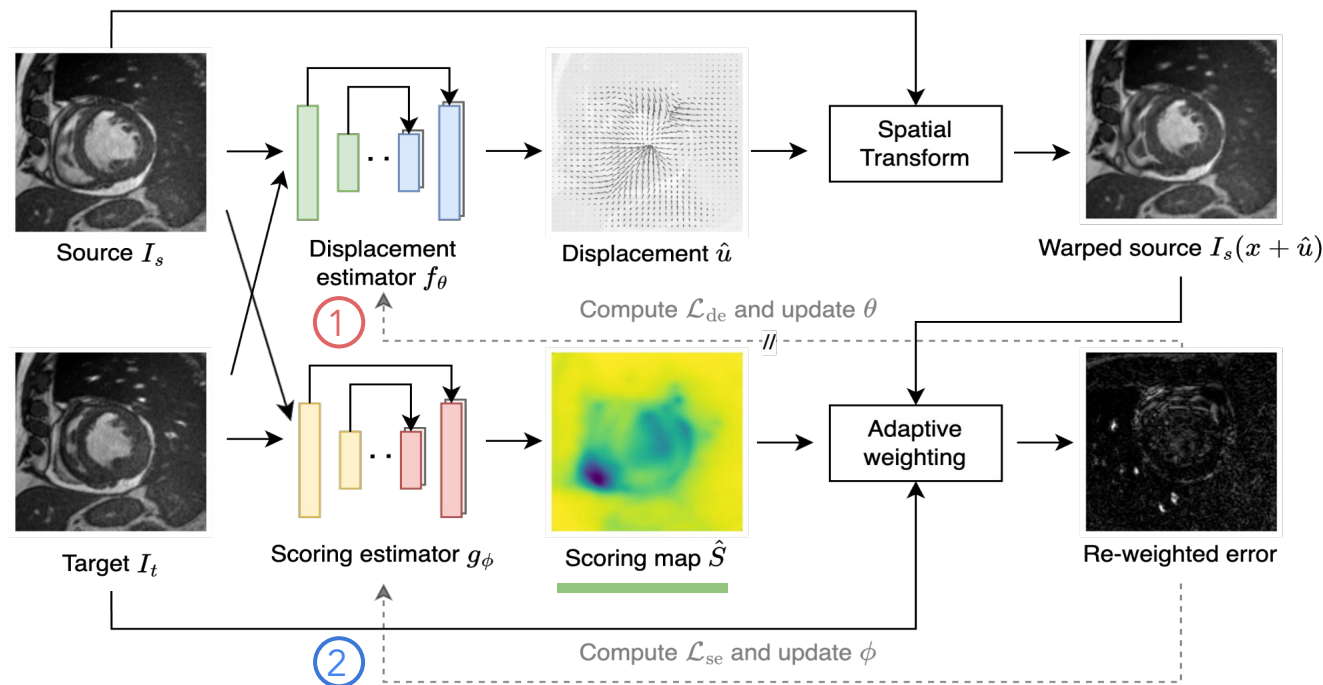
Avoid trivial solution:

$$\mathcal{L}_{\text{reg}} = \frac{1}{|\Omega|} \sum_{x \in \Omega} [1 - \hat{S}(x)]^2$$

Scores correspond to surfaces that are locally smooth:

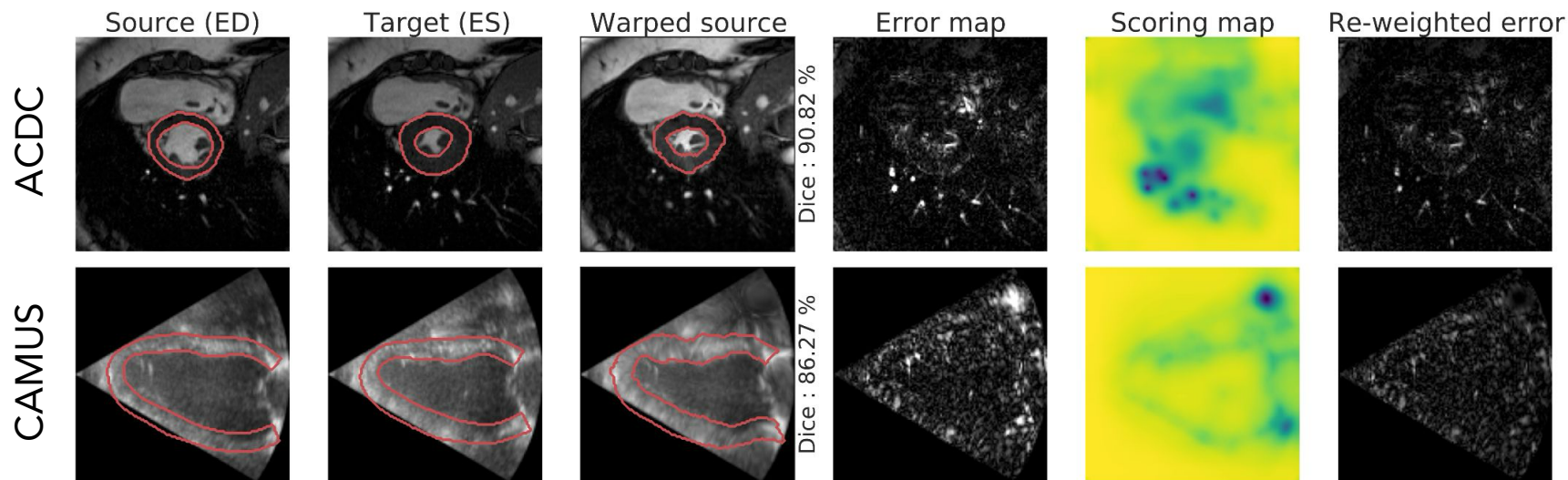
$$\mathcal{L}_{\text{smooth}} = m_T \frac{1}{|\Omega|} \sum_{x \in \Omega} \|\nabla \hat{S}(x)\|^2$$

Method



Scoring estimator loss: $\mathcal{L}_{se} = \mathcal{L}_{ucs} + \alpha \mathcal{L}_{reg} + \beta \mathcal{L}_{smooth}$

Results on ACDC and CAMUS



Our estimated scoring map identifies spurious error residuals and prevents parameter drift during training.

Results on ACDC and CAMUS

Quantitative evaluation

	ACDC			CAMUS		
	DSC ↑	HD ↓	ASD ↓	DSC ↑	HD ↓	ASD ↓
Undeformed	47.98	7.91	2.32	66.77	10.87	2.61
Elastix	77.26	4.95	1.28	80.18	10.02	1.81
Voxelmorph	79.48	4.79	1.27	81.50	8.72	1.74
NLL	76.49	5.46	1.45	75.24	11.05	2.20
β -NLL	78.74	5.07	1.33	79.75	9.39	1.93
AdaFrame	66.38	5.80	1.67	77.88	10.54	1.93
AdaReg	78.75	5.13	1.33	79.31	9.78	1.88
AdaCS (Ours)	80.50	4.69	1.23	81.74	8.55	1.72
Transmorph	76.94	5.51	1.30	79.24	10.30	1.79
NLL	73.12	7.22	1.27	75.08	11.60	1.79
β -NLL	75.74	6.12	1.29	77.39	10.99	1.86
AdaFrame	67.95	5.72	1.59	78.06	9.86	1.91
AdaReg	76.22	5.68	1.29	78.12	10.62	1.84
AdaCS (Ours)	78.39	5.40	1.32	79.64	9.85	1.79
Diffusemorph	67.38	5.80	1.67	75.23	9.80	2.07
NLL	66.24	5.84	1.73	74.78	10.62	2.15
β -NLL	66.31	5.93	1.74	73.27	9.85	2.25
AdaFrame	59.78	6.46	1.93	75.04	10.41	2.10
AdaReg	69.41	6.25	1.78	74.36	10.66	2.21
AdaCS (Ours)	72.09	5.35	1.53	77.65	9.82	1.99

Comparison to robust losses

	ACDC			CAMUS		
	DSC ↑	HD ↓	ASD ↓	DSC ↑	HD ↓	ASD ↓
NCC	78.55	4.94	1.29	77.01	10.23	1.89
MI	78.04	5.25	1.35	78.18	9.83	1.99
TBL	79.31	4.64	1.23	81.18	8.91	1.72
MAE	78.27	5.36	1.43	78.59	10.23	1.97
MSE	79.48	4.79	1.27	81.50	8.72	1.74
AdaCS	80.50	4.69	1.23	81.74	8.55	1.72
NCC	73.77	6.64	1.12	73.03	11.87	1.70
MI	73.57	6.57	1.11	74.83	11.94	1.83
TBL	78.23	5.11	1.27	79.12	9.75	1.84
MAE	74.30	6.36	1.28	75.96	11.35	1.89
MSE	76.94	5.51	1.30	79.24	10.30	1.79
AdaCS	78.39	5.40	1.32	79.64	9.85	1.79
NCC	70.25	5.29	1.58	75.67	10.75	2.06
MI	71.16	5.40	1.56	76.19	10.09	2.16
TBL	69.12	5.73	1.63	76.05	9.54	2.06
MAE	66.30	5.75	1.71	77.30	10.36	2.09
MSE	67.38	5.80	1.67	75.23	9.80	2.07
AdaCS	72.09	5.35	1.53	77.65	9.82	1.99

