Paper

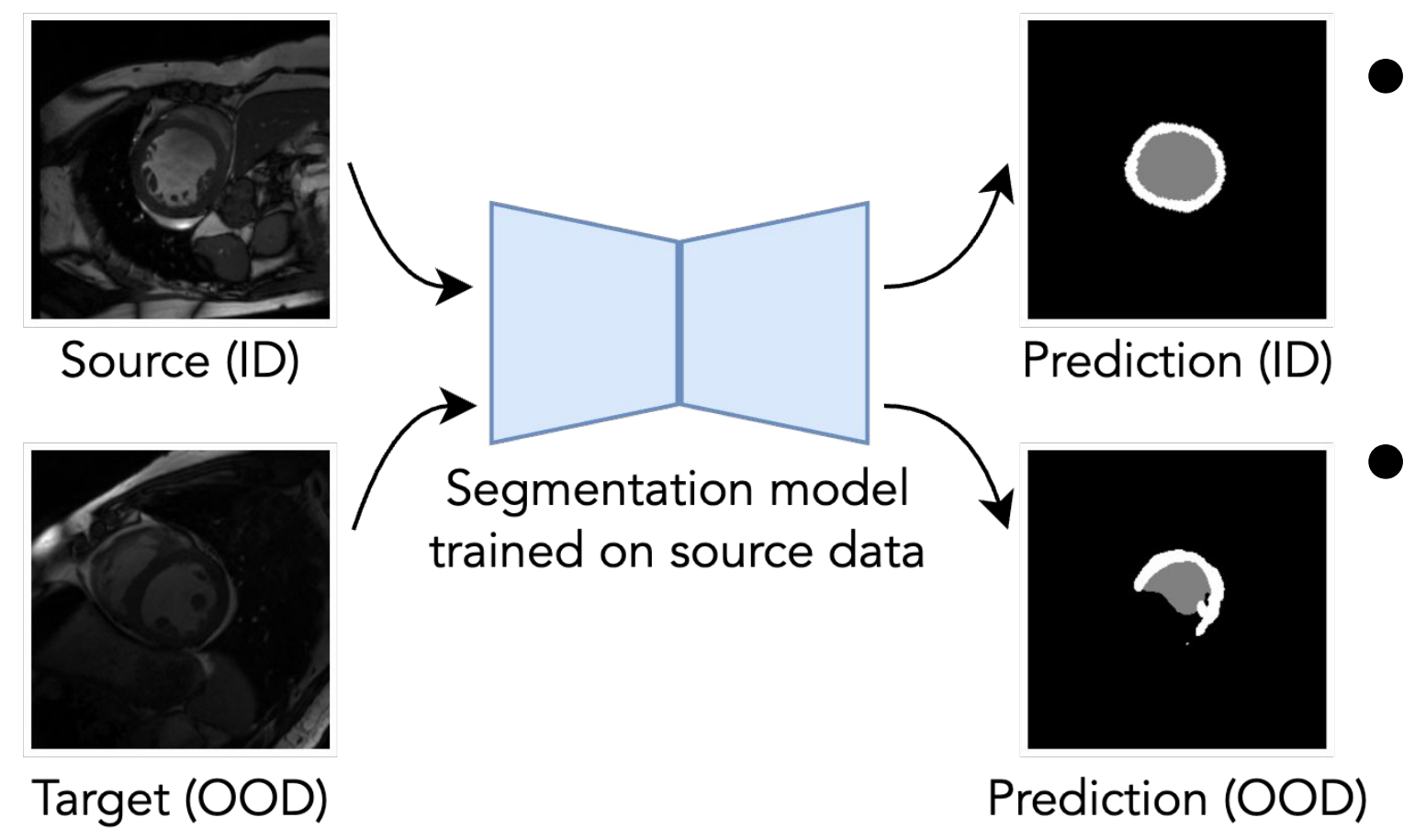
Project page

Progressive Test Time Energy Adaptation for Medical Image Segmentation

Xiaoran Zhang[1], Byung-Woo Hong[2], Hyoungseob Park[1], Daniel H. Pak[1], Anne-Marie Rickmann[1], Lawrence H. Staib[1], James S. Duncan[1*], Alex Wong[1*]
1. Yale University 2. Chung-Ang University *Joint supervision

