

Wake the Sleeping Weights: Sparsely-Activated Continual Test-Time Adaptation for Medical Image Segmentation

Jianhang Ji¹ Zhiming Cheng^{2†} Jianxiang Zhao² Bingtao Ma²
Hao Chen¹ Yuhan Gao^{2,3} Lian Zhang⁴ Zuobin Ying^{1†} Shuai Wang^{2†}

¹City University of Macau ²Hangzhou Dianzi University

³Lishui Institute of Hangzhou Dianzi University ⁴The First Hospital of Hebei Medical University

jijianhang.ai@gmail.com czming@hdu.edu.cn zbying@cityu.edu.mo shuaiwang.tai@gmail.com

Abstract

Medical image distributions continuously shift across centers, scanners, and acquisition protocols, undermining the reliable deployment of pre-trained segmentation models in clinical practice. Continual test-time adaptation (CTTA) updates models online using unlabeled test streams, but noisy supervision repeatedly overwrites pre-trained knowledge, causing error accumulation and degrading reliability, while repeated sample-specific updates induce catastrophic forgetting and undermine long-term stability. We observe that many parameters in a pre-trained model have near-zero magnitude and negligible influence: zeroing these inactive parameters barely changes performance, whereas tuning only them yields consistent gains. Based on this insight, we propose Sparsely-Activated Continual Test-Time Adaptation (TACT), a two-stage framework that updates only inactive parameters at test time. Stage 1 (Sparse Domain Alignment) builds shift-robust representations via style and structure perturbations with prediction- and feature-level consistency. Stage 2 (Progressive Sample Refinement) performs per-sample refinement with a hard-pixel contrastive loss, temporarily updating only parameters that remain inactive and resetting these changes after inference to avoid sample-specific interference. On multi-center medical image segmentation benchmarks, TACT achieves state-of-the-art adaptability and long-term stability under continual domain shifts. Code is available at <https://github.com/M4cheal/TACT>.

1. Introduction

Semantic segmentation is fundamental to computer-aided diagnosis, enabling automated delineation of anatomical structures in medical images [1, 2, 46, 60]. However, distribution shift is widespread in clinical practice, as data are

routinely collected from new scanners, protocols, or patient cohorts not seen during training. This shift severely limits the reliability of pre-trained segmentation models in real-world deployment [13, 45]. Domain adaptation (DA) aims to bridge this gap, but it relies on both labeled source data and large, representative target datasets for offline retraining [3, 7, 8, 14, 58, 59], conditions rarely met in practice due to privacy and data scarcity [33, 61]. Test-time adaptation (TTA) offers a more practical solution, adapting models online using only unlabeled test data [19, 39, 43, 55, 62, 63]. Nevertheless, most TTA methods assume a stationary target distribution, which seldom holds in dynamic clinical environments [9]. Continual test-time adaptation (CTTA) extends TTA by allowing models to adapt to continually changing, unlabeled target domains, closely reflecting real-world clinical deployment [9, 36, 56, 57]. However, CTTA remains challenging due to error accumulation [6, 26], where noisy pseudo-labels cause model degradation, and catastrophic forgetting, where sample-specific updates erase previously acquired knowledge [21, 40].

CTTA methods are typically grouped into three families: full parameter adaptation [34, 56], additive parameter adaptation [9, 30], and layer selection adaptation [36, 41, 55]. As illustrated in Fig. 1(a–c), under noisy supervision and continual shifts, the trainable parts in these CTTA schemes are repeatedly overwritten, eroding pre-trained knowledge and manifesting as the above error accumulation and catastrophic forgetting. Although additive methods freeze the backbone, the added parameters typically require source-based initialization, which is infeasible in clinical settings due to data privacy [34], and the attached modules are overwritten by noisy supervision over time. These limitations motivate a key question: *How can we achieve reliable and stable adaptation to continual distribution shifts while minimizing interference with pre-trained knowledge?*

Motivated by advances in network pruning, we revisit the source model and observe that many parameters have

[†] Corresponding authors.

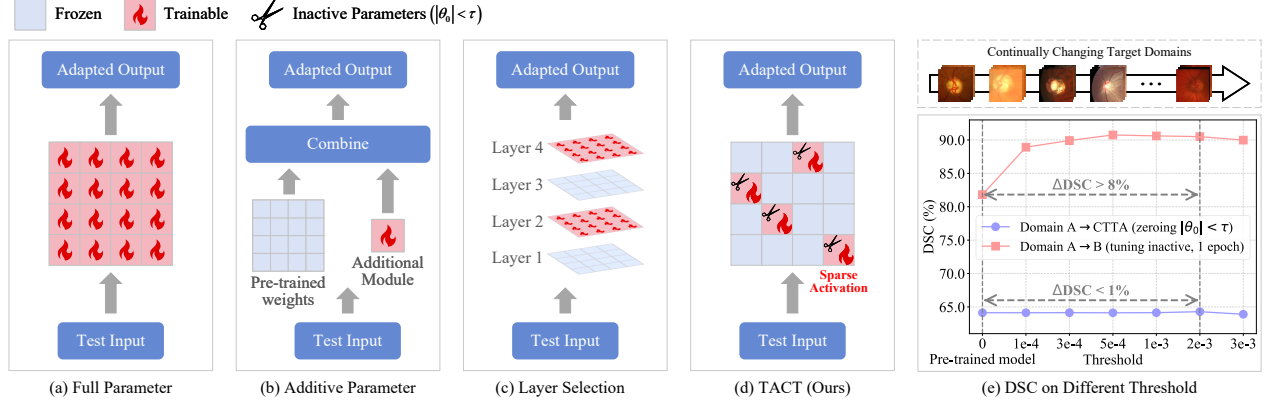


Figure 1. Comparison of mainstream CTTA strategies. (a) Full parameter adaptation updates all weights and is prone to error accumulation and catastrophic forgetting under noisy supervision. (b) Additive parameter adaptation (adapters/prompts atop a frozen backbone) may mitigate forgetting but is capacity-limited and sensitive to initialization, drifting over long streams. (c) Layer selection adaptation (e.g., BN/LN-only or dynamically selected layers) narrows the update scope yet still overwrites the trainable parts over time. (d) Our TACT confines updates to a sparse subspace of inactive parameters ($|\theta_0| < \tau$), preserving pre-trained knowledge while enabling robust continual adaptation. (e) Inactive-parameter analysis across thresholds τ : zeroing parameters with $|\theta_0| < \tau$ causes minimal change ($\Delta\text{DSC} < 1\%$), whereas supervised fine-tuning of only these parameters for one epoch on a new domain yields a substantial gain ($\Delta\text{DSC} > 8\%$).

near-zero magnitude ($|\theta_0| < \tau$) and negligible influence on predictions. We refer to these as inactive parameters and view them as a natural, low-interference subspace for adaptation. Empirically, across thresholds τ , zeroing these parameters has little effect on performance, while selectively fine-tuning them yields measurable gains (see Fig. 1(e)). This observation motivates TACT’s design, which confines updates to this sparse subset (see Fig. 1(d)), enabling robust adaptation with minimal disruption to pre-trained knowledge. In contrast to layer selection adaptation, we update at the parameter level and enforce sparsity within layers to reduce interference with pre-trained knowledge.