Smoothness

	ACDC			CAMUS		
	DSC ↑	$ J_{\hat{a}} \leq 0 \downarrow T_{\text{train}}$		DSC ↑	$ J_{\hat{a}} \leq 0 \downarrow T_{\text{train}}$	
Voxelmorph	79.48 ± 9.23	0.29	0.26	81.50 ± 5.58	0.60	0.26
AdaCS (Ours)	80.50 ± 8.58	0.22	0.43	81.74 ± 5.36	0.30	0.43
Transmorph	76.94 ± 8.93	0.76	0.60	79.24 ± 6.06	1.41	0.59
AdaCS (Ours)	78.39 ± 9.06	0.57	0.87	79.64 ± 6.37	0.70	0.70
Diffusemorph	67.38 ± 15.65	0.05	1.08	75.23 ± 8.71	0.05	1.06
AdaCS (Ours)	72.09 ± 13.60	0.06	1.86	77.65 ± 7.64	0.08	1.91

Ablation

	Loss		ACDC			CAMUS		
	$\mathcal{L}_{\text{reg}}$	$\mathcal{L}_{\text{smooth}}$	DSC ↑	HD ↓	ASD ↓	DSC ↑	HD ↓	ASD ↓
vxm	✓	✗	80.24	4.64	1.23	81.58	8.89	1.74
	✓	✓	80.50	4.69	1.23	81.74	8.55	1.72
tsm	✓	✗	77.84	5.41	1.33	79.58	10.17	1.81
	✓	✓	78.39	5.40	1.32	79.64	9.85	1.79
dfm	✓	✗	71.62	5.56	1.58	77.32	9.71	2.00
	✓	✓	72.09	5.35	1.53	77.65	9.82	1.99

Our proposed approach consistently outperforms baselines in various architectures and datasets and produces reasonably smooth displacement.

Conclusion

- We identify the limitation of the widely used unsupervised training objective
- We address this by proposing an adaptive correspondence scoring framework during training
- Our proposed approach can be plugged-and-played into existing frameworks with no extra cost during inference



Project page



Paper

A probabilistic framework for unsupervised registration

$$\text{Objective: } u^* = \arg \max_u p(u | I_t; I_s) = \arg \max_u p(I_t | u; I_s) p(u)$$

$$u^* = \arg \min_u \underbrace{-\log p(I_t|u; I_s)}_{\text{Likelihood}} - \underbrace{\log p(u)}_{\text{Prior}}$$

More plausible prior

- Biomechanics-informed Regularization (STACOM 2022 Oral)

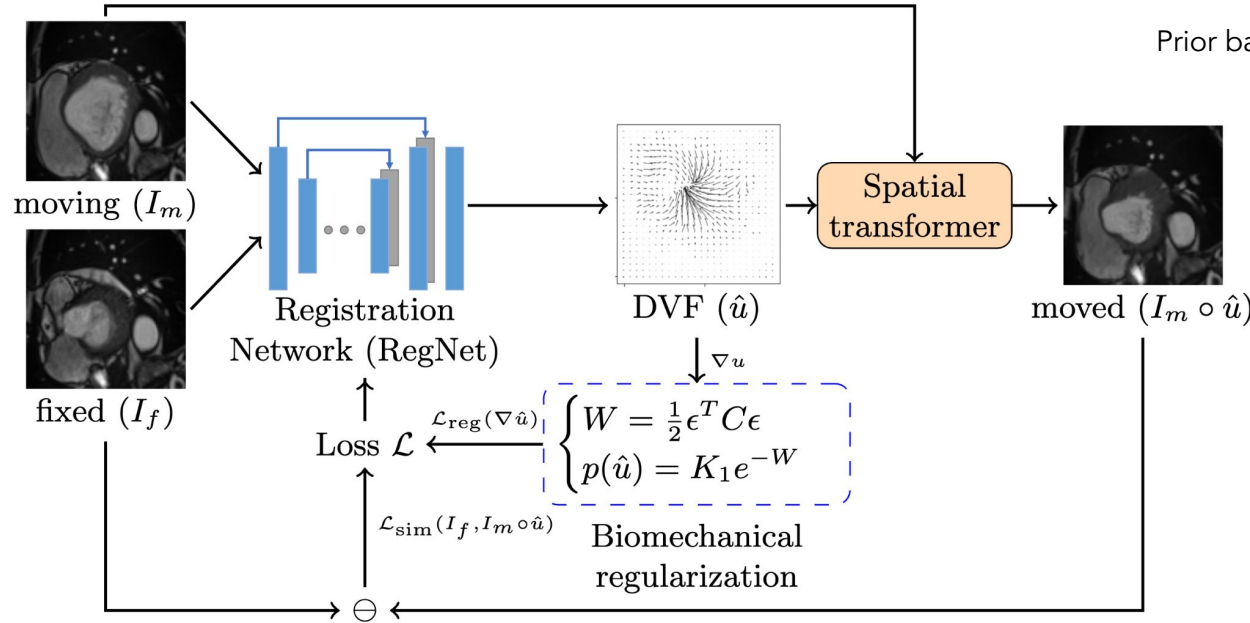
Yale

Learning Correspondences of Cardiac Motion from Images using Biomechanics-informed Modeling

Xiaoran Zhang, Chenyu You, Shawn Ahn, Juntang Zhuang, Lawrence Staib, James Duncan

Yale University

Method



$$u^* = \arg \min_u -\log p(I_t|u; I_s) - \log p(u)$$

Prior based on strain energy density

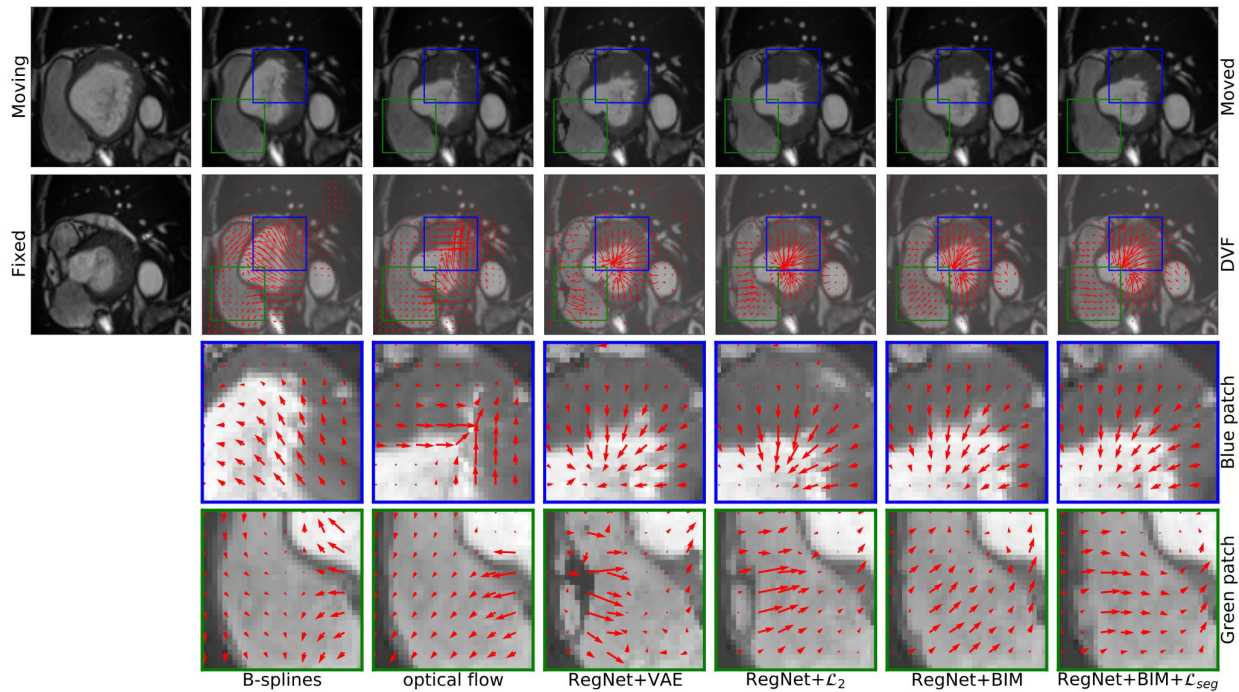
$$p(\hat{u}) = k_1 e^{-W}$$

$$W(x) = \frac{1}{2} \epsilon(x)^T C \epsilon(x)$$

$$\frac{1}{|\Omega|} \sum_{x \in \Omega} \|\epsilon(x)^T C \epsilon(x)\|_2$$

Biomechanics-informed
Regularization

Result



Quantitative evaluation

Method	Dice[%]↑	Jaccard[%]↑	HD[mm]↓	ASD[mm]↓
affine only	72.01	57.24	11.77	2.13
B-splines [13]	89.96	82.62	5.37	0.48
optical flow [14]	87.07	77.87	7.65	0.74
RegNet+VAE [6]	89.61	81.65	6.46	0.49
RegNet+ $\mathcal{L}_2$	89.60	81.68	8.00	0.54
RegNet+BIM (ours)	90.32	82.70	5.51	0.37
RegNet+BIM+$\mathcal{L}_{seg}$ (ours)	92.77	86.76	4.39	0.19
affine only	81.08	70.06	14.31	2.21
B-splines [13]	83.59	74.33	13.70	2.43
optical flow [14]	84.96	75.94	13.70	2.15
RegNet+VAE [6]	84.65	75.58	16.48	2.31
RegNet+ $\mathcal{L}_2$	84.80	75.73	14.75	2.25
RegNet+BIM (ours)	85.07	76.16	14.55	2.22
RegNet+BIM+$\mathcal{L}_{seg}$ (ours)	85.93	77.54	14.07	2.15
affine only	85.07	74.50	8.61	0.93
B-splines [13]	92.33	86.12	5.97	0.33
optical flow [14]	92.02	85.64	6.09	0.39
RegNet+VAE [6]	91.45	84.83	7.57	0.48
RegNet+ $\mathcal{L}_2$	89.92	82.22	7.77	0.59
RegNet+BIM (ours)	91.53	84.75	6.20	0.40
RegNet+BIM+$\mathcal{L}_{seg}$ (ours)	92.57	86.48	5.54	0.32

Smoothness evaluation

Method	$ \det(J(\hat{u})) - 1 $
B-splines [13]	0.2489 ± 0.2735
optical flow [14]	0.1283 ± 0.3432
RegNet+VAE [6]	0.0088 ± 0.0100
RegNet+ $\mathcal{L}_2$	0.0038 ± 0.0051
RegNet+BIM (ours)	0.0035 ± 0.0036
RegNet+BIM+ $\mathcal{L}_{seg}$ (ours)	0.0037 ± 0.0038

Overview

Develop *reliable* and *adaptable* models that generalize across domains without requiring extensive annotation.

