Progressive Test Time Energy Adaptation for Medical Image Segmentation

Supplementary Material

A. Dataset Details

A.1. ACDC [2]

The ACDC dataset, publicly accessible, comprises 2D cardiac MRI scans from 150 patients across five subgroups: (1) 30 normal patients, (2) 30 with previous myocardial infarction, (3) 30 with dilated cardiomyopathy, (4) 30 with hypertrophic cardiomyopathy, and (5) 30 with abnormal right ventricles. The acquisitions were obtained using two MRI scanners of different magnetic strengths (1.5 T and 3.0 T). Cine images were acquired in breath hold with an SSFP sequence in short-axis orientation. The spatial resolution ranges from 1.37 to 1.68 mm^2. The dataset includes 100 training and 50 testing subjects. Each sequence has end-diastole (ED) and end-systole (ES) frames with left ventricle and myocardium labels. Our training set includes 80 patients, the validation set 20 patients, and the testing set 50 patients. ED and ES image pairs are extracted slice-by-slice from 2D longitudinal stacks, center-cropped to 256 × 256 around the myocardium centroid in ED frames after resampling the pixel spacing to 1mm × 1mm. This process produces 1266 2D training images, 277 validation images, and 852 testing images. To reduce computational cost, approximately one-quarter of the testing set is randomly subsampled to 213 samples.

A.2. LVQuant [61]

The LVQuant dataset is publicly available and includes short-axis MR sequences from 56 subjects. The 2D cine MR images were collected from three hospitals during routine clinical practice without specific selection criteria, encompassing pathologies ranging from moderate to severe cardiac conditions. Each sequence contains 20 frames of mid-ventricle slices spanning a complete cardiac cycle, acquired with ECG-gating and breath-holding. Ground truth segmentations for the endocardium and epicardium are provided. The MR images have pixel spacings ranging from 0.6 mm to 2.08 mm. Following the preprocessing steps of ACDC, we resample the pixel spacings to 1mm × 1mm and apply a center crop to 256 × 256. End-diastole (ED) and end-systole (ES) frames are extracted from each sequence, and myocardium masks are generated by subtracting the endocardium mask from the epicardium. The dataset is randomly split into 60/20/20 for training, validation, and testing, resulting in 66 training images, 22 validation images, and 24 testing images. Due to the limited size of the training set, it is insufficient for developing a robust source segmentation model. Therefore, only the testing set from LVQuant is reserved for adaptation experiments.

A.3. MyoPS [28]

The MyoPS challenge dataset includes 45 paired three-sequence CMR images (bSSFP, LGE, and T2 CMR) acquired from the same patients, with 25 subjects publicly available. Each subject contains 2–6 slices, with an in-plane resolution ranging from 0.73 to 0.76 mm. The dataset provides gold-standard segmentations for the left ventricular blood pool, right ventricular blood pool, left ventricular myocardium, left ventricular myocardial scar, and edema. We randomly split the dataset into training, validation, and testing sets in a 60/20/20 ratio, resulting in 51 training images, 23 validation images, and 18 testing images. However, similar to the LVQuant dataset, the limited number of training examples makes it insufficient for developing a source segmentation model. Consequently, we only retain the testing set for adaptation experiments. For these experiments, we extract the bSSFP sequence as the input image and compose the ground truth myocardium by combining the myocardial scar and edema labels with the normal myocardium.

A.4. M&M [3]

The M&M dataset is publicly available, comprising 360 cases collected from four vendors (Siemens, Philips, GE Healthcare, and Canon), six centers, and three countries (Spain, Canada, and Germany). The dataset builds upon the labeling framework of the ACDC dataset [2], providing ground truth annotations for the left ventricle, myocardium, and right ventricle. The in-plane resolution ranges from 0.98 to 1.32 mm, and the longitudinal stack contains 10-12 slices. Following a preprocessing pipeline similar to ACDC, we extract the end-diastolic (ED) and end-systolic (ES) frames from each sequence slice-by-slice along the longitudinal stack. After resampling pixel spacings to 1mm × 1mm, we center-crop the frames to 256 × 256 around the myocardium mask in the ED frame. The dataset is randomly divided into training, validation, and testing sets in a 60/20/20 ratio, resulting in 2918 training images, 964 validation images, and 987 testing images. To reduce computational costs, the testing set is subsampled to approximately one-quarter of its size, yielding 246 samples.