ICCV OCT 19-23, 2025 HONOLULU HAWAII

Background



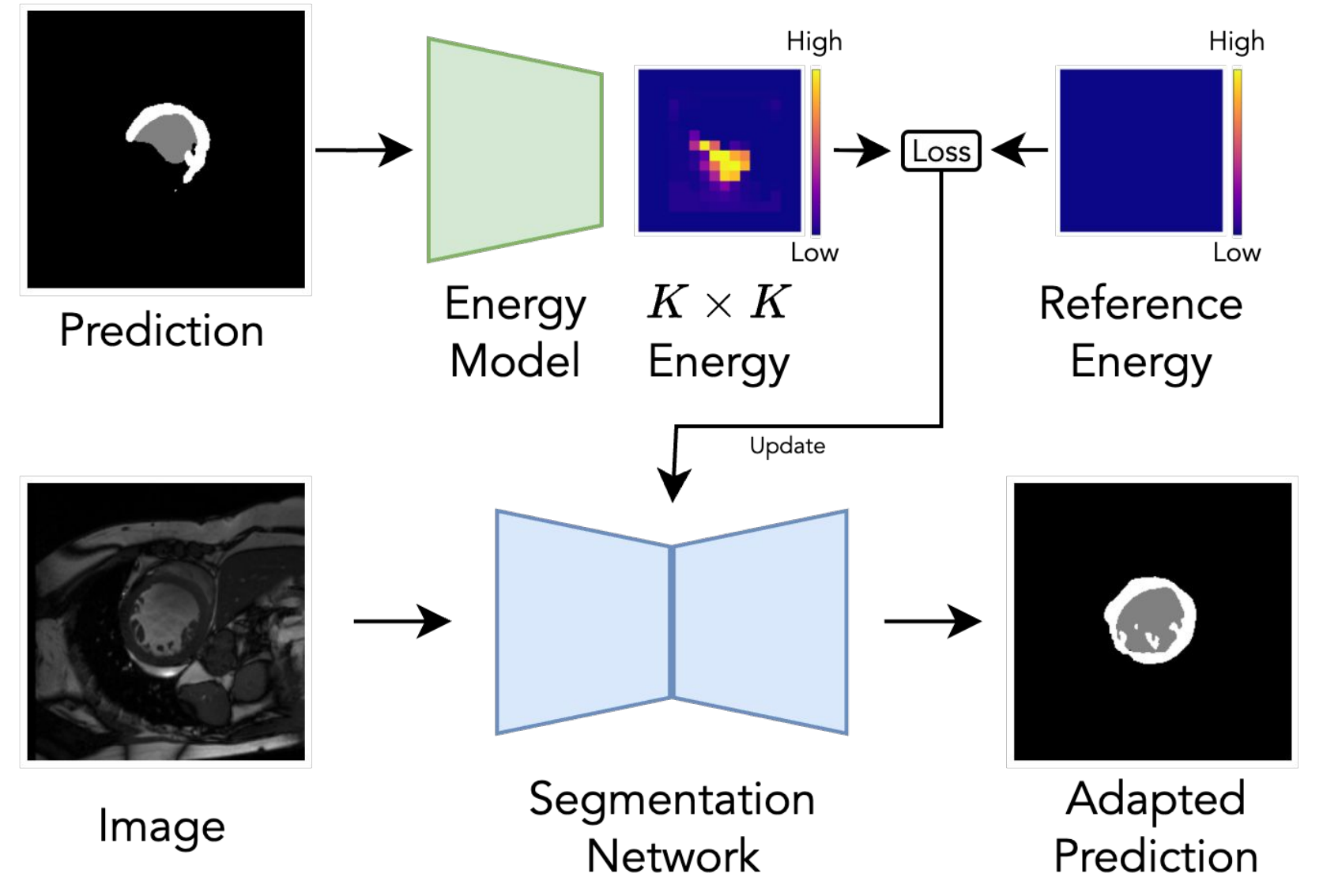
- Covariate shifts caused by nuisances such as heteroscedastic noise and inconsistent imaging protocols limit the fidelity of medical image segmentation models.
- Without assuming access to a pre-collected target dataset, which is often impractical, test-time adaptation (TTA) offers a practical solution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

Problem formulation

Assuming a segmentation model is solely trained on source dataset, our goal is to adapt the model to target data without access to the entire target dataset.

Key components

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.