1) Identify *reliable* supervision signals for unsupervised registration.

- I. More expressive likelihood
HeteroUnc, *MICCAI 2024*
AdaCS, *ECCV 2024*
- II. More plausible prior
BIM-reg, *STACOM 2022*

Overview

Develop *reliable* and *adaptable* models that generalize across domains without requiring extensive annotation.

1) Identify *reliable* supervision signals for unsupervised registration.

I. More expressive likelihood
HeteroUnc, MICCAI 2024
AdaCS, ECCV 2024

II. More plausible prior
BIM-reg, STACOM 2022

2) Build *adaptable* models for versatile medical data.

I. Training-time adaptation:
Adapt vision foundation model for ultrasound.



Adapting Vision Foundation Models for Real-time Ultrasound Image Segmentation

Xiaoran Zhang¹, Eric Z. Chen², Lin Zhao², Xiao Chen², Yikang Liu², Boris Maihe², James Duncan¹, Terrence Chen², Shanhui Sun²

1. Yale University 2. United Imaging Intelligence

Foundation models for segmentation - SAM series

SAM (2023):
segment anything
on images

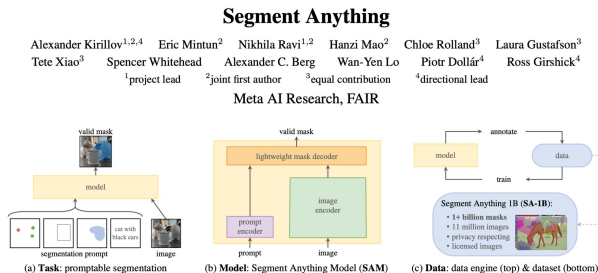
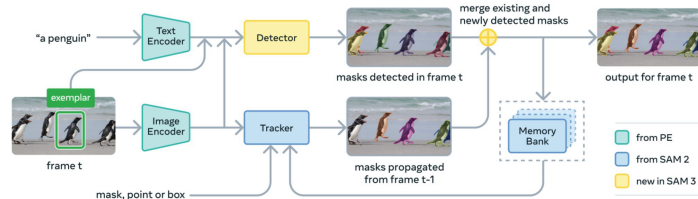
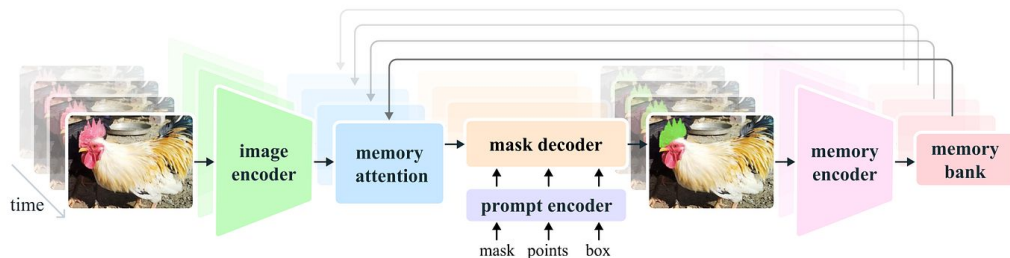


Figure 1: We aim to build a foundation model for segmentation by introducing three interconnected components: a promptable segmentation task, a segmentation model (SAM) that powers data annotation and enables zero-shot transfer to a range of tasks via prompt engineering, and a data engine for collecting SA-1B, our dataset of over 1 billion masks.

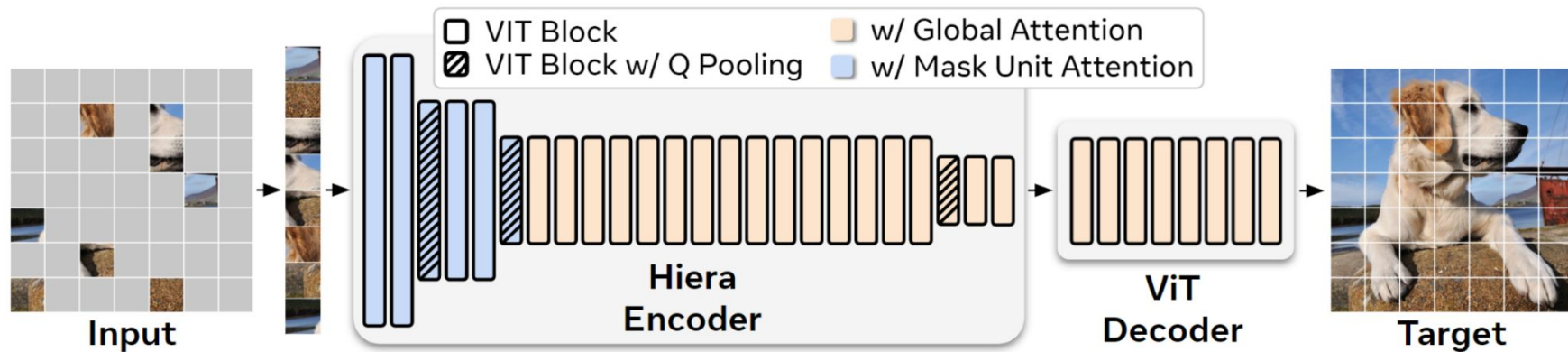
SAM3 (2025): segment
anything with concepts



SAM2 (2024): segment anything on
images and videos

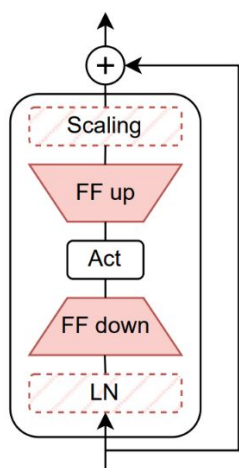


SAM2 image encoder: Hierarchical vision transformer

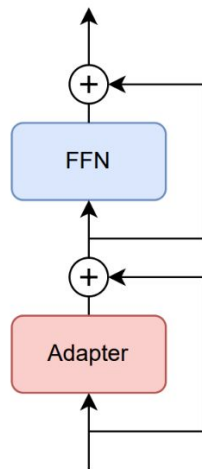


Hiera provides image features at multiple scales.

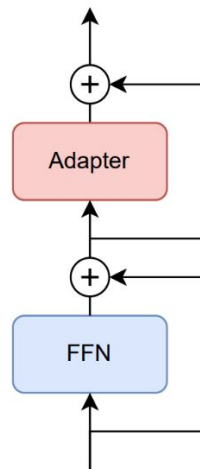
Parameter-efficient fine-tuning - Adapters