Based on this insight, we propose **Sparsely-Activated Continual Test-Time Adaptation (TACT)**, a two-stage framework for robust continual adaptation that preserves pre-trained knowledge. TACT confines all parameter updates to a sparse subspace of inactive parameters, minimizing interference with source-domain representations. In Stage 1 (Sparse Domain Alignment), the model applies style perturbation (FFT-based amplitude swap) and structure perturbation (entropy-based masking) to the inputs, and enforces prediction- and feature-level consistency to improve robustness to distribution shifts. However, domain shifts often introduce sample-specific ambiguities that consistency regularization alone cannot resolve. In Stage 2 (Progressive Sample Refinement), we apply a hard pixel contrastive loss and further refine the model for each target sample while temporarily updating only the parameters that remain inactive after Stage 1. After each inference, these per-sample updates are reset to preserve long-term knowledge and prevent sample-specific interference.

Our main contributions are as follows: 1) We identify the

inactive-parameter sparse subspace as a low-interference and principled locus for robust continual test-time adaptation. 2) We propose TACT, a progressive two-stage CTTA confined to the inactive-parameter subspace: Stage 1 stabilizes via input style and structure perturbation consistency, and Stage 2 performs per-sample refinement with hard-pixel contrastive learning and a reset after each inference. 3) TACT achieves state-of-the-art performance and long-term stability on multi-center medical segmentation benchmarks.

2. Related Work

2.1. Test-Time Adaptation (TTA)

Test-time adaptation (TTA) enables pre-trained models to adapt to unlabeled target domains during inference, without requiring source data [39, 40, 49, 53, 55, 62, 63]. Methods such as TENT [55] minimize prediction entropy, while DUA [37] and TTN [28] adapt batch normalization statistics. Although effective for isolated shifts, these approaches assume a stationary target distribution, which rarely holds in clinical applications with evolving domains.

Continual test-time adaptation (CTTA) generalizes TTA to non-stationary target streams [56], closely resembling real-world deployment. CTTA methods are grouped into three families according to which parameters are updated at test time. Full parameter adaptation updates all weights at test time [11, 34, 56]. CoTTA [56] uses an EMA teacher with augmentation-averaged pseudo labels and stochastic weight restoration. SURGEON [34] enforces dynamic activation sparsity to enable full parameter updates under tight memory budgets. These methods are highly plastic but vulnerable to noisy supervision and may accumulate

errors and forget. Additive parameter adaptation freezes the source backbone and attaches small trainable modules [9, 22, 30, 51]. VPTTA [9] uses visual prompts for 2D medical segmentation. ViDA [30] uses domain adapters with homeostatic fusion of low-rank and high-rank branches. Prompts have limited capacity, adapters often need source-based initialization, and attached modules are overwritten by noisy test-time supervision, causing forgetting and bias. Layer selection adaptation restricts updates to a subset of internal layers [36, 41, 44, 55]. SAR [41] updates only normalization layers and performs sharpness-aware reliable entropy minimization. PALM [36] dynamically selects layers by prediction uncertainty and adjusts learning rates by shift sensitivity. These methods reduce interference but the updated parts are still repeatedly overwritten during continual adaptation. In contrast, TACT confines updates to a sparse subspace of inactive parameters, preserving pre-trained knowledge while enabling robust continual adaptation.

2.2. Parameter-Efficient Fine-Tuning (PEFT)

Parameter-efficient fine-tuning (PEFT) reduces computation and memory by updating only a subset of model parameters [32, 47]. PEFT strategies include additive [23, 52], selective [18, 25], reparameterization [17, 32], and hybrid approaches [15]. Selective fine-tuning updates parameters based on specific criteria, such as magnitude, as in PaFi [27] and SaRA [18], or Fisher information, as in FishMask [54] and Fish-Dip [10]. However, most PEFT methods are designed for static or supervised settings, and do not address the online, unsupervised, and non-stationary requirements of CTTA. In this work, we leverage the inactive-parameter subspace for continual test-time adaptation and introduce a progressive, resettable two-stage mechanism, providing robust adaptation in challenging clinical environments.

3. Method

3.1. Problem Definition and Method Overview

Continual Test-Time Adaptation (CTTA) aims to deploy a source model f_{θ_0} , trained on labeled data $\mathcal{D}_S = \{X_S, Y_S\}$, in a sequence of unlabeled and non-stationary target domains $\{\mathcal{D}_{T_1}, \mathcal{D}_{T_2}, \dots\}$. In this setting, source data are not available and each target sample is encountered only once. The objective is to continually adapt the model to non-stationary target distributions while maintaining reliable performance. We propose TACT, a two-stage CTTA framework that confines adaptation to a sparse subspace of inactive parameters in the pre-trained model, thus minimizing interference with source-domain knowledge. An overview of TACT is illustrated in Fig. 2.

3.2. Inactive Parameter Selection

TACT restricts adaptation to a sparse subspace of inactive parameters, defined as those with small magnitudes and minimal contribution to source-domain predictions. We formally define the sparse inactive-parameter mask as

$$\mathcal{M} = \mathbb{I}(|\theta_0| < \tau), \quad (1)$$

where τ is the sparse threshold. During adaptation, only entries selected by $\mathcal{M}$ are updated:

$$\theta \leftarrow \theta - \eta (\nabla_{\theta} \mathcal{L}) \odot \mathcal{M}, \quad (2)$$

where $\odot$ denotes element-wise multiplication. This selective update strategy enables continual adaptation while minimizing interference with source-domain knowledge.

3.3. Stage 1: Sparse Domain Alignment

Stage 1 aims to improve robustness to distribution shifts by applying style and structure perturbations and enforcing prediction- and feature-level consistency.

Style Perturbation. We introduce style variation in the frequency domain to encourage robustness to style shifts. For each input x , we perform a Fast Fourier Transform (FFT) to obtain its amplitude $\mathcal{F}^A(x)$ and phase $\mathcal{F}^P(x)$. A central low-frequency patch $\mathcal{F}_{\mathcal{R}}^A(x)$ is extracted, where the patch height and width are α times those of the input. We maintain a style memory bank $\mathcal{B}_{style}$ (size N_{style}) updated in First In First Out (FIFO). For each x , we retrieve the nearest neighbor q^* of $\mathcal{F}_{\mathcal{R}}^A(x)$ in $\mathcal{B}_{style}$ by cosine similarity, replace the patch, and update the bank accordingly. We choose the nearest to keep the perturbation small and realistic, and to stabilize consistency training (see Sec. 4.4). The perturbed image is then reconstructed by inverse FFT:

$$\tilde{x} = \mathcal{F}^{-1} \left(\mathcal{F}^A(x) \big|_{\mathcal{R} \leftarrow q^*}, \mathcal{F}^P(x) \right). \quad (3)$$

where $\mathcal{F}^{-1}$ denotes the inverse FFT. This induces plausible style changes while preserving spatial structure, improving robustness to distribution shifts.