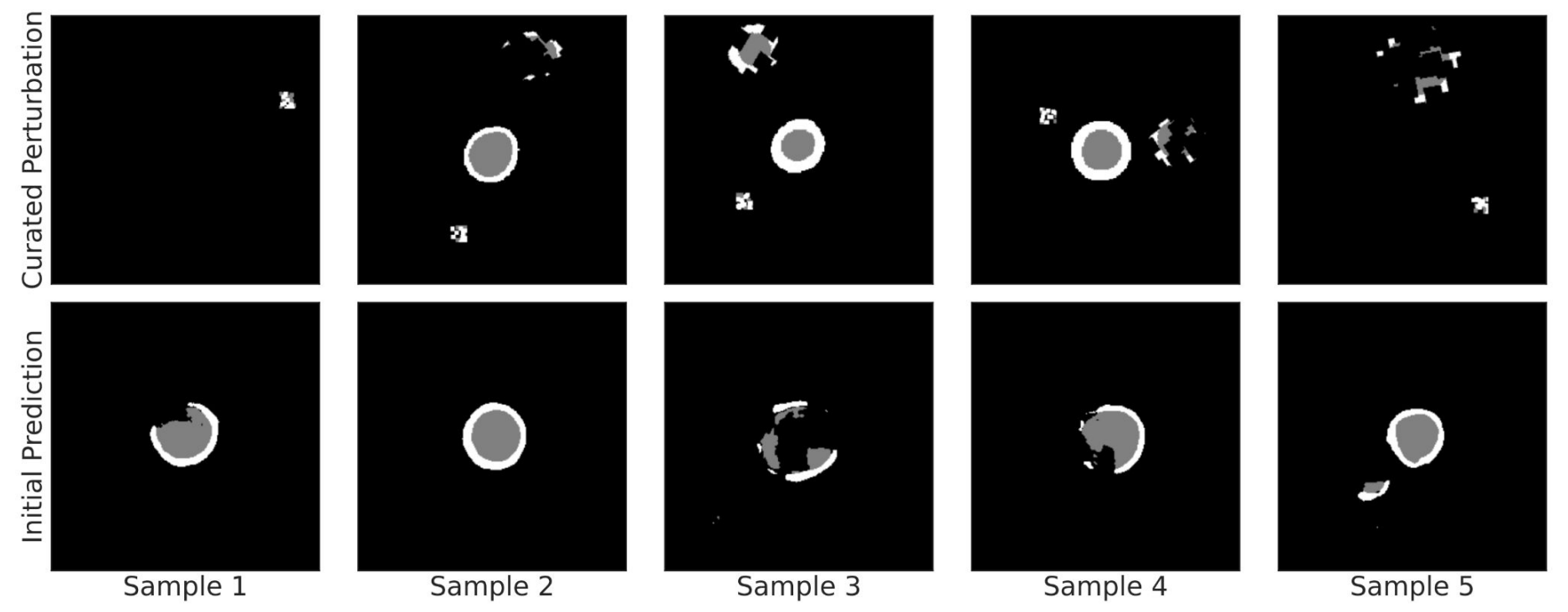
$$p_\phi(x) = \frac{\exp(-E_\phi(x))}{Z(\phi)}$$

- 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.

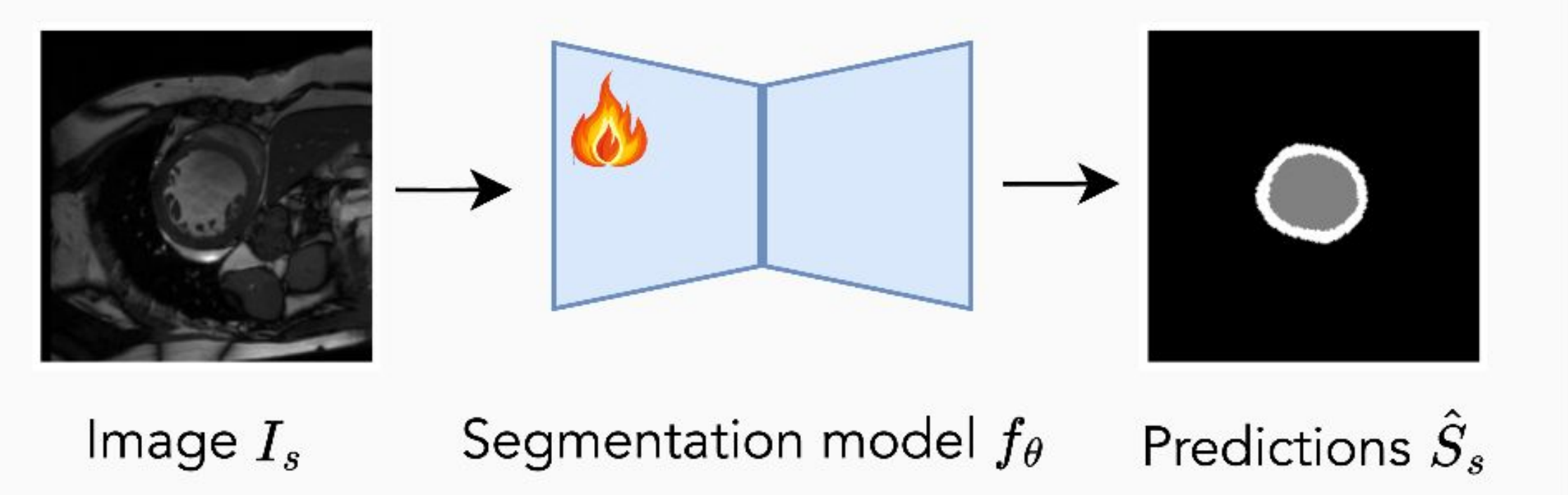
- Our formulation assumes two collections of examples, one following the desired distribution (of shapes) and the other out-of-distribution (OOD).
- In addition, the input distribution to the energy model is constrained by the predictions afforded by the segmentation model.