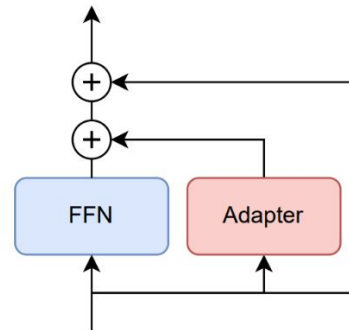
(a) Inner structure



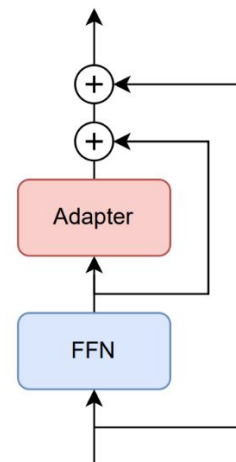
(b) Pre



(c) Post

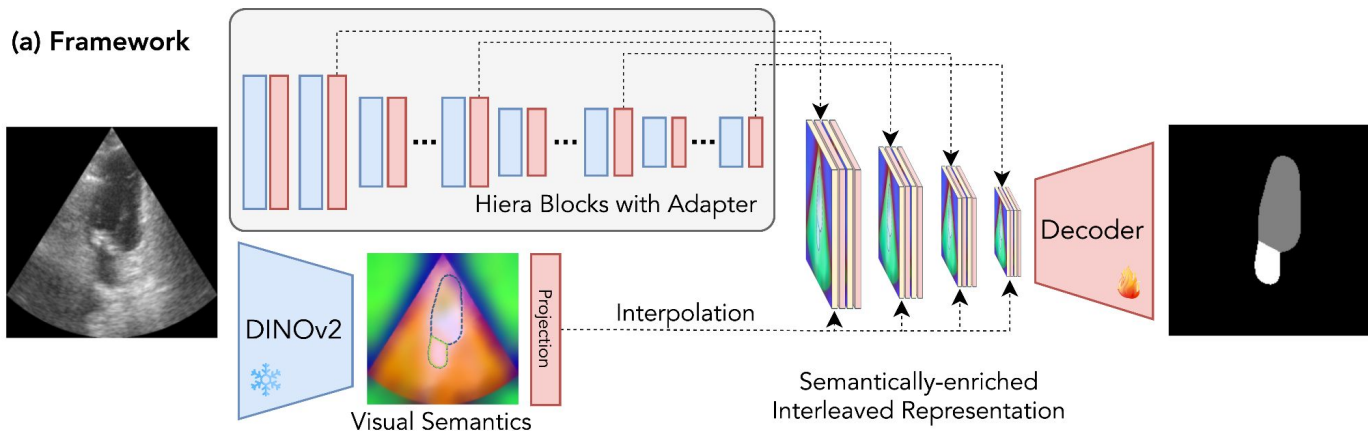


(d) Parallel



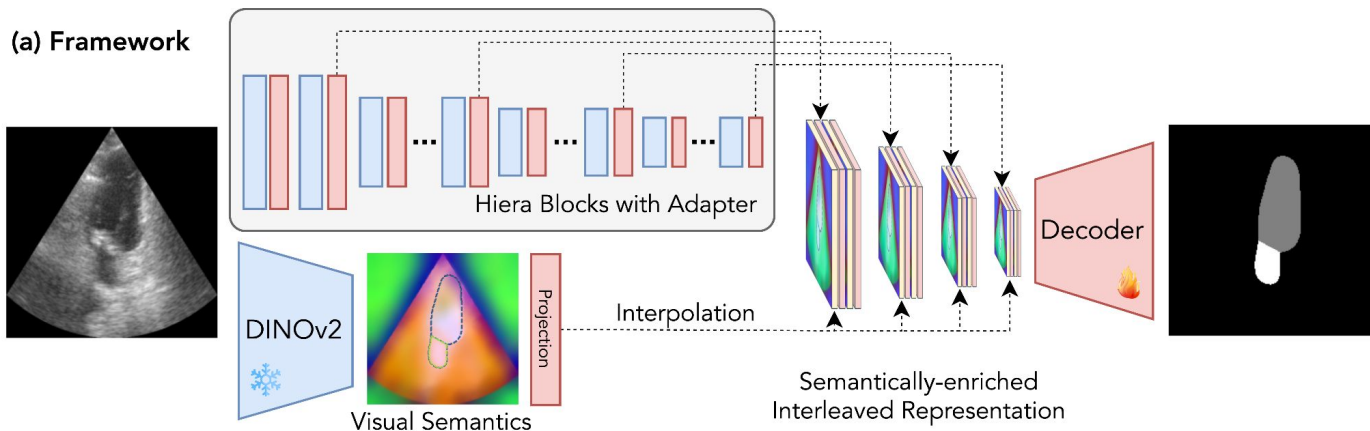
(e) Intermediate

Method

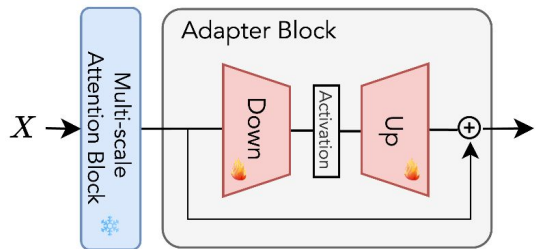


We adapt Hiera in SAM2 for multi-scale feature extraction and integrate DINOv2 image encoder to enhance visual feature representation, providing a richer and more robust understanding of ultrasound image

Method

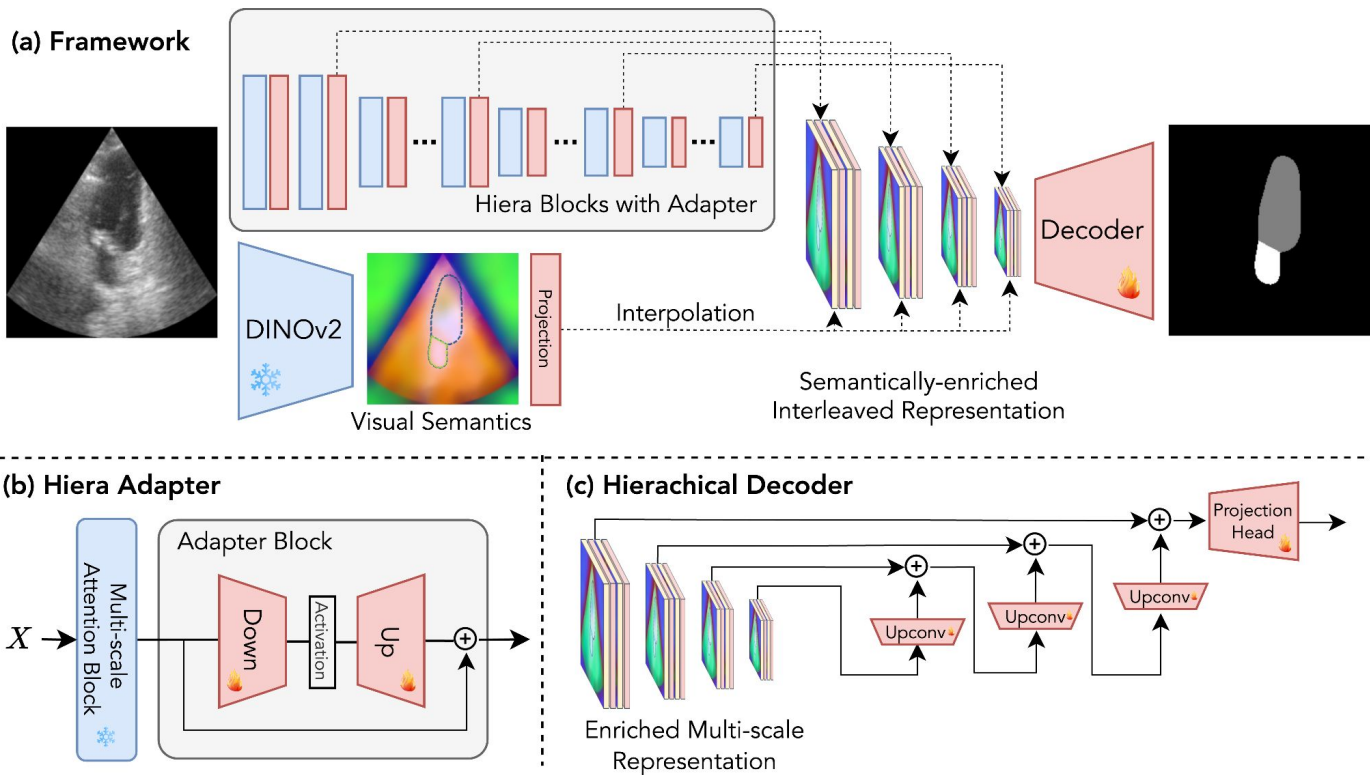


(b) Hiera Adapter

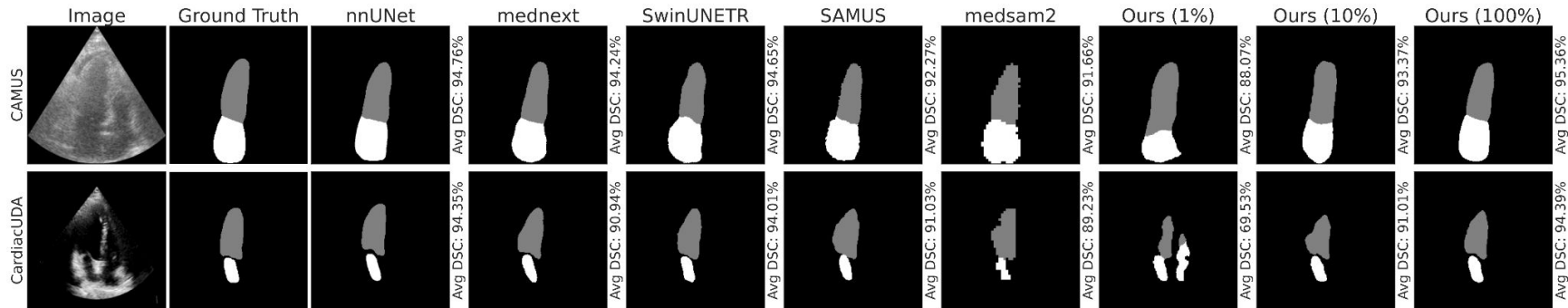


$$\text{Adapter}(X) = \sigma(XW_{\text{down}} + b_{\text{down}})W_{\text{up}} + b_{\text{up}} + X$$