Structure Perturbation. To improve robustness to domain shifts, we mask high-uncertainty regions in the input, prompting the model to reconstruct ambiguous areas using contextual cues. For each x , we compute a normalized entropy map $E_{\theta_0}(x)$ from the source model. Pixels above the 90th percentile are treated as high-uncertainty. Denoting the number of high-uncertainty pixels as n_{high} , we randomly select $N = \max(1, \lfloor 0.1 n_{high} \rfloor)$ centers, each masked by a $k \times k$ patch, where $k = \max(1, \lfloor \sqrt{n_{high}} \rfloor)$. This choice increases the masking strength with the amount of uncertainty through N , and sets the patch side length proportional to $\sqrt{n_{high}}$ so that the mask size adapts to the spatial extent of the uncertain area. We apply these $k \times k$ masks to the style-perturbed $\tilde{x}$ to obtain $\tilde{x}_{mask}$, which encourages the model to infer uncertain regions from reliable context.

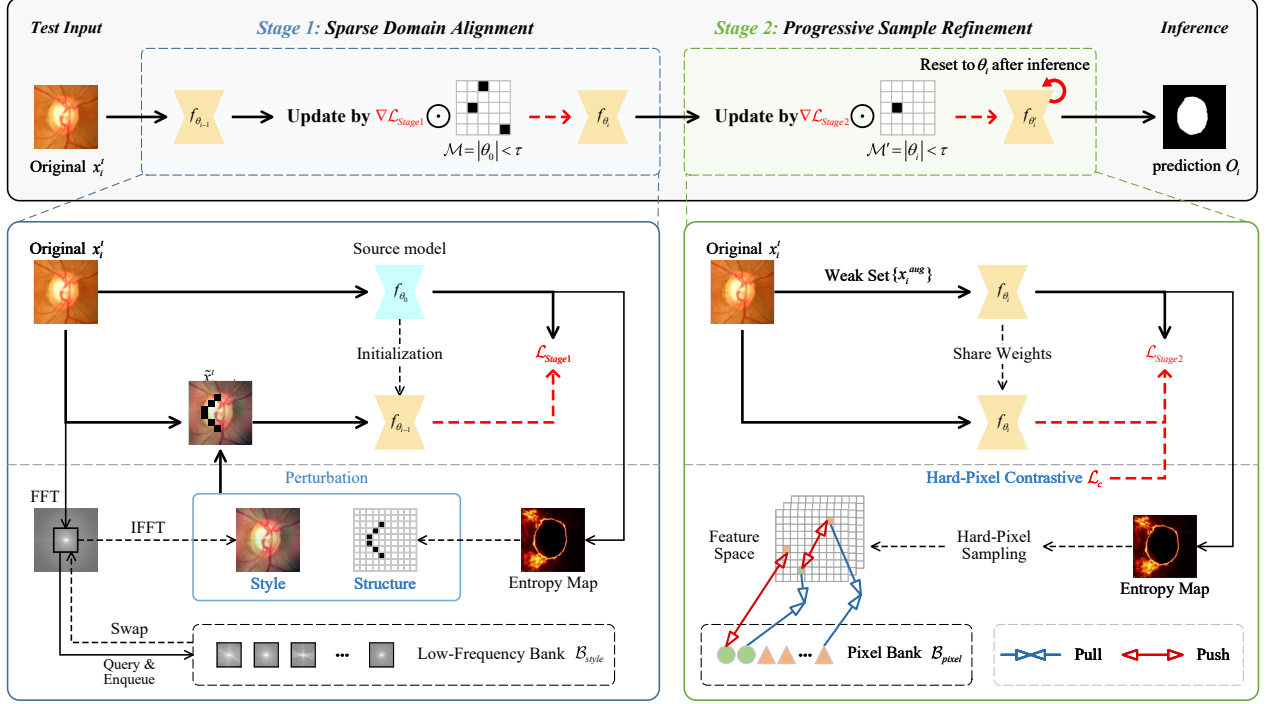


Figure 2. Overview of TACT. For each test image x_i^t , adaptation is restricted to the inactive parameters $\mathcal{M} = |\theta_0| < \tau$. Stage 1 (Sparse Domain Alignment) applies style perturbation (FFT-based amplitude swap) and structure perturbation (entropy-based masking), and enforces consistency losses $\mathcal{L}_{con}^{pred}$ and $\mathcal{L}_{con}^{feat}$, updating parameters with $\nabla \mathcal{L}_{Stage1}$. Stage 2 (Progressive Sample Refinement) selects hard pixels from the entropy map and applies a pixel-wise contrastive loss $\mathcal{L}_c$ and prediction consistency $\mathcal{L}_{ent}$; parameters in $\mathcal{M}' = |\theta_i| < \tau$ are updated by minimizing $\mathcal{L}_{Stage2}$, and all sample-specific updates θ_i' are reset to θ_i after inference.

Consistency Regularization. To stabilize adaptation under perturbations, TACT enforces consistency at both the prediction and feature levels. For each target input, pseudo-labels $\hat{y}$ are generated by combining high-confidence predictions and high-entropy regions:

$$\hat{y} = \mathbb{I}(p \geq \tau_p) \vee \mathbb{I}(E_{\theta_0}(x) \geq \tau_e), \quad (4)$$

where $p = f_{\theta_0}(x)$ denotes the source model's sigmoid probabilities on the unperturbed input, and τ_p and τ_e are thresholds. Consistency is enforced with two losses:

$$\begin{aligned} \mathcal{L}_{con}^{pred} &= \mathcal{L}_{seg}(\tilde{p}, \hat{y}) \\ \mathcal{L}_{con}^{feat} &= \|\mathcal{F}(\tilde{\phi}) - \mathcal{F}(\phi)\|_1 \end{aligned} \quad (5)$$

where $\tilde{p}$ and $\tilde{\phi}$ are the prediction and feature map from the perturbed input $\tilde{x}_{mask}$ using $f_{\theta_{i-1}}$, ϕ is the feature map from the original input x using f_{θ_0} , and $\mathcal{F}$ denotes the FFT. Throughout, ϕ denotes the final decoder feature map before the prediction head. The total loss for Stage 1 is:

$$\mathcal{L}_{Stage1} = \mathcal{L}_{con}^{pred} + \lambda \mathcal{L}_{con}^{feat}, \quad (6)$$

where λ is a weighting coefficient.

3.4. Stage 2: Progressive Sample Refinement

Stage 1 improves robustness but may miss sample-specific details crucial for target performance. To address this, we introduce progressive sample refinement, which adapts the model to each target input while preserving accumulated knowledge. After Stage 1 we obtain parameters θ_i . For each target sample x , we compute a temporary sparse mask $\mathcal{M}' = \mathbb{I}(|\theta_i| < \tau)$ and update only the entries selected by $\mathcal{M}'$. This produces sample-adapted parameters θ_i' . After inference, we evaluate $f_{\theta_i'}(x)$ and then reset $\theta_i' \leftarrow \theta_i$, which preserves accumulated knowledge and yields fine-grained refinement without sacrificing long-term stability.