A.5. GMSC [42]

The GMSC dataset is a multi-center, multi-vendor collection of spinal cord MRI anatomical images, comprising healthy subjects from four sites: (1) Site 1 from University College London, acquired using a 3T Philips Achieva T1-weighted MRI; (2) Site 2 from Polytechnique Montreal, using a 3T Siemens TIM Trio with T1-weighted; (3) Site

3 from the University of Zurich, using a 3T Siemens Skyra T2-weighted MRI; and (4) Site 4 from Vanderbilt University, acquired with a 3T whole-body Philips scanner for T2-weighted. Each site contains 10 subjects, with manual annotations from four raters. Following [7], we preprocess the data by center-cropping each slice to 144×144 after normalizing image intensity to $[0, 1]$ and randomly split it at the subject level into training, validation, and testing sets using a 60/20/20 ratio, resulting in 21/3/6 for Site 1, 76/13/24 for Site 2, 125/18/36 for Site 3, and 95/12/26 for Site 4. This structured distribution ensures balanced evaluation across imaging centers and scanner variations.

A.6. CHN [21]

The Shenzhen (CHN) chest X-ray dataset, created by the Third People's Hospital of Shenzhen City and Guangdong Medical College in collaboration with the Department of Health and Human Services, is publicly available. It consists of 566 chest X-ray images, including both normal and abnormal cases with tuberculosis manifestations, accompanied by radiologist readings. We preprocess the images by resizing them to 128×128 and randomly splitting the dataset into training, validation, and testing sets using a 60/20/20 ratio, resulting in 339/113/114 images, respectively.

A.7. MCU [21]

The Montgomery County dataset (MCU) is publicly available and was created through a collaboration between the National Library of Medicine and the Montgomery County Department of Health and Human Services. It comprises 138 chest X-ray images, including 80 normal cases and 58 with tuberculosis-related abnormalities. Following the same preprocessing as CHN, we resize each image to 128×128 and randomly split the dataset into training, validation, and testing sets using a 60/20/20 ratio, yielding 82/28/28 images, respectively.

A.8. JSRT [46]

The Japanese Society of Radiological Technology (JSRT) dataset is a publicly available collection of posteroanterior chest X-ray (CXR) images, widely used for lung segmentation and nodule detection research. It comprises 199 images for training and 40 for testing. Following standard protocols from other lung segmentation datasets, we randomly split the dataset into training, validation, and testing sets using a 60/20/20 ratio, yielding 159/40/48 images, respectively. Each image is resized to 128×128. This structured split ensures a balanced evaluation across different dataset partitions, facilitating robust model training and validation.

B. Implementation Details

B.1. Evaluation Metrics

Dice Score For a pair of predicted segmentation mask $\hat{S} \in \{0,1\}^{H \times W}$ and ground truth mask $S \in \{0,1\}^{H \times W}$, the Dice score is defined to measure the ratio of overlap:

$$\mathrm{Dice}(\hat{S}, S) = \frac{2|\hat{S} \cap S|}{|\hat{S}| + |S|}. \tag{9}$$

Here, $|\hat{S} \cap S|$ denotes the number of overlapping elements between the predicted mask $\hat{S}$ and the ground truth mask S, while $|\hat{S}|$ and $|S|$ represent the total number of elements in the predicted mask and ground truth mask, respectively. The Dice score quantifies the similarity between the predicted and ground truth masks, ranging from 0 to 1. A Dice score of 1 indicates perfect overlap, while a score of 0 indicates no overlap.

Average Surface Distance We compute the Average Surface Distance (ASD) to measure the mean boundary deviation between the predicted and ground truth segmentation masks. Formally, ASD is defined as:

$$d_{\mathrm{ASD}}(\hat{S}, S) = \frac{1}{|\partial\hat{S}| + |\partial S|} \left(\sum_{x \in \partial\hat{S}} \min_{y \in \partial S} \|x - y\| \right. \tag{10}$$

$$\left. + \sum_{y \in \partial S} \min_{x \in \partial\hat{S}} \|y - x\| \right). \tag{11}$$

Here, $\partial\hat{S}$ and ∂S represent the boundary points of the predicted segmentation and the ground truth segmentation, respectively. The term $\|x-y\|$ denotes the Euclidean distance between two points, while $\min_{y \in \partial S} \|x - y\|$ computes the shortest distance from a boundary point $x \in \partial\hat{S}$ to the closest point in ∂S, ensuring an accurate local correspondence. The final ASD value represents the mean of these shortest distances, summing over both segmentations.