Method



Data efficiency and adaptability on cardiac ultrasound

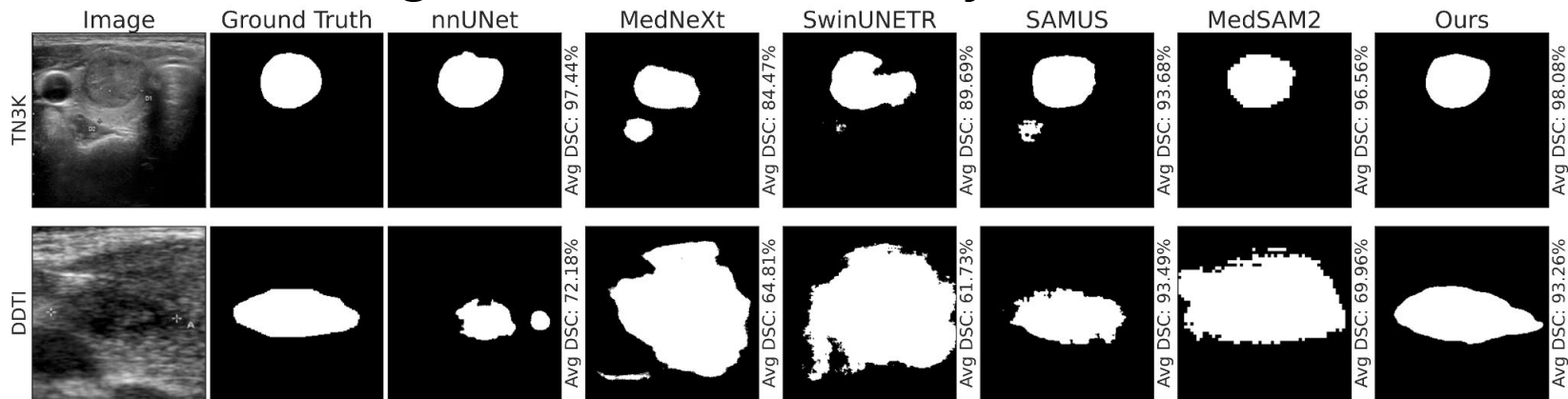


	CAMUS [20]							CardiacUDA [46]						
	1%		10%		100%			1%		10%		100%		
	DSC ↑	HD ↓	DSC ↑	HD ↓	DSC ↑	HD ↓		DSC ↑	HD ↓	DSC ↑	HD ↓	DSC ↑	HD ↓	
UNet [35]	59.49	51.51	80.14	26.36	87.70	15.05		34.65	41.04	72.83	26.24	81.69	20.78	
nnUNet [16]	67.02	54.22	<u>89.36</u>	<u>8.83</u>	<u>91.69</u>	6.35		35.48	58.94	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]	<u>75.00</u>	<u>26.29</u>	87.43	16.24	91.11	8.79		<u>63.65</u>	<u>23.91</u>	<u>82.10</u>	10.15	86.14	<u>7.70</u>	
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	<u>6.75</u>		65.84	23.57	82.38	<u>11.07</u>	<u>87.44</u>	6.45	

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	<u>93.31</u>	<u>87.61</u>	3.82	1.57
Ours	93.80	88.49	<u>4.80</u>	<u>1.90</u>

Our approach remains highly effective under limited supervision, significantly outperforming baselines when trained with only 1% and 10% of the training data.

Cross-dataset generalization on thyroid ultrasound

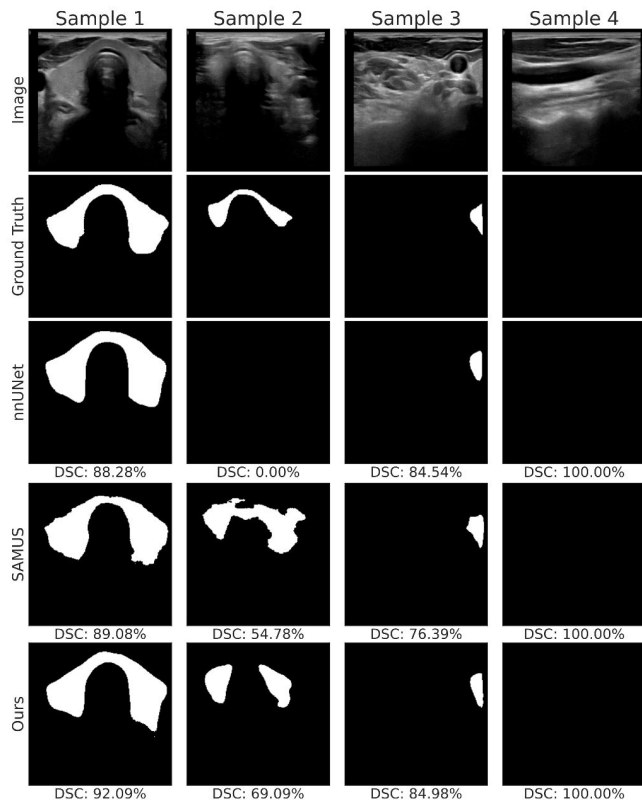


	TN3K [12]		DDTI [30]		Stanford [1]		TG3K [12]		In-house	
	DSC ↑	HD95 ↓	DSC ↑	HD95 ↓	DSC ↑	HD95 ↓	DSC ↑	HD95 ↓	DSC ↑	HD95 ↓
UNet [35]	67.93	41.26	48.43	52.60	89.09	18.96	70.27	65.88	68.50	60.72
nnUNet [16]	<u>85.13</u>	<u>17.24</u>	73.26	40.43	96.92	2.72	<u>77.50</u>	22.52	79.38	<u>18.59</u>
MedNeXt [36]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

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	<u>13.23</u>
SHAN	<u>96.73</u>	<u>73.59</u>	<u>84.61</u>	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

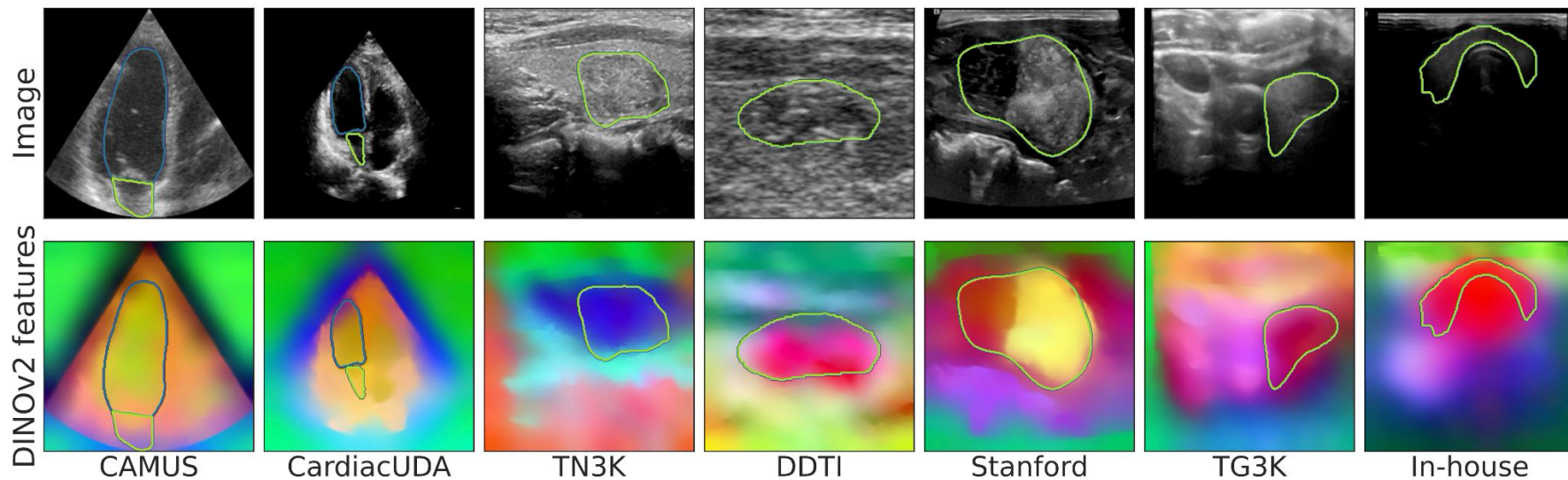


In house dataset:

Model	Precision $\uparrow$	Recall $\uparrow$	Specificity $\uparrow$	F1-Score $\uparrow$
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

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.

Feature visualization



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

Ablation study

H-Dec.	H-Adp.	DINOv2	Interleave	DSC $\uparrow$	HD $\downarrow$
$\times$	$\times$	$\times$	$\times$	78.23	30.98
$\checkmark$	$\times$	$\times$	$\times$	90.85	9.42
$\checkmark$	$\checkmark$	$\times$	$\times$	91.89	6.84
$\checkmark$	$\checkmark$	$\checkmark$	$\times$	91.90	6.78
$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	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

Conclusion

- We propose a method for adapting vision foundation models to ultrasound segmentation.
- Our method demonstrates strong efficiency when trained with 1% or 10% of data in cardiac ultrasound and achieves state of the art generalization result in thyroid ultrasound.
- Our approach enables real-time segmentation at ~77 FPS with TensorRT.

Overview

Develop *reliable* and *adaptable* models that generalize across domains without requiring extensive annotation.

1) Identify *reliable* supervision signals for unsupervised registration.

I. More expressive likelihood
HeteroUnc, *MICCAI 2024*
AdaCS, *ECCV 2024*

II. More plausible prior
BIM-reg, *STACOM 2022*

2) Build *adaptable* models for versatile medical data.

I. Training-time adaptation:
Adapt vision foundation model for ultrasound.
(AdaSeg, *MICCAI 2025*)

Overview

Develop *reliable* and *adaptable* models that generalize across domains without requiring extensive annotation.

1) Identify *reliable* supervision signals for unsupervised registration.

I. More expressive likelihood
HeteroUnc, *MICCAI 2024*
AdaCS, *ECCV 2024*

II. More plausible prior
BIM-reg, *STACOM 2022*

2) Build *adaptable* models for versatile medical data.