3.4.1. Hard-Pixel Contrastive Loss.

To effectively capture sample-specific ambiguities in the target domain, we introduce a hard-pixel contrastive loss that explicitly drives the model to focus on the most uncertain and informative pixels. Specifically, we define hard pixels as those with high entropy and disagreement between the predicted mask $\bar{y} = \mathbb{I}(\bar{p} \geq \tau_p)$ (with $\bar{p} = f_{\theta_i}(x)$) and a low-entropy pseudo label $\check{y} = \mathbb{I}(\bar{p} \geq \tau_p) \wedge \mathbb{I}(E_{\theta_i}(x) < \tau_e)$, where $\wedge$ denotes element-wise AND. For each class k , we identify hard pixels with $\bar{y}_i^{(k)} = 1$ and $\check{y}_i^{(k)} = 0$, and select

Algorithm 1 TACT Algorithm

Require: Frozen source model θ_0 , inactive mask $\mathcal{M}$, sparse threshold τ , memory bank $\mathcal{B}_{style}$, memory bank $\mathcal{B}_{pixel}$

Input: For each time step i , current target model θ_{i-1} , current test image x_i^t

Output: Adapted prediction O_i

▷ **Stage 1 - Sparse Domain Alignment:**

- 1: Obtain $\tilde{x}_i^t$ using $\mathcal{B}_{style}$ by Eq. (3)
- 2: Obtain $\tilde{x}_{i,mask}^t$ using $E_{\theta_0}(x_i^t)$ by masking $\tilde{x}_i^t$
- 3: Forward $f_{\theta_{i-1}}(\tilde{x}_{i,mask}^t)$ and $f_{\theta_0}(x_i^t)$, compute $\mathcal{L}_{Stage1}$ by Eq. (6), and obtain θ_i by updating θ_{i-1} under $\mathcal{M}$ using Eq. (2)

▷ **Stage 2 - Progressive Sample Refinement:**

- 4: Obtain temporary sparse mask $\mathcal{M}'$ from θ_i and τ (same form as Eq. (1))
- 5: Forward $f_{\theta_i}(x_i^t)$ and $f_{\theta_i}(\{x_i^{t,aug}\})$, compute $\mathcal{L}_{Stage2}$ by Eq. (10), and obtain θ'_i by updating θ_i under $\mathcal{M}'$ using Eq. (2)

▷ **Inference:**

- 6: Forward $O_i = f_{\theta'_i}(x_i^t)$
 - 7: Update the memory banks $\mathcal{B}_{style}$ and $\mathcal{B}_{pixel}$
 - 8: Reset $\theta'_i \leftarrow \theta_i$
 - 9: **return** O_i
-

the top n by entropy:

$$\mathcal{S}_{hard}^{(k)} = \text{Top}_n\left(\{i \mid \bar{y}_i^{(k)} = 1, \check{y}_i^{(k)} = 0\}; E_{\theta_i}(x)_i\right). \quad (7)$$

Here $\text{Top}_n(S; w)$ returns the n elements $i \in S$ with the largest scores $w(i)$. Each selected hard-pixel feature acts as an anchor in contrastive learning. We maintain a per-class pixel memory bank $\mathcal{B}_{pixel}^{(k)}$ with capacity N_{pixel} , updated in FIFO order. For each class k , features from the selected hard pixels, with pseudo labels computed by the current model (as in Eq. (4) but using f_{θ_i}), are detached and enqueued into $\mathcal{B}_{pixel}^{(k)}$ to provide positives and negatives for contrastive matching. The loss is:

$$\mathcal{L}_c = -\frac{1}{N} \sum_{i=1}^N \frac{1}{|P_i|} \sum_{z_i^+} \log \frac{e^{z_i \cdot z_i^+ / \tau}}{e^{z_i \cdot z_i^+ / \tau} + \sum_{z_i^-} e^{z_i \cdot z_i^- / \tau}}, \quad (8)$$

where N is the number of anchors and z_i is the anchor feature, with P_i and N_i the positive and negative sets. The inner sums run over $z_i^+ \in P_i$ and $z_i^- \in N_i$, and τ is the temperature. This loss encourages class-consistent, target-specific representations.

3.4.2. Prediction Consistency.

To further stabilize adaptation, we apply prediction consistency between each input and its geometric augmentations $\{x^{aug}\}$ (e.g., flips, rotations). For each input x , let $\bar{p}$ denote

the predicted probabilities and let $\bar{p}^{aug}$ be the mean prediction over the de-augmented augmentations. The loss is

$$\mathcal{L}_{ent} = -\frac{1}{K} \sum_{c=1}^K \bar{p}_c \log \bar{p}_c^{aug}, \quad (9)$$

where K is the number of classes. The final Stage 2 loss is

$$\mathcal{L}_{Stage2} = \beta \mathcal{L}_{ent} + \gamma \mathcal{L}_c, \quad (10)$$

where β and γ are weights. After inference, we reset $\theta'_i \leftarrow \theta_i$ so that sample-specific updates do not accumulate across samples. The complete TACT procedure is detailed in Algorithm 1.

4. Experiments

We evaluate TACT on three medical image segmentation benchmarks: the joint optic disc (OD) and cup (OC) segmentation task, the prostate segmentation task, and the polyp segmentation task.

4.1. Datasets and Evaluation Metrics

The OD/OC segmentation benchmark [9] uses five publicly available fundus image datasets as distinct domains: A (RIM-ONE-r3 [12]), B (REFUGE [42]), C (ORIGA [64]), D (REFUGE validation/test [42]), and E (Drishti-GS [50]). The five domains contain 159, 400, 650, 800, and 101 images, respectively. All images are resized to 512×512 . We report the Dice coefficient (DSC) as the evaluation metric.

The prostate segmentation benchmark [35] uses T2-weighted MRI from three publicly available prostate datasets, partitioned into six domains: A (BIDMC), B (BMC), C (HK), D (I2CVB), E (RUNMC), and F (UCL) [5, 24, 29, 31]. The six domains contain 261, 384, 158, 468, 421, and 175 slices, respectively. Each 2D slice is resized to 384×384 . We report the DSC for evaluation.

The polyp segmentation benchmark [9] uses four publicly available datasets as distinct domains: A (BKAI-IGH-NeoPolyp [38]), B (CVC-ClinicDB [4]), C (ETIS-LaribPolypDB [48]), and D (Kvasir-Seg [20]). The four domains contain 1000, 612, 196, and 1000 images, respectively. All images are resized to 352×352 . We report three metrics: DSC for region overlap, structure similarity index measure (S_α) for structural consistency, and enhanced-alignment measure ($E_\phi^{\max}$) for boundary alignment.

4.2. Implementation Details

In the CTTA setting, we take each domain in turn as the source and train a source model. We then evaluate on the remaining domains, which arrive sequentially as a continuously changing target stream. We report the mean performance over the target domains. For source training, we follow [9] to train a ResUNet-34 [16] backbone as the baseline on all three segmentation benchmarks. During test-time

Table 1. Comparison with state-of-the-art CTTA methods on OD/OC and prostate segmentation. Best and second-best scores are highlighted in **bold** and underline, respectively. In OD/OC segmentation, Do. A (Domain A) denotes the scenario with source A and targets B $\rightarrow$ C $\rightarrow$ D $\rightarrow$ E, and results are averaged over the target domains.