Hardware and Hyperparameters All our experiments were implemented using PyTorch on NVIDIA A5000 GPUs with 24 GB memory. We choose patch size as $1/16$ of the image size and use the mean absolute difference as the distance metric in Eq. (7) with $\tau = 50$. We use Adam optimizer [23] for Eq. (8) in the main paper.

B.2. Training Details

Source Segmentation Model All experiments are implemented in Python using the PyTorch framework. We train the segmentation models from scratch on the source datasets ACDC and M&M for each architecture, utilizing a hybrid segmentation loss combining Dice and cross-entropy. During training, we apply random data augmentation with a probability of 0.5, including random flipping, translation, and rotation. The models are optimized using the Adam optimizer with a learning rate of 1×10^{-4}. To align with the convention of adapting BatchNorm layers [55, 56, 63] in segmentation models, we replace GroupNorm in MedNeXt and InstanceNorm in SwinUNETR with BatchNorm. All segmentation models are trained with 150 epochs with a batch size of 8. We want to emphasize that our test-time adaptation framework operates under the assumption that the source segmentation model is pre-trained and provided.

Shape Energy Model Our region-based shape energy model is implemented as a simple convolutional neural network (CNN), designed to capture the localized nature of shape energy. The model comprises four convolutional layers, each with a kernel size of 5, stride of 2, and padding of 2. Each convolutional layer is followed by a LeakyReLU activation function with a negative slope of 0.2, as well as a BatchNorm layer for regularization and stability. Finally, the output is projected to a single channel using an additional convolutional layer as logits. For spatial augmentations, we introduce handcrafted spatial affine transformation and pixel-wise noise with probability p across all samples in the minibatch. Additionally, we apply patch-wise dropout to create holes in the initially augmented masks. These altered predictions are then added back to the original masks for further augmentation. We use the `BCEWithLogitsLoss` to exploit the logsumexp trick for training numerical stability. We use one-hot encoding for the ground truth mask and apply softmax activation for segmentation prediction logits as inputs for the shape energy model. We train our proposed region-based shape energy model using the Adam optimizer with 150 epochs. We use the cosine decay learning rate scheduler with a warm-up stage including 1000 steps. During the adaptation stage, we employ the Adam optimizer to update the collected BatchNorm parameters from the source segmentation model.

C. Additional Results

Quantitative Results We begin by presenting the quantitative evaluation of the source segmentation models, as shown Tab. 8. All models demonstrate strong performance in LV segmentation, achieving Dice scores above 90%, and myocardium segmentation, with Dice scores exceeding 80%. Myocardium segmentation is inherently more challenging due to its complex structure, being a thin, crescent-shaped layer surrounding the LV. Both MedNeXt and SwinUNETR achieve lower average surface distances compared to the vanilla UNet, highlighting the advantages of their advanced architectures. MedNeXt benefits from its ConvNeXt backbone, while SwinUNETR leverages transformer layers, both of which excel at capturing global context. This enables them to handle fine details more effectively and produce smoother, more accurate contours, particularly for complex structures like the myocardium.

Architecture	ACDC				M&M			
	LV		Myo		LV		Myo	
	DSC ↑	ASD ↓	DSC ↑	ASD ↓	DSC ↑	ASD ↓	DSC ↑	ASD ↓
UNet	90.16	3.81	82.56	4.21	92.99	2.75	84.45	2.38
MedNeXt	90.06	1.63	80.62	1.64	94.00	1.13	84.29	1.25
SwinUNETR	91.95	1.36	84.39	1.38	94.41	1.19	85.25	1.24

Table 8. Quantitative results for the source segmentation models (UNet, MedNeXt, and SwinUNETR) trained on the ACDC and M&M datasets are presented. The evaluation metrics include the DSC (%) and ASD (px).