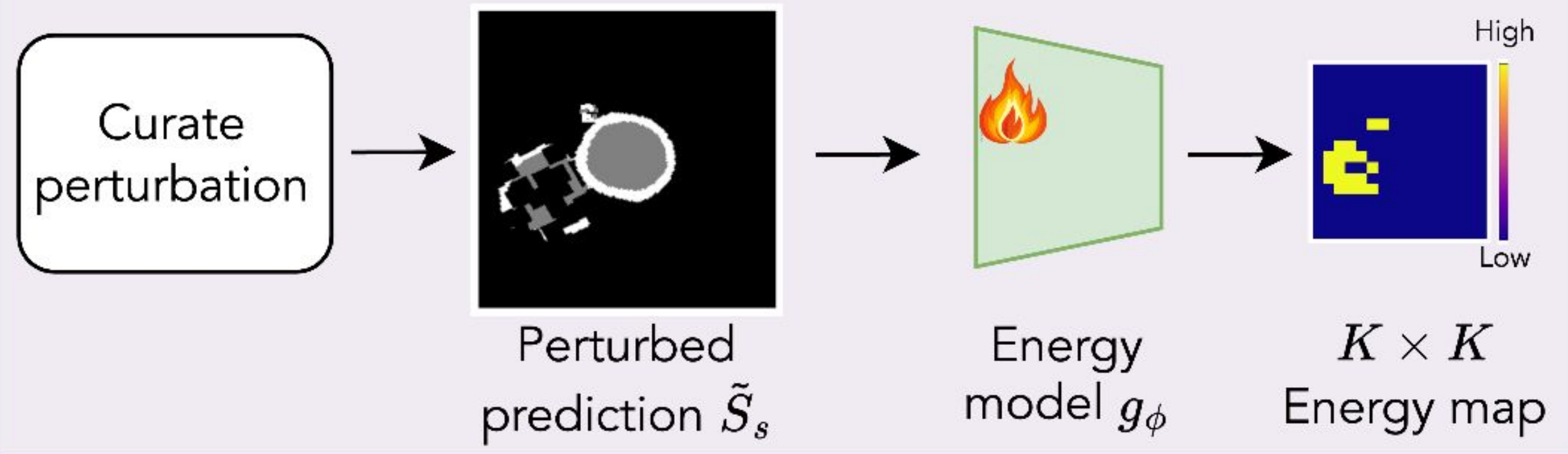
We propose to explore data space by probing the segmentation model with inputs optimized to simulate OOD examples.

Methods

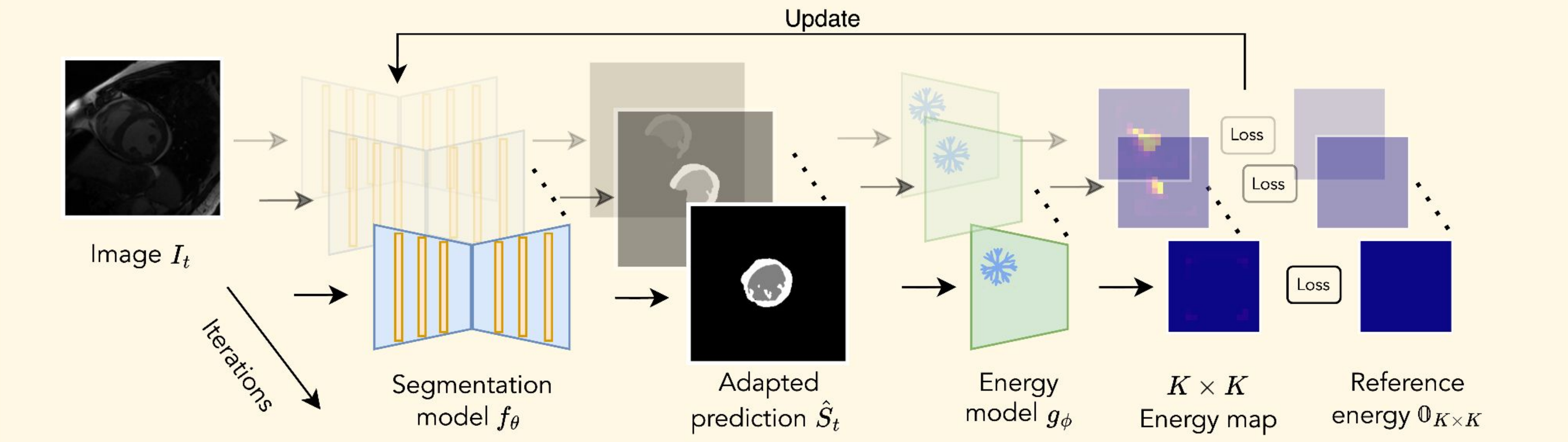
Given $f_\theta(\cdot)$ trained on source data $(I_s, S_s) \in \mathcal{D}_s$



(a) Training $g_\phi(\cdot)$ on source data $(I_s, S_s) \in \mathcal{D}_s$



(b) Progressive test time adaptation on target data $I_t \in \mathcal{D}_t$


Update
Image I_t
Iterations
Segmentation model f_θ
Adapted prediction $\hat{S}_t$
Energy model g_ϕ
Loss
$K \times K$ Energy map
Reference energy $0_{K\times K}$

Perturbation curation

We generate negative (implausible) examples by applying FGSM adversarial noise and spatial affine transformations to the input images.