Methods	Venue	OD/OC Segmentation						Prostate Segmentation							
		Do. A	Do. B	Do. C	Do. D	Do. E	Average	Do. A	Do. B	Do. C	Do. D	Do. E	Do. F	Average	
		<i>DSC</i>	<i>DSC</i>	<i>DSC</i>	<i>DSC</i>	<i>DSC</i>	<i>DSC</i> ↑	<i>DSC</i>	<i>DSC</i>	<i>DSC</i>	<i>DSC</i>	<i>DSC</i>	<i>DSC</i>	<i>DSC</i> ↑	
Source Only	-	64.13	74.23	69.99	56.09	65.17	65.92	58.49	59.23	27.95	26.84	53.54	48.79	45.81	
TENT [55]	ICLR'21	68.40	78.72	72.62	47.63	75.45	68.57	69.95	40.36	11.50	43.07	66.69	16.79	41.39	
CoTTA [56]	CVPR'22	63.44	71.90	65.47	66.81	59.74	65.47	63.44	59.13	45.93	40.65	63.54	61.65	55.72	
SAR [41]	ICLR'23	66.33	78.35	70.04	62.37	67.96	69.01	66.08	61.05	50.65	39.55	64.12	60.65	57.02	
ViDA [30]	ICLR'24	64.87	71.57	69.13	58.40	63.59	65.51	57.86	59.56	28.64	27.97	55.27	51.10	46.73	
VPTTA [9]	CVPR'24	66.95	76.34	71.24	60.05	68.11	68.54	66.68	62.46	41.12	32.47	61.38	60.95	54.17	
PALM [36]	AAAI'25	66.55	78.59	68.32	58.36	72.45	68.85	66.87	61.10	43.80	40.35	66.10	58.96	56.20	
SURGEON [34]	CVPR'25	67.70	78.41	69.11	59.14	70.50	68.97	66.50	60.96	50.42	41.74	66.22	60.23	57.68	
TACT	This paper	71.81	78.80	75.60	60.20	76.12	72.51	66.97	64.47	45.63	47.38	67.33	63.12	59.15	

Table 2. Comparison with state-of-the-art CTTA methods on polyp segmentation.

Methods	Venue	Domain A			Domain B			Domain C			Domain D			Average		
		<i>DSC</i>	$E_{\phi}^{\max}$	S_{α}	<i>DSC</i>	$E_{\phi}^{\max}$	S_{α}	<i>DSC</i>	$E_{\phi}^{\max}$	S_{α}	<i>DSC</i>	$E_{\phi}^{\max}$	S_{α}	<i>DSC</i> $\uparrow$	$E_{\phi}^{\max} \uparrow$	$S_{\alpha} \uparrow$
		<i>DSC</i>	$E_{\phi}^{\max}$	S_{α}	<i>DSC</i>	$E_{\phi}^{\max}$	S_{α}	<i>DSC</i>	$E_{\phi}^{\max}$	S_{α}	<i>DSC</i>	$E_{\phi}^{\max}$	S_{α}	<i>DSC</i> $\uparrow$	$E_{\phi}^{\max} \uparrow$	$S_{\alpha} \uparrow$
Source Only	-	73.93	83.35	80.95	62.48	78.60	74.42	62.35	80.94	73.06	73.10	84.10	82.56	67.96	81.75	77.75
TENT [55]	ICLR'21	66.24	78.55	75.63	64.45	79.23	75.37	24.64	46.26	55.61	68.72	79.74	79.43	56.01	70.94	71.51
CoTTA [56]	CVPR'22	72.14	83.53	78.69	59.00	73.43	70.67	65.21	82.99	74.26	56.26	68.67	70.93	63.15	77.16	73.64
SAR [41]	ICLR'23	70.30	82.14	77.62	62.34	76.64	72.98	64.14	80.93	74.27	61.20	73.01	74.31	64.49	78.18	74.79
ViDA [30]	ICLR'24	75.42	84.56	81.89	61.17	76.93	73.58	62.18	80.49	72.80	72.86	83.94	82.38	67.91	81.48	77.66
VPTTA [9]	CVPR'24	74.35	84.01	80.75	64.55	80.91	75.64	65.48	82.31	75.22	72.66	84.87	81.76	<u>69.26</u>	<u>83.02</u>	<u>78.35</u>
PALM [36]	AAAI'25	69.72	81.69	77.31	63.59	77.67	73.93	63.26	79.86	73.87	62.55	74.25	75.20	64.78	78.37	75.08
SURGEON [34]	CVPR'25	70.31	82.11	77.67	63.50	77.37	73.78	63.55	79.97	74.04	61.86	73.57	74.69	64.80	78.26	75.04
TACT	This paper	74.98	84.37	81.50	67.02	81.17	77.20	67.82	83.60	76.97	74.70	86.60	82.64	71.13	83.94	79.58

adaptation, online updates are performed per target image with batch size 1. All competing methods are evaluated under the same protocol to ensure a fair comparison. For TACT, we use AdamW with a learning rate of 1×10^{-4} and set the inactivity threshold to $\tau = 2 \times 10^{-4}$. We use the following hyperparameters across all segmentation tasks: $\alpha = 0.01$ (low-frequency patch ratio), $N_{style} = 40$ (style bank size), $N_{pixel} = 500$ (pixel bank size), $\tau_p = 0.5$ (prediction threshold), $\tau_e = 0.75$ (entropy threshold), $T = 0.5$ (temperature), and $\lambda = \beta = \gamma = 0.1$ (loss weights). TACT is implemented in PyTorch 2.2.2, and all experiments are conducted on a single NVIDIA RTX 3090 GPU.

4.3. Comparison with State-of-the-Art Methods

We evaluate TACT on three CTTA benchmarks. “Source Only” means direct testing without adaptation. We compare against seven representative baseline methods: full parameter adaptation (CoTTA [56] and SURGEON [34]), layer selection adaptation (TENT [55] and SAR [41] updating only normalization layers, and PALM [36] dynamically selecting layers), and additive parameter adaptation (freezing the source model and attaching small trainable modules, with VPTTA [9] employing visual prompts for 2D medical image segmentation and ViDA [30] adopting domain adapters). We reimplement all competing methods using the same source model as TACT. For each domain, results are computed by using that domain as the source and applying

the pretrained model to the remaining domains in sequence.

OD/OC segmentation. Table 1 compares TACT with the “Source Only” baseline and seven CTTA methods. Most adaptation methods outperform “Source Only”, indicating that test-time adaptation helps under cross-domain shifts. TACT achieves the best overall performance and remains competitive across all domains, both where baselines are already strong and where several are weaker.

Prostate segmentation. Table 1 presents results on the prostate benchmark, which shows larger variation across centers and more pronounced domain shifts than OD/OC. On challenging domains, the entropy minimization method TENT can fall below “Source Only”. TACT improves over “Source Only” and outperforms the compared methods, achieving the best mean performance.