Iterations (i)	LV		Myo	
	DSC ↑	ASD ↓	DSC ↑	ASD ↓
$i = 1$	64.85	16.26	51.20	13.81
$i = 3$	74.05	10.68	57.83	**9.48**
$i = 5$	73.94	11.60	58.66	10.42
$i = 10$	**76.93**	**8.77**	**59.43**	11.68

Table 9. Effect of the number of iterations of the proposed method during test-time adaptation on the UNet architecture for the ACDC $\mapsto$ LVQuant task. Evaluation metrics include the DSC (%) and ASD (px), with the best-performing results highlighted in bold.

We further analyze the effect of the number of iterations in the proposed method, as shown in Tab. 9. The results indicate that the Dice score improves consistently with an increasing number of iterations, with multiple iterations significantly outperforming a single update. This improvement occurs because a single gradient-based update represents only a linear step, which is insufficient to reach the typically nonlinear local minima required for optimal performance. We observe that the average surface distance for the myocardium increases as the number of iterations grows. This can be attributed to a limitation in our proposed approach, which lacks explicit regularization of the image during adaptation. Consequently, this may lead to the generation of shapes in unintended regions, ultimately contributing to the observed increase in the average surface distance.

Qualitative Results We present additional qualitative results of our proposed approach in Fig. 6, demonstrating its ability to effectively refine initial predictions and produce more plausible shapes after adaptation. The refined shapes are visually closer to the ground truth compared to other baselines, showcasing the effectiveness of our method. To