I. Training-time adaptation:
Adapt vision foundation model for ultrasound.
(AdaSeg, *MICCAI 2025*)

II. Test-time energy adaptation for segmentation.

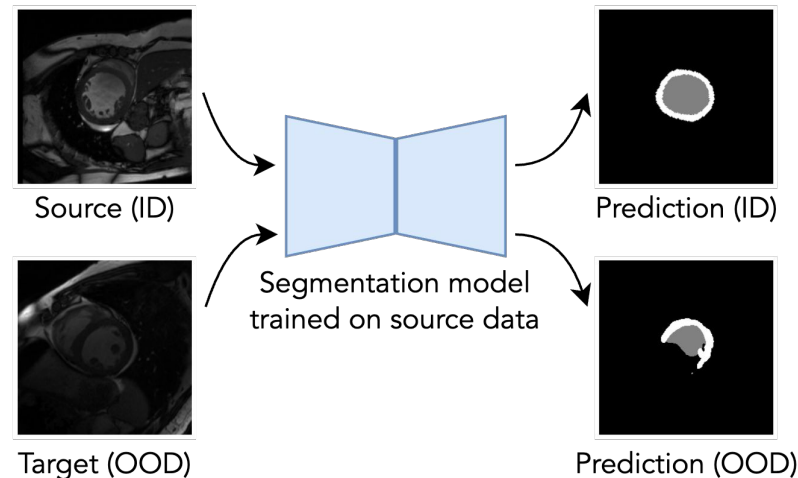
Progressive Test Time Energy Adaptation for Medical Image Segmentation

Xiaoran Zhang¹, Byung-Woo Hong², Hyungseob Park¹, Daniel H. Pak¹, Anne-Marie Rickmann¹,
Lawrence H. Staib¹, James S. Duncan¹, Alex Wong¹

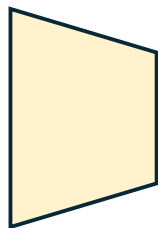
1. Yale University 2. Chung-Ang University

Motivation

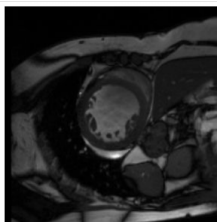
- Covariate shifts caused by nuisances such as heteroscedastic noise and inconsistent imaging protocols limit the fidelity of medical image segmentation models.
- Without assuming access to a pre-collected target dataset, which is often impractical, test-time adaptation (TTA) offers a practical solution to calibrate models on-the-fly during inference.



Test-time adaptation



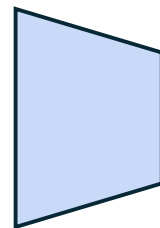
Source pre-trained
model



Target only



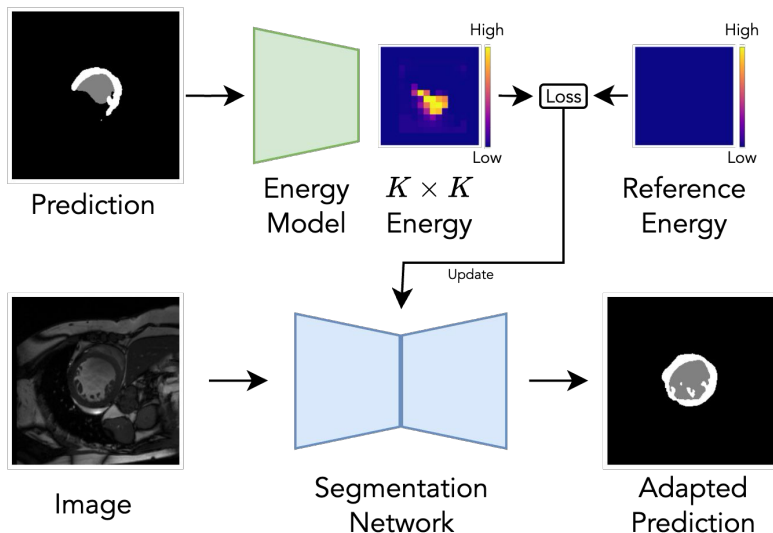
A single
iteration pass



Adapted
model

Test-Time Adaptation is the task of adapting a pretrained source model to the target domain in an unsupervised manner, i.e., without accessing the source data.

Region-based shape energy model



Why energy? Quantifies distribution misalignment at test-time.

- Energy-based models naturally capture distribution changes by reflecting sample likelihood, make them suited for TTA.

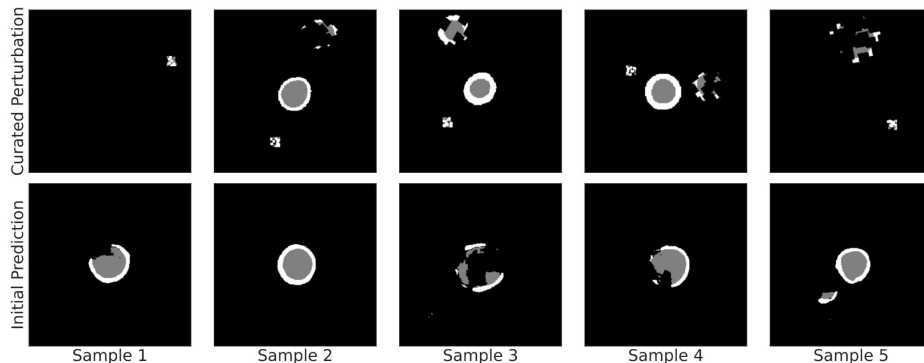
$$p_{\phi}(x) = \frac{\exp(-E_{\phi}(x))}{Z(\phi)}$$

- We propose a shape energy model trained on source data, which assigns an energy score at the region level:
 - low energy -> ID (accurate) shapes
 - high energy -> OOD (erroneous) predictions

Curate negative examples for energy model training

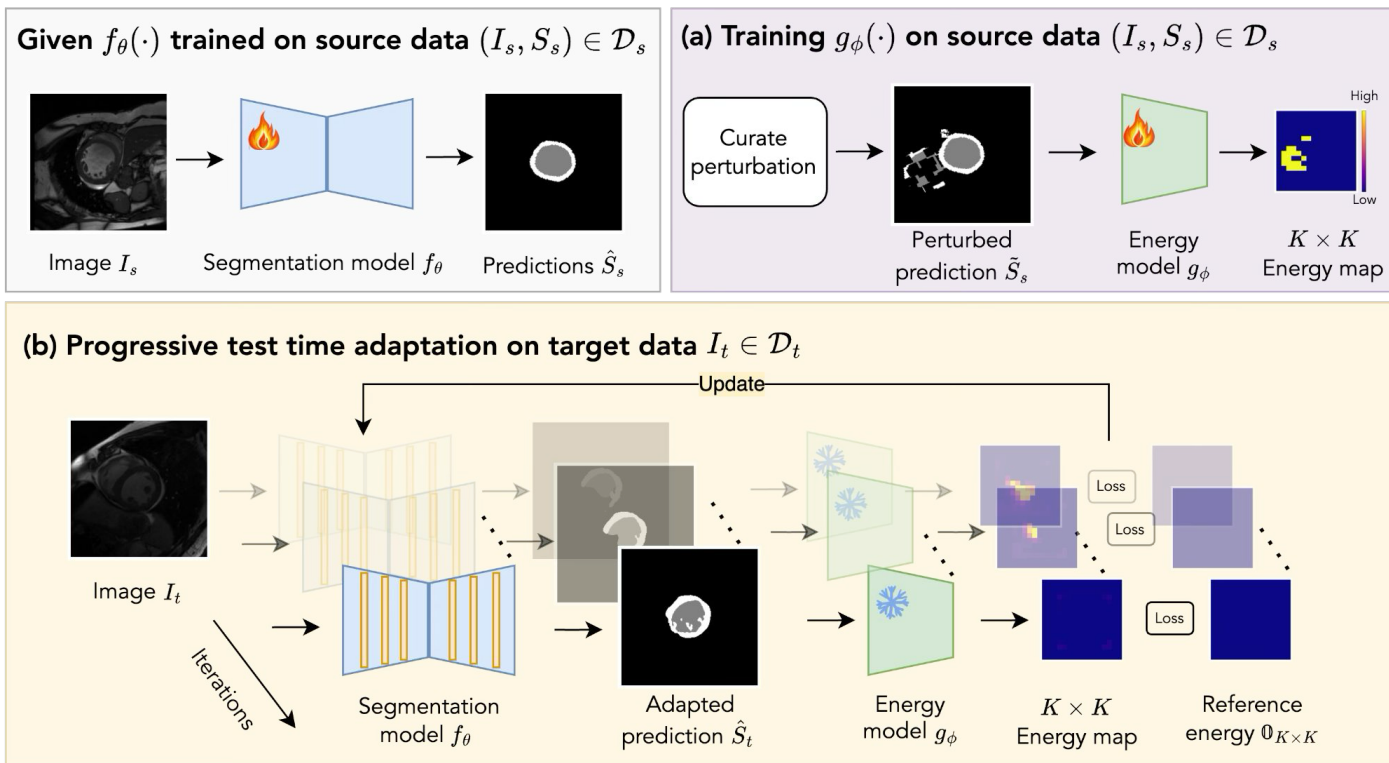
Impossible to have real negative examples? Curate them instead.

- Our formulation assumes two collections of examples, one following the desired distribution (of shapes) and the other out-of-distribution (OOD).
- In addition, the input distribution to the energy model is constrained by the predictions afforded by the segmentation model.



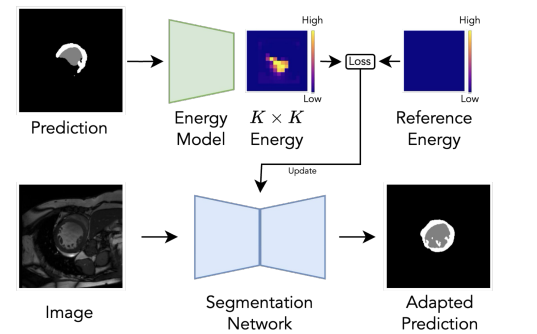
We propose to explore data space by probing the segmentation model with inputs optimized to simulate OOD examples.