Polyp segmentation. Table 2 reports results on the polyp benchmark. Full parameter adaptation (CoTTA, SURGEON) and layer selection adaptation (TENT, SAR, PALM) often underperform “Source Only”. This pattern can be attributed to the nature of polyp segmentation, where targets are often small and concealed and, under shift, models tend to abstain, producing low-entropy yet incorrect predictions that yield misleading adaptation gradients and erode pre-trained knowledge. Additive parameter adaptation (ViDA, VPTTA) freezes the source model and attaches small trainable modules, which avoids harmful supervision and remains effective, though capacity is limited and per-

Table 3. Comparison with state-of-the-art CTTA methods on OD/OC under long-term CTTA using a ViT backbone (Segformer-B0).

Time	→					
Round	1	2	3	Average	Gain	Perform.
Methods	<i>DSC</i>	<i>DSC</i>	<i>DSC</i>	<i>DSC</i> ↑	<i>DSC</i> ↑	Degra. ↓
Source Only	68.98	68.98	68.98	68.98	-	-
TENT	69.42	66.54	62.40	66.12	-2.86	3.30
CoTTA	65.09	61.86	61.04	62.66	-6.32	2.43
SAR	69.89	69.26	69.50	69.55	0.57	0.34
ViDA	69.43	69.82	69.93	69.73	0.74	-0.30
PALM	70.54	70.47	70.31	70.44	1.46	0.10
SURGEON	69.74	69.86	70.06	69.89	0.91	-0.14
TACT	72.71	73.87	74.01	73.53	4.55	-0.82

Table 4. Results of ablation study on OD/OC segmentation.

$\mathcal{L}_{Stage1}$		$\mathcal{L}_{Stage2}$		Average
Style Perturbation	Structure Perturbation	$\mathcal{L}_e$	Reset	<i>DSC</i> ↑
-	-	-	-	65.92
✓	-	-	-	67.14
✓	✓	-	-	70.49
✓	✓	✓	-	<u>71.56</u>
✓	✓	-	✓	71.05
✓	✓	✓	✓	72.51

formance can be sensitive to initialization. Our TACT updates only a sparse set of inactive parameters to preserve pre-trained knowledge and applies resettable sample-wise refinement, thereby surpassing all competing methods and achieving superior performance across the three metrics.

Long-term CTTA. We evaluate long-term adaptation on OD/OC segmentation over three rounds. As shown in Fig. 3, TENT and CoTTA exhibit noticeable degradation, indicating error accumulation and forgetting across rounds, while more conservative or additive approaches (e.g., SAR, PALM, SURGEON, ViDA, VPTTA) reduce drift but yield smaller gains. By restricting updates to inactive parameters and applying resettable sample-wise refinement, TACT preserves pre-trained knowledge and remains stable under continual shifts, achieving the highest “Average Gain” and the lowest “Performance Degradation” among all methods.

Different backbones. We further evaluate OD/OC segmentation with a ViT backbone (SegFormer-B0). For ViT, we set the sparsity threshold to $\tau = 6 \times 10^{-4}$ to update $\approx 1\%$ of parameters (those with $|\theta_0| < \tau$), and we use a learning rate of 1×10^{-5} . All other hyperparameters follow the CNN setting. For the baselines, TENT and SAR update layer normalization (LN), and VPTTA is inapplicable because it relies on batch normalization (BN) statistics. As shown in Table 3, TACT achieves the largest “Average Gain” and the smallest (negative) “Performance Degradation”, indicating effective transfer to transformer backbones while maintaining stability.

Efficiency–accuracy trade-off. Figure 4 compares average *DSC* against peak GPU memory on OD/OC seg-

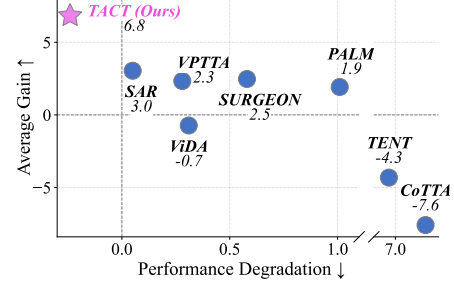


Figure 3. Our TACT achieves the highest gain with the lowest degradation on OD/OC segmentation under long-term CTTA. “Average Gain” is the mean *DSC* over three rounds minus “Source Only”. “Performance Degradation” is the first-round *DSC* minus the three-round mean.

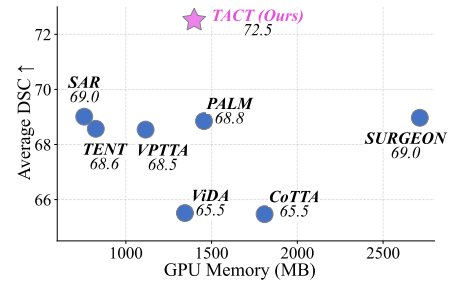


Figure 4. Efficiency–accuracy trade-off on OD/OC segmentation.

mentation. TACT attains the highest average *DSC* with a moderate memory footprint. SAR and TENT update only normalization layers, giving the lowest memory but lower accuracy. Full parameter methods (SURGEON and CoTTA) consume more memory. Prompt and adapter methods (VPTTA and ViDA) lie in between. Under the same protocol, TACT updates about 1.6% of parameters, supporting its balance between accuracy and memory.

4.4. Ablation Studies

We perform ablations on the OD/OC benchmark to assess the contribution of each component in TACT.

Effectiveness of each module. Table 4 reports ablations on OD/OC. The results reveal three main findings: (1) Stage 1 introduces style and structure perturbations and regularizes predictions under appearance and structure changes, yielding clear performance gains; (2) adding the hard-pixel contrastive loss in Stage 2 further enhances sample-specific adaptation, enabling the model to better handle challenging target samples; (3) the reset mechanism is critical for long-term stability because it restores parameters after sample-wise updates and prevents interference from these updates. Therefore, each module plays a vital role in achieving robust and adaptive continual adaptation, and the full model with all components performs best.

Sparsity threshold τ and scheduling. We study the

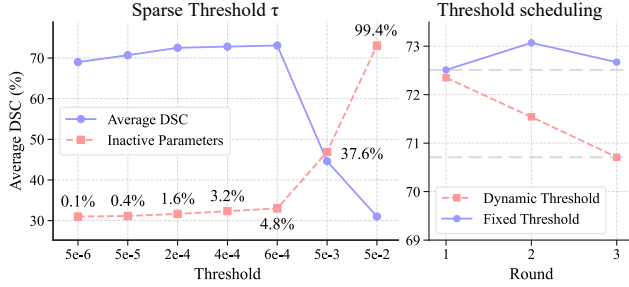


Figure 5. Effect of the sparsity threshold τ and its scheduling on OD/OC segmentation. Left: Sensitivity to τ . Numbers denote the trainable parameter ratio. Right: Long-term CTTA with a fixed τ compared with a dynamic schedule.

Table 5. Ablations with alternative choices on OD/OC segmentation. Results are averaged over all domains.