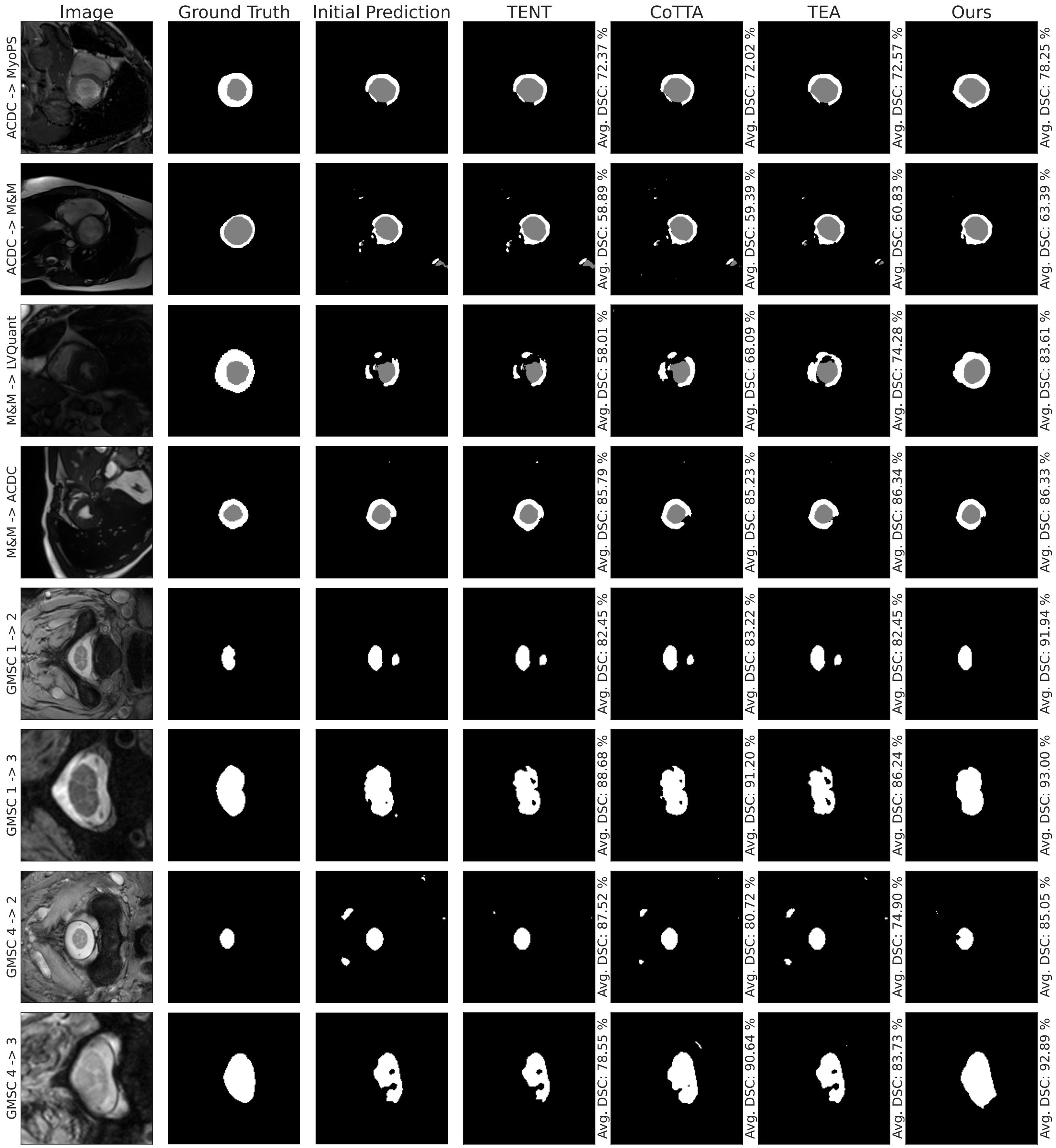

Figure 6. Qualitative evaluation of our proposed approach using the UNet architecture compared to baseline methods. The top four rows depict adapted cardiac segmentation, while the bottom four rows show lung segmentation from chest X-rays. Our approach effectively refines incomplete initial segmentations, generating more anatomically plausible shapes after adaptation.

further evaluate the performance of our approach in scenarios where imaging semantics are misaligned, we conduct an additional analysis by training on the public 2D ultrasound dataset CAMUS, which contains two-chamber views, and adapting to the CardiacUDA Site G dataset, which consists of four-chamber views. The qualitative results are shown

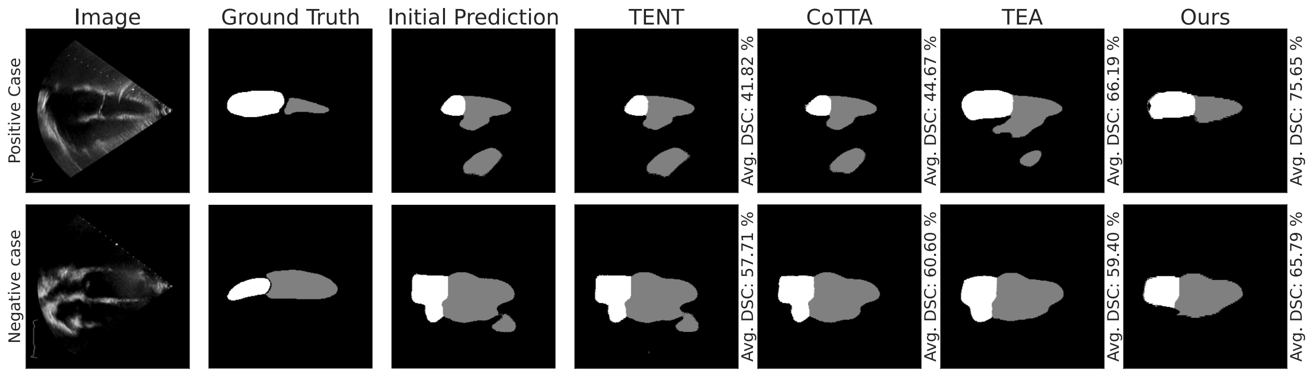

Figure 7. Qualitative evaluation of adaptation performance on source and target datasets with misaligned semantics is presented. Our proposed approach is trained on a 2D ultrasound dataset from CAMUS and adapted to the CardiacUDA Site G dataset. In the top row, we showcase a positive example where the initial prediction accurately identifies the right chambers. Conversely, the bottom row illustrates a negative case where the initial prediction incorrectly identifies the chamber locations. These examples highlight the challenges of semantic misalignment and the variability in adaptation outcomes.

Methods	UNet	MedNeXt	SwinUNETR
TENT	0.18 (-21.74%)	0.19 (-82.24%)	0.18 (-48.57%)
CoTTA	1.76 (+665.22%)	3.47 (+224.30%)	1.99 (+468.57%)
TEA	0.25 (+8.7%)	4.15 (+287.85%)	0.63 (+80.00%)
Ours	0.23	1.07	0.35

Table 10. Inference time per sample (in seconds) measured on a single NVIDIA RTX 2080 Ti GPU with 11 GB memory.

in Fig. 7. In the top row example, when the initial prediction correctly identifies the chamber, our proposed approach successfully removes outliers and generates more plausible shapes compared to the baselines. However, in cases where the initial prediction incorrectly identifies the chamber (bottom row example), our approach fails to correct the misprediction. We plan to address this limitation in future work by introducing enhanced regularization on the image during the adaptation process.