- Apply FGSM adversarial noise: $\epsilon = \delta\,\mathrm{sign}\left(\nabla_{I_s}\mathcal{L}(f_\theta(I_s), S_s)\right)$
- Apply random affine transforms
- Generate perturbed segmentation: $\tilde{S}_s = f_\theta(I_s + \epsilon)$

Shape energy model training

A region-based model learns patchwise energy values, assigning high energy to implausible regions and low energy to anatomically valid ones.

Loss: $\mathcal{L}_\phi = \frac{1}{N_p}\sum_{i=1}^{N_p}\Big(-y_s^i \log\sigma(-g_\phi(s_s^i)) - (1-y_s^i)\log(1-\sigma(-g_\phi(s_s^i)))\Big)$

Label curation

For each perturbed segmentation, we compare it with ground truth and assign categorical energy labels to each region, where regions dissimilar to the ground truth are labeled as high-energy. $y_s = 1 - \mathbf{1}(d(\tilde{s}_s, s_s) < \tau)$

Progressive test-time adaptation

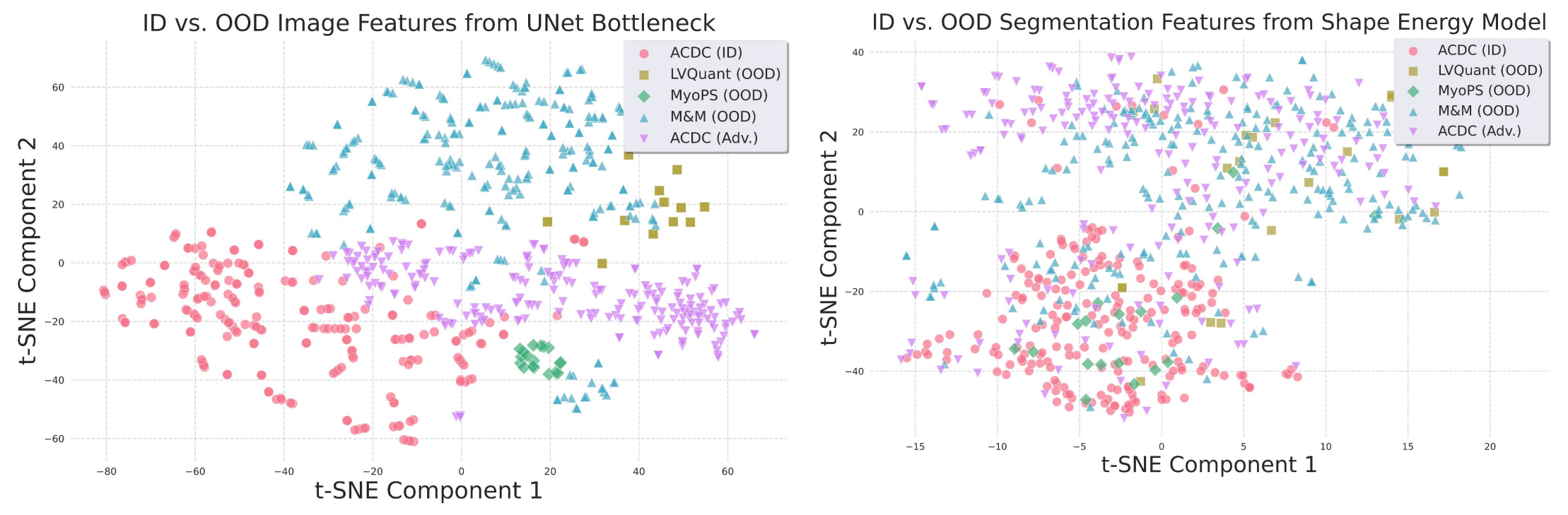
At inference, the segmentation model is iteratively updated to minimize the predicted energy, aligning outputs with plausible anatomical shapes.