Ablation	Variant	DSC
Structure perturbation	Random mask	68.37
Structure perturbation	Ours	72.51
Style perturbation	Random style augmentation	71.67
Style perturbation	Ours	72.51
Style bank sampling	Farthest	70.53
Style bank sampling	Nearest (Ours)	72.51

effect of the inactivity threshold τ and its scheduling on OD/OC segmentation. In Fig. 5 (left), small thresholds update only a small subset of inactive parameters and keep DSC high, which indicates that restricting adaptation to inactive weights preserves pretrained knowledge. As τ increases, the trainable set expands and starts to include parameters that encode pretrained knowledge, which leads to a sharp drop in DSC . In Fig. 5 (right), we adjust τ during adaptation using prediction entropy. The dynamic schedule matches the fixed threshold in the first CTTA round, but in long-term CTTA it shows a pronounced and sustained decline because it is sensitive to the selection process and may update parameters that carry pretrained knowledge when the domain shift is large. A fixed and moderate threshold performs a single mask selection at the beginning, remains more stable across rounds, and achieves higher DSC .

Ablations with alternative choices. Table 5 compares TACT with alternative choices on OD/OC segmentation. Replacing entropy-guided structure masking with random masking lowers DSC , which shows the benefit of focusing perturbations on high-uncertainty regions. Swapping our FFT-based style perturbation for standard online style augmentation (e.g., brightness or contrast jitter, hue or saturation changes) is consistently weaker, suggesting that low-frequency amplitude exchange yields more realistic and effective appearance variation. For style bank sampling, using the nearest style outperforms the farthest, indicating that overly large low-frequency swaps can introduce artifacts and hinder adaptation.

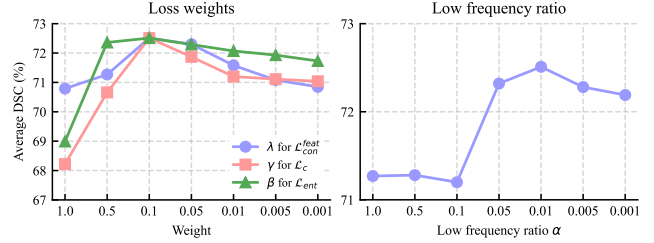


Figure 6. Sensitivity to hyperparameters on OD/OC segmentation.

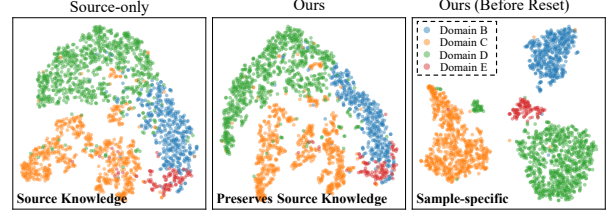


Figure 7. t-SNE visualization on OD/OC segmentation. Domain A as the source and all other domains as targets.

Hyperparameter sensitivity. We study two key hyperparameters on OD/OC segmentation (Fig. 6): the loss weights (λ for $\mathcal{L}_{con}^{feat}$, γ for $\mathcal{L}_c$, and β for $\mathcal{L}_{ent}$) and the low-frequency ratio α in the FFT-based style perturbation. Moderate loss weights produce a clear performance plateau, whereas very small or very large weights reduce DSC . Smaller low-frequency ratios work best, and larger ratios inject high-frequency content that hurts accuracy. We therefore fix a single default configuration for all experiments without per-dataset tuning.

t-SNE analysis. t-SNE on OD/OC segmentation (Fig. 7) shows that TACT closely matches the ‘‘Source Only’’ distribution, indicating that restricting updates to inactive parameters reduces interference with pre-trained knowledge. After per-sample refinement (before reset), target clusters become more separated, showing that the refinement captures sample-specific cues.

5. Conclusion

In this work, we propose TACT, a progressive two-stage continual test-time adaptation framework that restricts model updates to a sparse subspace of inactive parameters. Specifically, TACT employs style and structure perturbations to stabilize adaptation and introduces a hard-pixel contrastive loss for sample-specific refinement. A reset mechanism prevents interference from sample-specific updates, ensuring long-term stability. Extensive experiments on three multi-center medical image segmentation benchmarks demonstrate that TACT consistently outperforms state-of-the-art methods in both stability and adaptability across various network backbones and domain shifts.

Acknowledgement

This research was supported by National Natural Science Foundation of China (Grant No. 62571173), Zhejiang Provincial Natural Science Foundation of China (Grant Nos. LD25F020005 and LQN25F030009), and NSFC-FDCT under its Joint Scientific Research Project Fund (Grant No. 0004/2025/AFJ).

References

- [1] Reza Azad, Ehsan Khodapanah Aghdam, Amelie Rauland, Yiwei Jia, Atlas Haddadi Avval, Afshin Bozorgpour, Sanaz Karimijafarbigloo, Joseph Paul Cohen, Ehsan Adeli, and Dorit Merhof. Medical image segmentation review: The success of u-net. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):10076–10095, 2024. 1
- [2] Wenjia Bai, Ozan Oktay, Matthew Sinclair, Hideaki Suzuki, Martin Rajchl, Giacomo Tarroni, Ben Glocker, Andrew King, Paul M Matthews, and Daniel Rueckert. Semi-supervised learning for network-based cardiac mr image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 253–260. Springer, 2017. 1
- [3] Mathilde Bateson, Herve Lombaert, and Ismail Ben Ayed. Test-time adaptation with shape moments for image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 736–745. Springer, 2022. 1
- [4] Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, Debora Gil, Cristina Rodríguez, and Fernando Vilariño. Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized medical imaging and graphics*, 43:99–111, 2015. 5
- [5] B. Nicholas Bloch, Anant Madabhushi, Henkjan Huisman, John Freymann, Justin Kirby, Michael Grauer, Andinet Enquobahrie, Carl Jaffe, Larry Clarke, and Keyvan Farahani. Nci-isbi 2013 challenge: Automated segmentation of prostate structures (isbi-mr-prostate-2013), 2015. 5
- [6] Paola Cascante-Bonilla, Fuwen Tan, Yanjun Qi, and Vicente Ordonez. Curriculum labeling: Revisiting pseudo-labeling for semi-supervised learning. In *Proceedings of the AAAI conference on artificial intelligence*, pages 6912–6920, 2021. 1
- [7] Cheng Chen, Qi Dou, Hao Chen, Jing Qin, and Pheng Ann Heng. Unsupervised bidirectional cross-modality adaptation via deeply synergistic image and feature alignment for medical image segmentation. *IEEE transactions on medical imaging*, 39(7):2494–2505, 2020. 1
- [8] Lin Chen, Huaian Chen, Zhixiang Wei, Xin Jin, Xiao Tan, Yi Jin, and Enhong Chen. Reusing the task-specific classifier as a discriminator: Discriminator-free adversarial domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7181–7190, 2022. 1
- [9] Ziyang Chen, Yongsheng Pan, Yiwen Ye, Mengkang Lu, and Yong Xia. Each test image deserves a specific prompt: Continual test-time adaptation for 2d medical image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11184–11193, 2024. 1, 3, 5, 6
- [10] Sarkar Snigdha Sarathi Das, Ranran Haoran Zhang, Peng Shi, Wenpeng Yin, and Rui Zhang. Unified low-resource sequence labeling by sample-aware dynamic sparse finetuning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6998–7010, 2023. 3
- [11] Mario Döbler, Robert A Marsden, and Bin Yang. Robust mean teacher for continual and gradual test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7704–7714, 2023. 2
- [12] Francisco Fumero, Silvia Alayón, José L Sanchez, Jose Sigut, and M Gonzalez-Hernandez. Rim-one: An open retinal image database for optic nerve evaluation. In *2011 24th international symposium on computer-based medical systems (CBMS)*, pages 1–6. IEEE, 2011. 5
- [13] Mohsen Ghafoorian, Alireza Mehrtash, Tina Kapur, Nico Karssemeijer, Elena Marchiori, Mehran Pesteie, Charles RG Guttman, Frank-Erik De Leeuw, Clare M Tempny, Bram Van Ginneken, et al. Transfer learning for domain adaptation in mri: Application in brain lesion segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 516–524. Springer, 2017. 1
- [14] Hao Guan and Mingxia Liu. Domain adaptation for medical image analysis: a survey. *IEEE Transactions on Biomedical Engineering*, 69(3):1173–1185, 2021. 1
- [15] Zeyu Han, Chao Gao, Jinyang Liu, Jeff Zhang, and Sai Qian Zhang. Parameter-efficient fine-tuning for large models: A comprehensive survey. *arXiv preprint arXiv:2403.14608*, 2024. 3
- [16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5
- [17] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3, 2022. 3
- [18] Teng Hu, Jiangning Zhang, Ran Yi, Hongrui Huang, Yabiao Wang, and Lizhuang Ma. High-efficient diffusion model fine-tuning with progressive sparse low-rank adaptation. In *13th International Conference on Learning Representations, ICLR 2025*, pages 92066–92078. International Conference on Learning Representations, ICLR, 2025. 3
- [19] Zixuan Hu, Yichun Hu, Xiaotong Li, Shixiang Tang, and Ling-Yu Duan. Beyond entropy: Region confidence proxy for wild test-time adaptation. *arXiv preprint arXiv:2505.20704*, 2025. 1
- [20] Debesh Jha, Pia H Smedsrud, Michael A Riegler, Pål Halvorsen, Thomas De Lange, Dag Johansen, and Håvard D Johansen. Kvasir-seg: A segmented polyp dataset. In *International conference on multimedia modeling*, pages 451–462. Springer, 2019. 5