Computation Efficiency We evaluate the running time of our proposed approach and compare it with existing methods in Tab. 10. Measured on a single GPU during inference, our approach achieves an average speedup of 4.5× over CoTTA and 1.3× over TEA, while remaining comparable to TENT. This efficiency stems from our trained energy model, which requires only a forward pass, eliminating the need for extensive augmentation averaging in CoTTA or the stochastic gradient Langevin dynamics used to generate synthetic samples in TEA.

Representativeness of Simulated Negative Examples We further perform a t-SNE analysis (Fig. 8) of both image and segmentation features, using ACDC as in-distribution (ID) and others as OOD. Fig. 8 (left) shows a t-SNE plot of features encoded by the pretrained source model on ACDC (source) as in-distribution (ID) images, adversarially perturbed ACDC images (Adv.), and OOD images (real-world testing sets with covariate shifts). Fig. 8 (right) shows the t-SNE of energy model features of the segmentation produced by the pretrained model. Adversarially perturbed images and their resulting segmentations align with OOD images and their segmentations. This validates that indeed our perturbations model real covariate shifts.

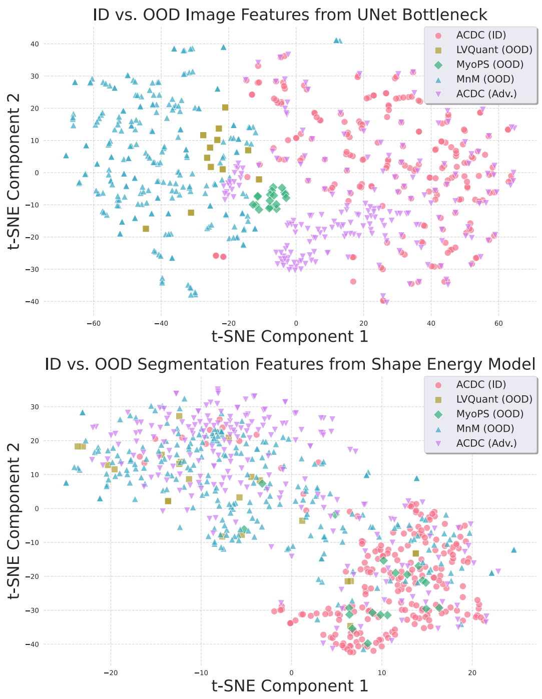

Figure 8. T-SNE analysis of both image (top) and segmentation features (bottom).

	† is proposed.	$1 \mapsto 2$	$1 \mapsto 3$	$1 \mapsto 4$	$4 \mapsto 1$	$4 \mapsto 2$	$4 \mapsto 3$	Avg.
Patch Size	4×4	69.1	73.2	93.4	89.4	42.5	85.3	75.5
	$9 \times 9^{\dagger}$	**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

Table 11. Hyperparameter sensitivity analysis on GMSC dataset, with sites 1 and 4 as source. Reported metrics include DSC (%).

$\Delta\phi$	0	0.01	0.03	0.05	0.07
Pretrained	58.98	55.20	47.95	48.29	35.39
TENT	65.78	67.20	58.30	52.39	45.42
Ours	**76.93**	**72.47**	**62.01**	**58.07**	**54.70**

Table 12. Quantitative analysis under varying degrees of simulated motion artifacts ($\Delta\phi$). Reported metric includes Dice (%).

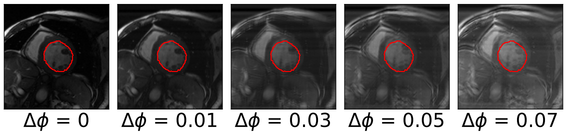

Figure 9. Visualization of simulated motion artifacts ($\Delta\phi$).

Hyperparameter Sensitivity Our choice of hyperparameters was determined empirically. We provide a sensitivity analysis in Tab. 11 of patch size, threshold τ, and perturbation magnitude δ on multiple adaptation scenarios on GMSC dataset (sites 1-4). While our method is not too sensitive to hyperparamters, our proposed choice achieves the overall best performance.