Method

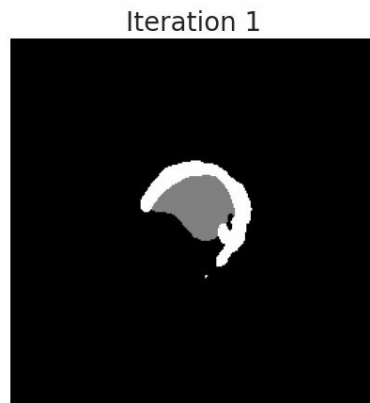
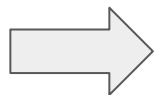


Results

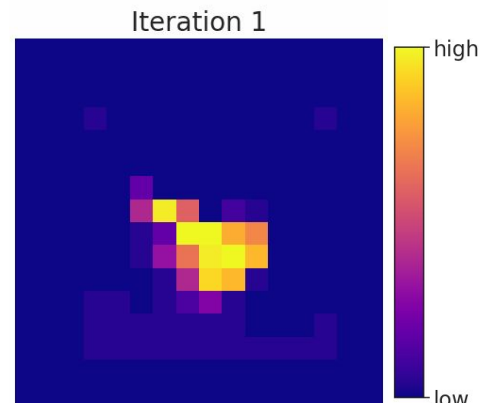
Adapt progressively



Initial prediction



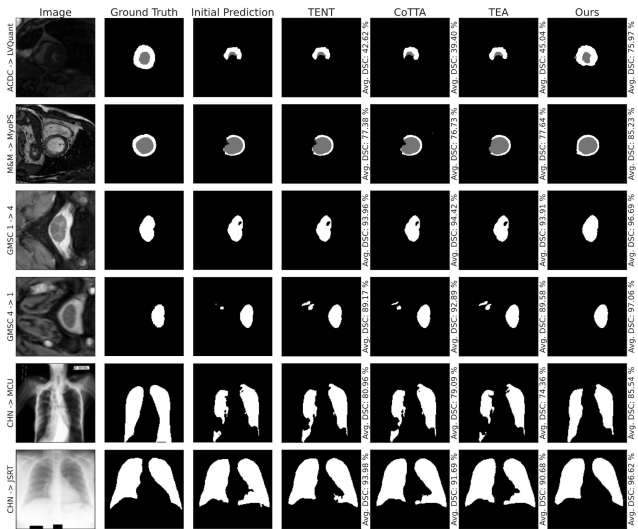
Adapted prediction



Estimated energy

Results

		ACDC [2] $\mapsto$ LVQuant [61]				ACDC [2] $\mapsto$ MyoPS [28]				ACDC [2] $\mapsto$ M&M [3]			
		LV		Myo		LV		Myo		LV		Myo	
		DSC $\uparrow$	ASD $\downarrow$	DSC $\uparrow$	ASD $\downarrow$	DSC $\uparrow$	ASD $\downarrow$	DSC $\uparrow$	ASD $\downarrow$	DSC $\uparrow$	ASD $\downarrow$	DSC $\uparrow$	ASD $\downarrow$
UNet	Pretrained [43]	58.98	24.40	42.52	19.37	85.69	2.99	72.91	2.26	47.69	24.11	41.19	<u>15.89</u>
	TENT [55]	65.78	<u>15.37</u>	51.57	12.78	85.63	<u>2.94</u>	73.49	3.24	57.01	<u>21.15</u>	<u>48.26</u>	19.99
	CoTTA [56]	64.58	17.69	50.52	13.80	85.64	2.96	73.47	3.24	<u>52.98</u>	27.55	46.72	24.65
	TEA [63]	67.96	16.42	<u>54.10</u>	11.17	85.88	3.21	73.98	<u>2.86</u>	52.83	38.43	48.06	29.32
	Ours	76.93	8.77	59.43	<u>11.68</u>	86.06	2.93	78.89	1.91	61.84	19.28	53.13	15.88
MedNeXt	Pretrained [44]	57.55	8.67	42.26	4.80	84.39	3.39	75.77	2.07	78.43	5.48	61.06	2.95
	TENT [55]	75.10	6.10	54.91	3.97	<u>84.48</u>	<u>3.35</u>	75.92	2.04	83.18	<u>4.53</u>	67.56	<u>2.70</u>
	CoTTA [56]	74.57	6.32	54.85	3.93	84.46	3.36	<u>75.95</u>	<u>2.03</u>	82.90	4.83	<u>67.93</u>	2.89
	TEA [63]	<u>75.85</u>	<u>5.96</u>	<u>55.32</u>	<u>3.88</u>	84.12	3.44	75.25	2.07	<u>83.53</u>	4.64	67.84	2.77
	Ours	76.22	5.29	57.29	3.70	84.78	3.28	76.44	1.98	83.82	4.11	68.40	2.49
SwinUNETR	Pretrained [16]	68.44	<u>5.92</u>	47.64	4.20	84.84	3.26	76.35	1.99	81.92	<u>3.52</u>	61.83	<u>3.03</u>
	TENT [55]	74.06	6.64	54.15	4.18	85.06	3.20	77.38	1.98	83.27	4.02	67.26	3.59
	CoTTA [56]	73.41	6.38	54.19	4.19	85.18	<u>3.19</u>	77.72	<u>1.91</u>	83.43	3.87	67.61	3.47
	TEA [63]	<u>74.32</u>	5.99	54.73	<u>4.11</u>	85.04	<u>3.19</u>	<u>77.79</u>	<u>1.91</u>	83.93	3.90	68.60	3.58
	Ours	76.05	5.79	<u>54.22</u>	3.98	85.22	3.17	77.87	1.90	<u>83.80</u>	3.13	<u>68.15</u>	2.66



Our proposed approach can be plugged-and-played into three existing architectures and we consistently outperform baselines in eight datasets.

Conclusion

- We propose a energy-based formulation for TTA in medical image segmentation, leveraging adversarial perturbations to explore the energy landscape and enforce shape priors for adapting to covariate shifts.
- Our shape energy model exhibits strong out-of-distribution detection capabilities at test time, accurately identifying erroneous segmentation regions over 92% accuracy.
- We consistently outperforms baselines across diverse medical datasets, achieving state-of-the-art results.

Overview

Develop *reliable* and *adaptable* models that generalize across domains without requiring extensive annotation.

1) Identify *reliable* supervision signals for unsupervised registration.

I. More expressive likelihood
HeteroUnc, *MICCAI 2024*
AdaCS, *ECCV 2024*

II. More plausible prior
BIM-reg, *STACOM 2022*

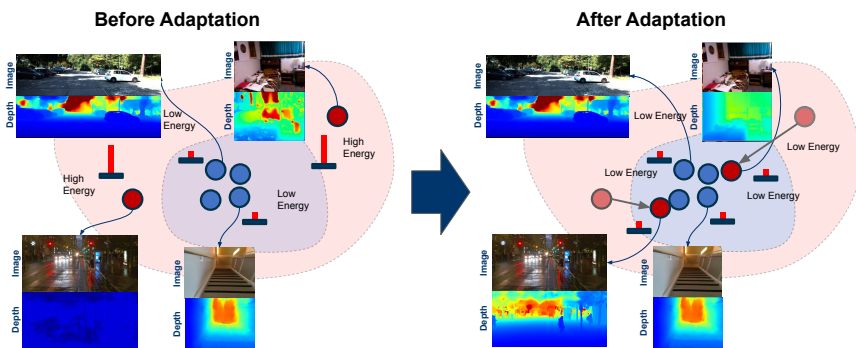
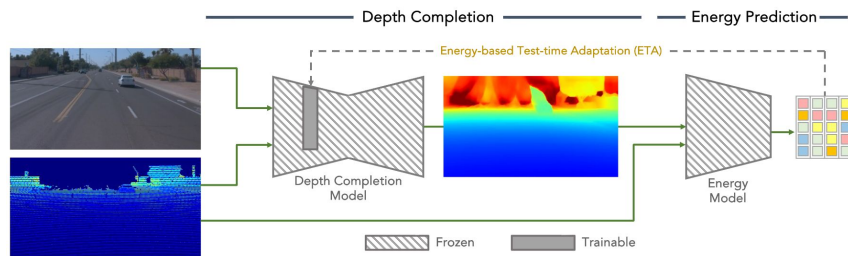
2) Build *adaptable* models for versatile medical data.