Update rule: $\theta^* = \arg\min_\theta - \sum_{i=1}^{B_t}\log(1-\sigma(-g_\phi(\hat{s}_t^i)))$

Datasets: (1) Cardiac (2D MRI)): ACDC, LVQuant, MyoPS, M&M (2) Spinal cord (2D MRI): GMSC (sites 1-4) (3) Lung (2D X-ray): CHN, MCU, JSRT. Metrics: (1) Dice coefficient score (DSC, %) (2) average surface distance (ASD, %)

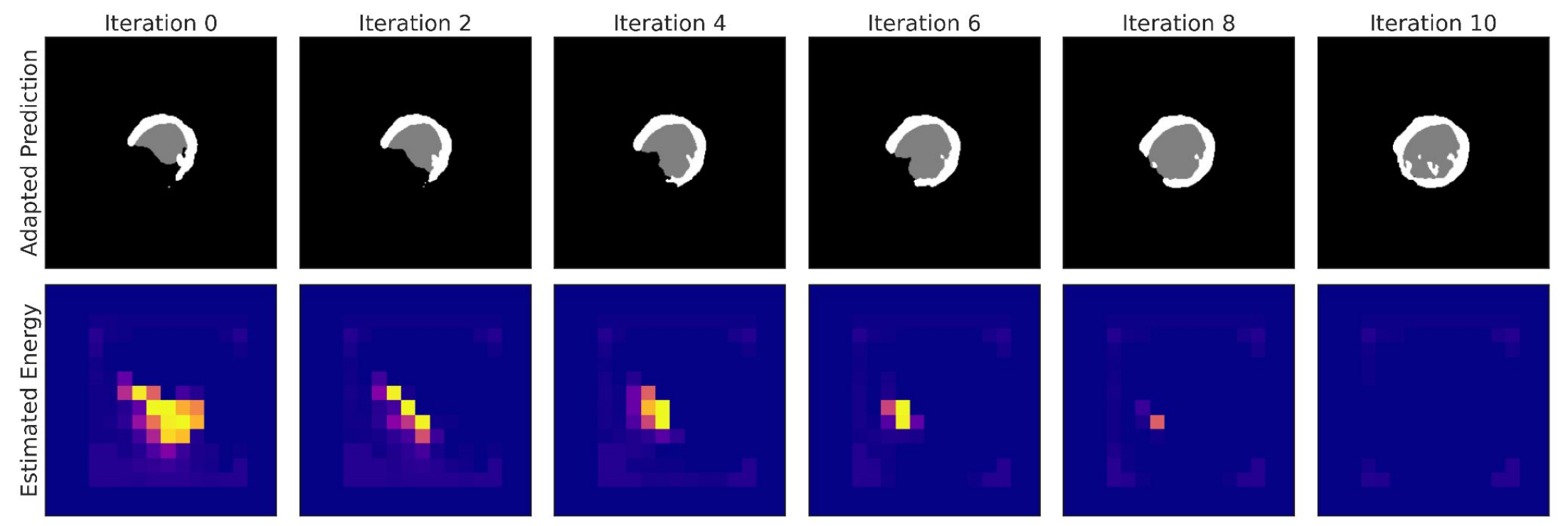
Results

T-SNE analysis of curated perturbation

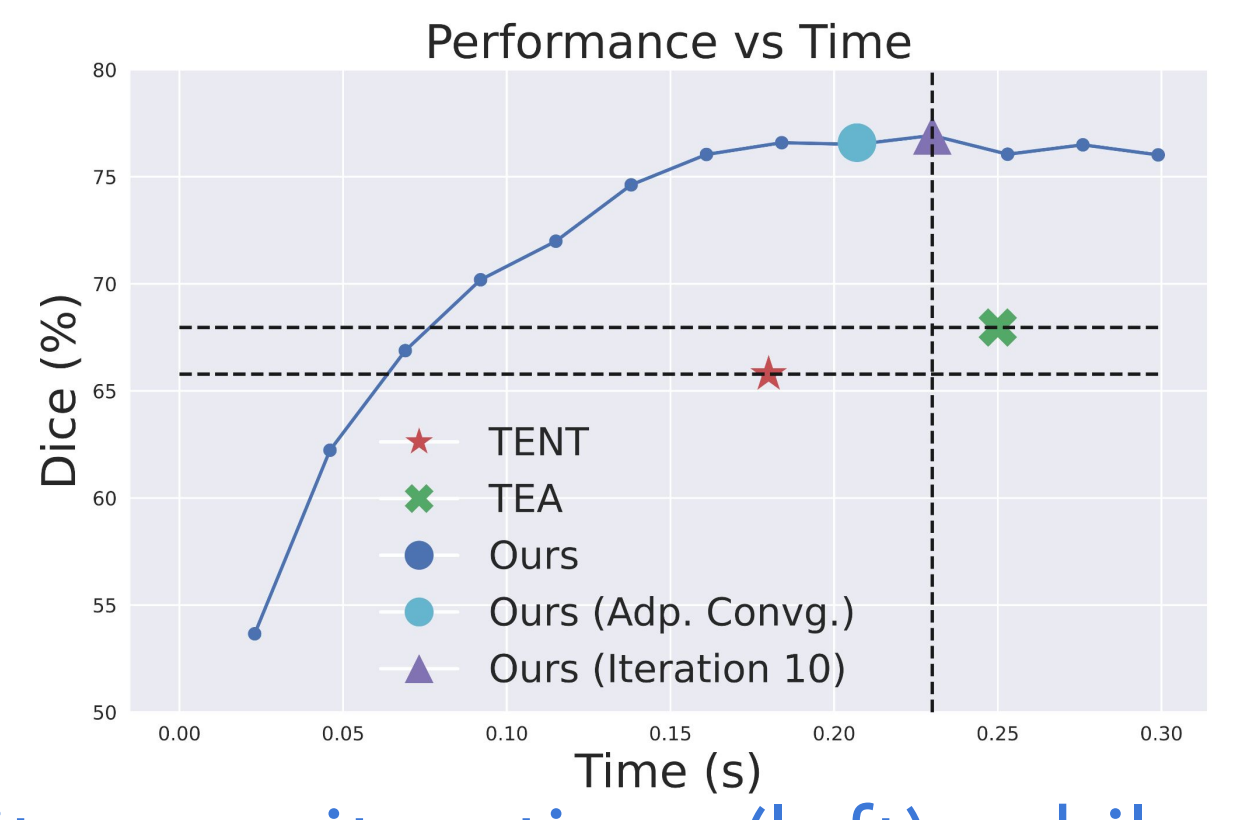


Our adversarial perturbation strategy produces images and segmentations that align with OOD cases, validating its effectiveness in modeling real covariate shifts.