Performance Analysis Under Varying degress of artifacts We present an analysis under varying realistic motion artifacts. We simulate motion blur following [64] by applying FFT and introducing random phase shifts in k-space ($\Delta\phi \in [0, 0.07]$), capturing mild to severe artifacts as shown in Fig. 9. From Tab. 12, although performance of all methods degrades gracefully with increasing artifact severity, ours remains most robust.

	† is proposed.	$1 \mapsto 2$	$1 \mapsto 3$	$1 \mapsto 4$	$4 \mapsto 1$	$4 \mapsto 2$	$4 \mapsto 3$	Avg.
Comparison	DeTTA [1]	66.7	70.3	91.9	91.3	55.2	87.1	77.1
	Ours + $\mathcal{L}_{\text{s}}$	73.1	77.5	**95.4**	95.2	54.4	87.6	80.5
	Ours (multi)	**74.2**	74.8	94.4	94.3	54.0	**88.7**	80.1
	Ours (single)†	73.6	**77.7**	95.3	**95.1**	**56.2**	87.2	**80.9**

Table 13. Quantitative comparisons of adapted predictions for spinal cord MRI segmentation, with sites 1 and 4 in GMSC as the source dataset. Reported metrics include DSC (%).

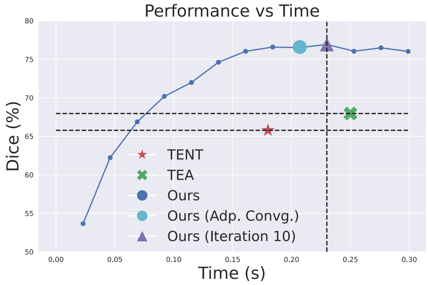

Figure 10. Convergence analysis under the same compute budget.

More Comparisons We further discuss relevant works and present quantitative comparisons across different configurations. Unlike DeTTA [59], which is UNet-specific and requires separate denoising pretraining, our method is model-agnostic. Methods [24, 31, 50] operate under different settings. Post-DAE [24] is post-processing without model adaptation, while MAS [50] and SFDA [31] rely on multiple passes over unlabeled target data, violating TTA assumptions. Amongst which, DeTTA provided code, so we compare with them in Tab. 13 where our method outperforms. We also evaluate effect of using multi-scale energy functions. Tab. 13 shows comparison with single-scale and multi-scale energy (scoring) function with output at spatial scales 4×4, 9×9, 18×18, where multiscale performs comparably. We assess the impact of adding a structure-aware smoothness loss. Tab. 13 shows optimizing a (image) structure-aware local smoothness loss ($\mathcal{L}_{\text{s}}$) along with binary cross entropy yields little difference. We hypothesize that the small difference is due to the architectural inductive bias from the convolutional feature pyramid, which captures spatial and scale continuity.

Fairness in Compute Budget and Convergence Per convention in prior works, there is an allowable time budget for updates rather than a fixed number of updates. Fig. 10 shows that we outperform others under the same time budget in ACDC $\mapsto$ LVQuant. An adaptive criterion stopping after small energy changes halts at 9 steps, which results in a minimal Dice change and still outperforms other methods. There are diminished gains over longer runs.

Calibration Analysis We perform a calibration analysis for our adapted left ventricle segmentation in ACDC $\mapsto$ LVQuant using UNet, which shows our approach is reasonably calibrated as shown in Fig. 11.

Applicability to 3D Volumes. Our method is viable for real-time deployment. Average 3D MRI acquisition time is ∼6 mins. We analyzed time complexity by exponentially scaling batch size (1-8), recording runtimes of 0.16, 0.40,

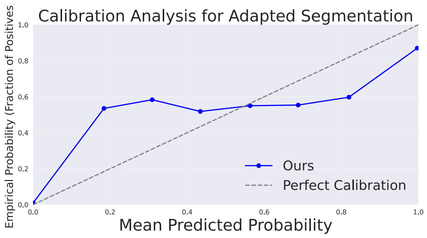

Figure 11. Calibration analysis (ACDC↦LVQuant, UNet).

1.81, 4.99 mins on 3D volumes of 128×128×128, confirming $O(N)$ w.r.t. size N. Thus, our adaptation completes before the next scan finishes acquisition.