I. Training-time adaptation:
Adapt vision foundation model for ultrasound.
(AdaSeg, *MICCAI 2025*)

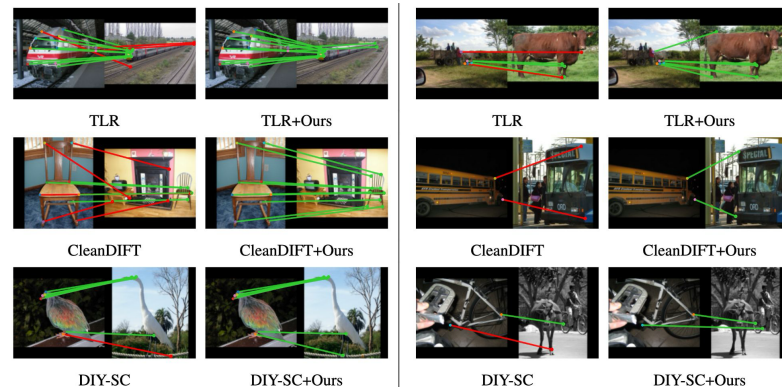
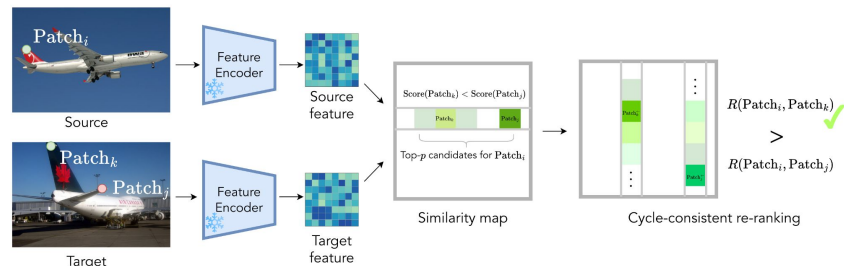
II. Testing-time adaptation:
Test-time energy adaptation for segmentation.
(PTTEA, *ICCV 2025*)

Beyond medical imaging

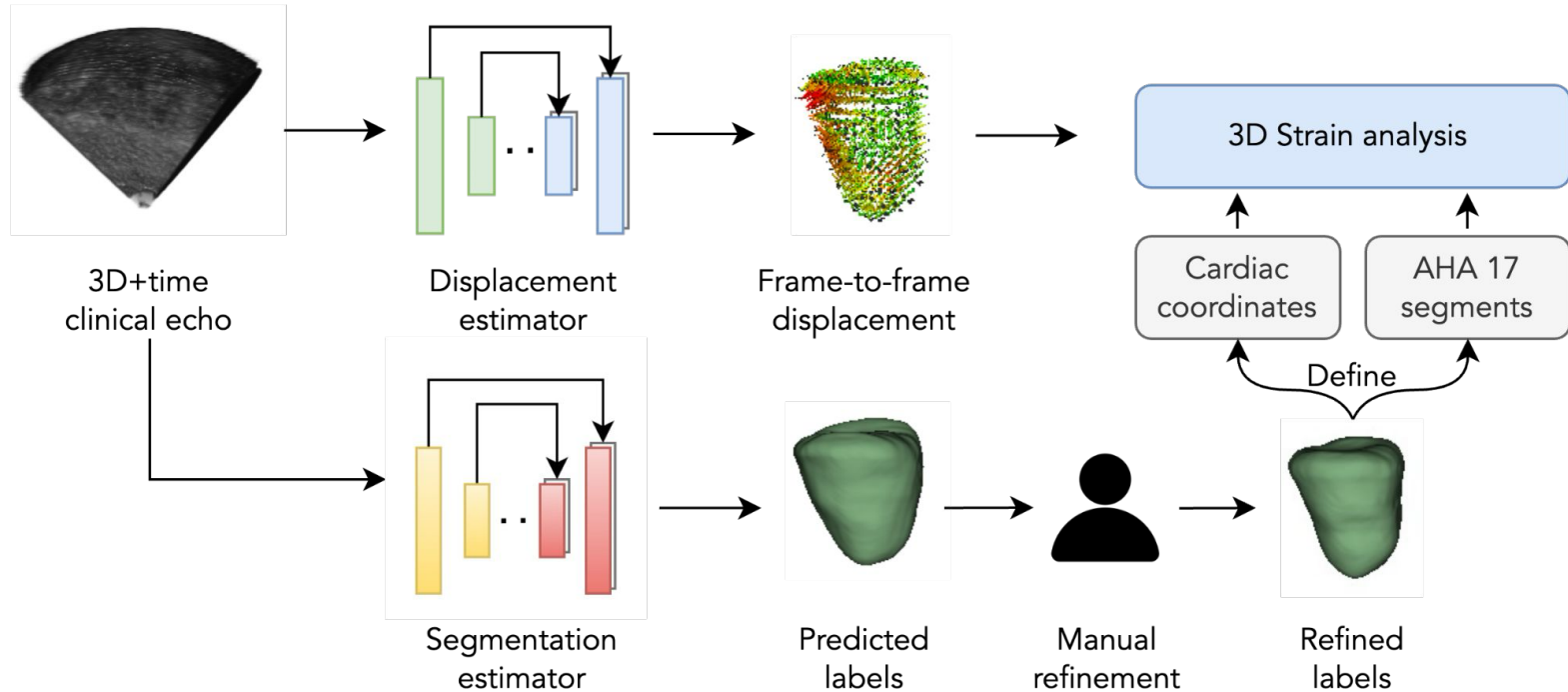
Test-time Energy Adaptation for Depth Completion (ICCV 2025)



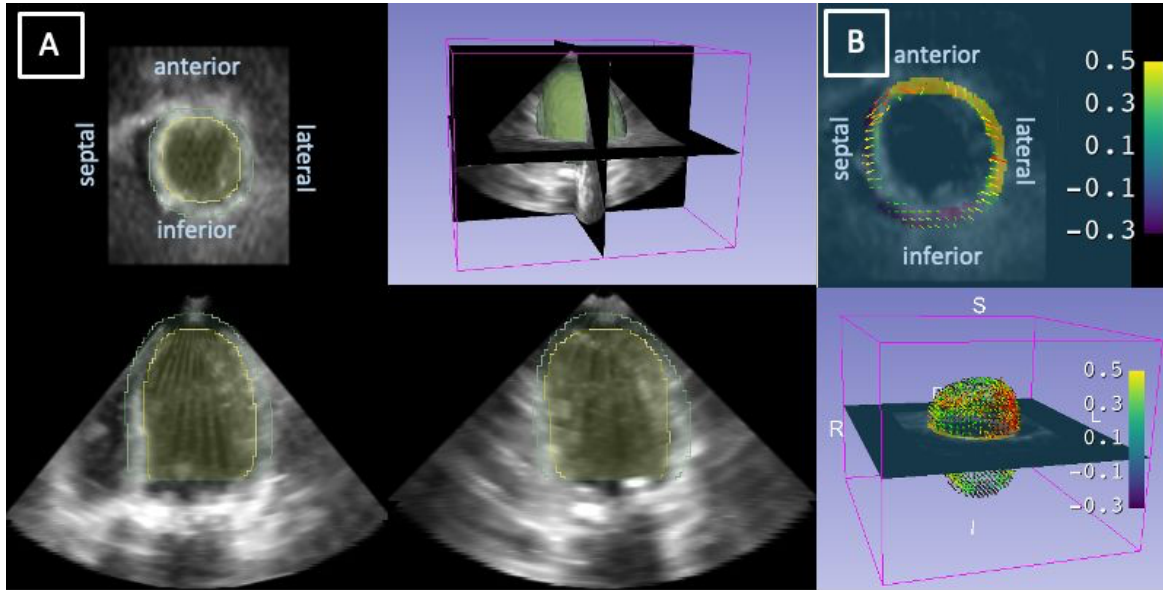
Deja Vu: Cycle-consistent Re-ranking for Semantic Correspondence (Under review)



Application - cardiac strain analysis



Application - Cardiac Strain Analysis



(A) Segmented clinical echo (rest)

(B) Rest radial strain overlaid with estimated displacement revealing akinetic septal and inferior walls

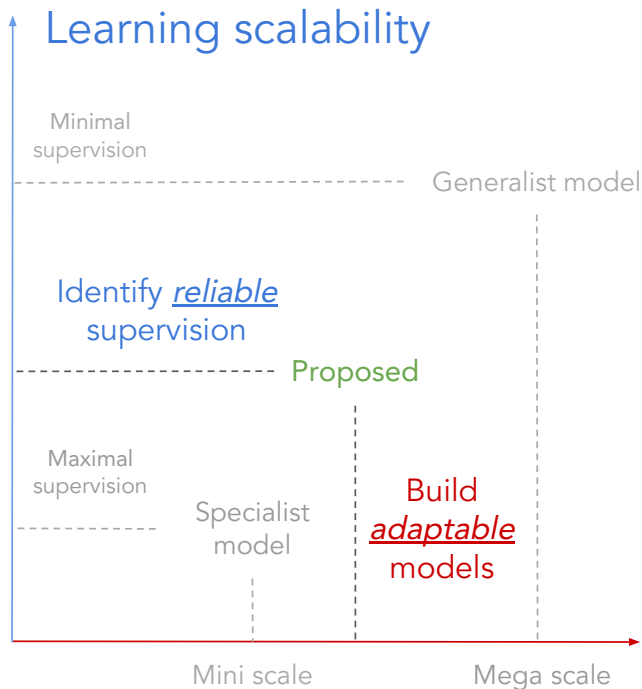
Future works

- Extend and evaluate the framework on 3D data and broader modalities.
- Expand clinical validation and calibrate regional strain values and improve robustness.
- Incorporate physics-based constraints and material properties to improve physiological plausibility.

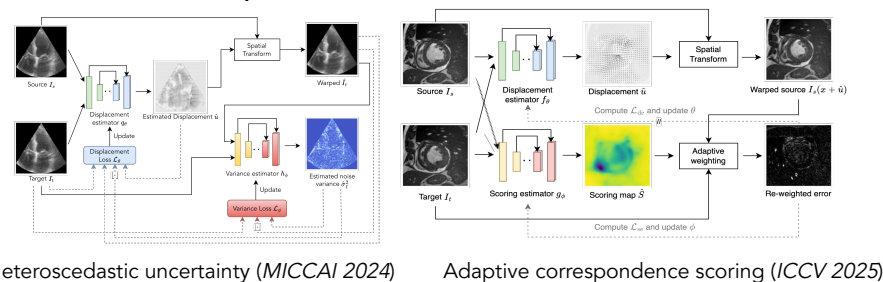
When you only need an apple, grow an apple tree not the whole orchard.



Thank you! Q & A



- 1) Identify reliable supervision signals for unsupervised learning.



- 2) Build adaptable models for versatile medical data.