Progressive update visualization



Convergence analysis

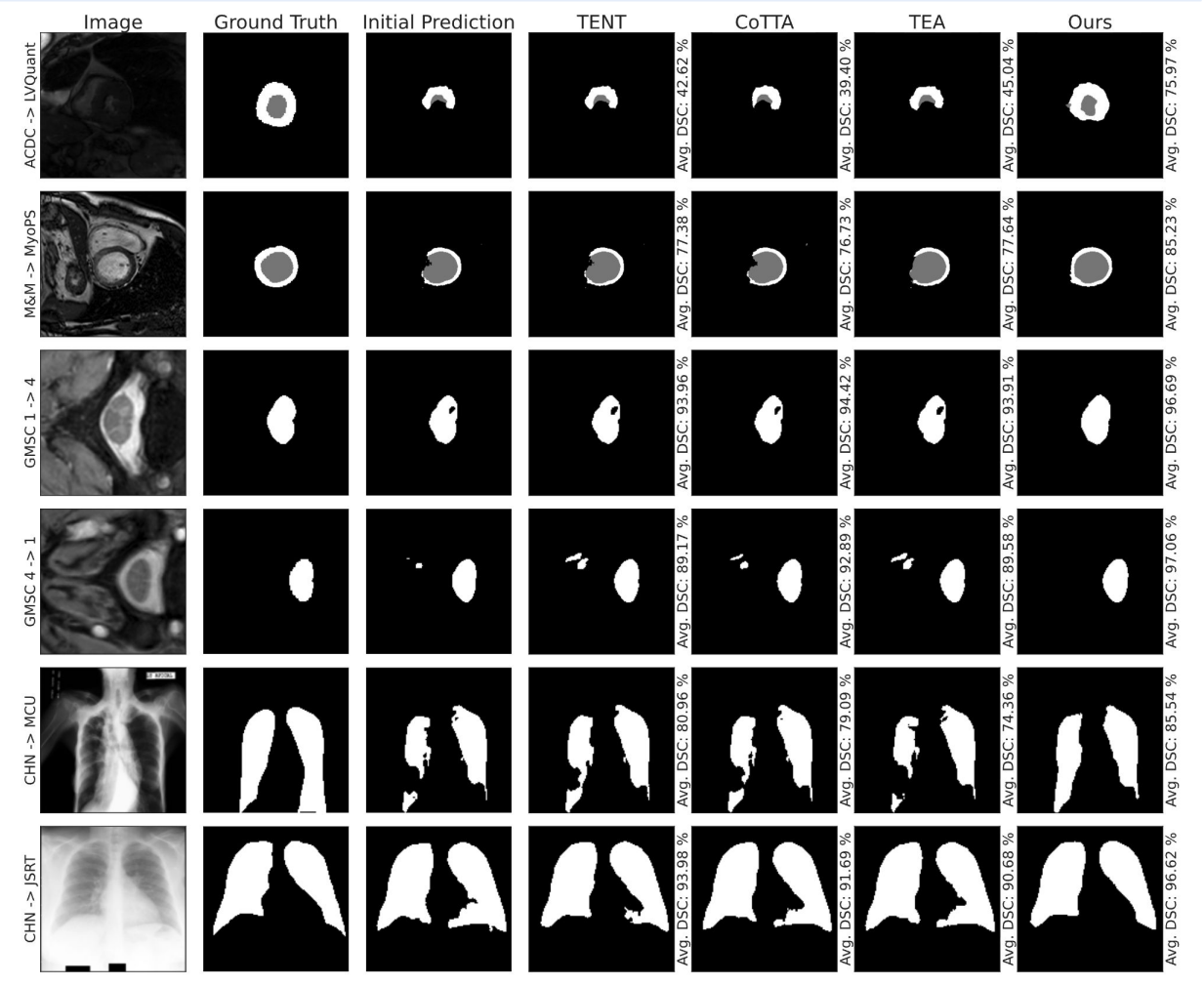


Our method progressively refines segmentation quality over iterations (left), while achieving better convergence under the same time budget (right).

Quantitative evaluation

		ACDC [2] ↦ LVQuant [61]				ACDC [2] ↦ MyoPS [28]				ACDC [2] ↦ M&M [3]			
		LV		Myo		LV		Myo		LV		Myo	
		DSC ↑	ASD ↓	DSC ↑	ASD ↓	DSC ↑	ASD ↓	DSC ↑	ASD ↓	DSC ↑	ASD ↓	DSC ↑	ASD ↓
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	15.89
	TENT [55]	65.78	15.37	51.57	12.78	85.63	2.94	73.49	3.24	57.01	21.15	48.26	19.99
	CoTTA [56]	64.58	17.69	50.52	13.80	85.64	2.96	73.47	3.24	52.98	27.55	46.72	24.65
	TEA [63]	67.96	16.42	54.10	**11.17**	85.88	3.21	73.98	2.86	52.83	38.43	48.06	29.32
	Ours	**76.93**	**8.77**	**59.43**	11.68	**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	84.48	3.35	75.92	2.04	83.18	4.53	67.56	2.70
	CoTTA [56]	74.57	6.32	54.85	3.93	84.46	3.36	75.95	2.03	82.90	4.83	67.93	2.89
	TEA [63]	75.85	5.96	55.32	3.88	84.12	3.44	75.25	2.07	83.53	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	5.92	47.64	4.20	84.84	3.26	76.35	1.99	81.92	3.52	61.83	3.03
	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	3.19	77.72	1.91	83.43	3.87	67.61	3.47
	TEA [63]	74.32	5.99	**54.73**	4.11	85.04	3.19	77.79	1.91	**83.93**	3.90	**68.60**	3.58
	Ours	**76.05**	**5.79**	54.22	**3.98**	**85.22**	**3.17**	**77.87**	**1.90**	83.80	**3.13**	68.15	**2.66**

Adaptation visualization



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

High-energy corresponds to test-time segmentation errors

Method	ACDC ↦ LVQuant	ACDC ↦ MyoPS	ACDC ↦ M&M
UNet	93.64	96.55	94.53
MedNeXt	92.17	95.83	93.66
SwinUNETR	92.01	96.42	93.97

Our shape energy model achieves over 92% accuracy across different segmentation models, confirming its effectiveness in identifying errors at test-time.

Hyperparameter sensitivity

	† is proposed.	1 ↦ 2	1 ↦ 3	1 ↦ 4	4 ↦ 1	4 ↦ 2	4 ↦ 3	Avg.
Patch Size	4 × 4	69.1	73.2	93.4	89.4	42.5	85.3	75.5
	9 × 9†	**73.6**	77.7	**95.3**	**95.1**	56.2	87.2	**80.9**
	18 × 18	73.0	**77.9**	94.5	94.7	**57.2**	**87.7**	80.8
	36 × 36	69.4	75.5	93.1	88.4	45.9	87.4	76.6
Threshold	$\tau = 25$	70.5	73.1	91.3	94.2	54.1	86.3	78.3
	$\tau = 50^\dagger$	**73.6**	77.7	**95.3**	**95.1**	**56.2**	87.2	**80.9**
	$\tau = 75$	73.0	**79.1**	95.1	94.7	52.6	**87.3**	80.3
	$\tau = 100$	71.6	75.7	95.1	94.6	52.5	86.7	79.4
Pert. Mag.	$\delta = 0.1$	70.8	76.2	94.2	95.1	54.5	87.0	79.6
	$\delta = 0.05^\dagger$	**73.6**	**77.7**	**95.3**	**95.1**	**56.2**	**87.2**	**80.9**
	$\delta = 0.01$	73.3	73.1	93.6	94.8	53.5	86.4	79.1

Paper: https://arxiv.org/abs/2312.00837

Code: https://github.com/Voldemort108X/pttea_seg

Funding: NIH/NHLBI R01HL121226, NSF 2112562 Athena, IITP-RS-2021-II211341, NRF-RS-2023-00251366