- [21] Christiaan Lamers, René Vidal, Nabil Belbachir, Niki Van Stein, Thomas Bäck, and Paris Giampouras. Clustering-based domain-incremental learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3384–3392, 2023. 1
- [22] Daeun Lee, Jaehong Yoon, and Sung Ju Hwang. Beccotta: Input-dependent online blending of experts for continual test-time adaptation. *arXiv preprint arXiv:2402.08712*, 2024. 3
- [23] Tao Lei, Junwen Bai, Siddhartha Brahma, Joshua Ainslie, Kenton Lee, Yanqi Zhou, Nan Du, Vincent Zhao, Yuxin Wu, Bo Li, et al. Conditional adapters: Parameter-efficient transfer learning with fast inference. *Advances in Neural Information Processing Systems*, 36:8152–8172, 2023. 3
- [24] Guillaume Lemaître, Robert Martí, Jordi Freixenet, Joan C Vilanova, Paul M Walker, and Fabrice Meriaudeau. Computer-aided detection and diagnosis for prostate cancer based on mono and multi-parametric mri: a review. *Computers in biology and medicine*, 60:8–31, 2015. 5
- [25] Dongze Lian, Daquan Zhou, Jiashi Feng, and Xinchao Wang. Scaling & shifting your features: A new baseline for efficient model tuning. *Advances in Neural Information Processing Systems*, 35:109–123, 2022. 3
- [26] Jian Liang, Ran He, Zhenan Sun, and Tieniu Tan. Exploring uncertainty in pseudo-label guided unsupervised domain adaptation. *Pattern Recognition*, 96:106996, 2019. 1
- [27] Baohao Liao, Yan Meng, and Christof Monz. Parameter-efficient fine-tuning without introducing new latency. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4242–4260, 2023. 3
- [28] Hyesu Lim, Byeonggeun Kim, Jaegul Choo, and Sungha Choi. Ttn: A domain-shift aware batch normalization in test-time adaptation. *arXiv preprint arXiv:2302.05155*, 2023. 2
- [29] Geert Litjens, Robert Toth, Wendy Van De Ven, Caroline Hoeks, Sjoerd Kerkstra, Bram Van Ginneken, Graham Vincent, Gwenael Guillard, Neil Birbeck, Jindang Zhang, et al. Evaluation of prostate segmentation algorithms for mri: the promise12 challenge. *Medical image analysis*, 18(2):359–373, 2014. 5
- [30] Jiaming Liu, Senqiao Yang, Peidong Jia, Renrui Zhang, Ming Lu, Yandong Guo, Wei Xue, and Shanghang Zhang. Vida: Homeostatic visual domain adapter for continual test time adaptation. *arXiv preprint arXiv:2306.04344*, 2023. 1, 3, 6
- [31] Quande Liu, Qi Dou, and Pheng-Ann Heng. Shape-aware meta-learning for generalizing prostate mri segmentation to unseen domains. In *International conference on medical image computing and computer-assisted intervention*, pages 475–485. Springer, 2020. 5
- [32] Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. Dora: Weight-decomposed low-rank adaptation. In *Forty-first International Conference on Machine Learning*, 2024. 3
- [33] Yang Liu, Wei Zhang, and Jun Wang. Source-free domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1215–1224, 2021. 1
- [34] Ke Ma, Jiaqi Tang, Bin Guo, Fan Dang, Sicong Liu, Zhui Zhu, Lei Wu, Cheng Fang, Ying-Cong Chen, Zhiwen Yu, et al. Surgeon: Memory-adaptive fully test-time adaptation via dynamic activation sparsity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 30514–30523, 2025. 1, 2, 6
- [35] Qinghe Ma, Jian Zhang, Zekun Li, Lei Qi, Qian Yu, and Yinghuan Shi. Steady progress beats stagnation: Mutual aid of foundation and conventional models in mixed domain semi-supervised medical image segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5175–5185, 2025. 5
- [36] Sarthak Kumar Maharana, Baoming Zhang, and Yunhui Guo. Palm: Pushing adaptive learning rate mechanisms for continual test-time adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 19378–19386, 2025. 1, 3, 6
- [37] M Jehanzeb Mirza, Jakub Micorek, Horst Possegger, and Horst Bischof. The norm must go on: Dynamic unsupervised domain adaptation by normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14765–14775, 2022. 2
- [38] Phan Ngoc Lan, Nguyen Sy An, Dao Viet Hang, Dao Van Long, Tran Quang Trung, Nguyen Thi Thuy, and Dinh Viet Sang. Neounet: Towards accurate colon polyp segmentation and neoplasm detection. In *International Symposium on Visual Computing*, pages 15–28. Springer, 2021. 5
- [39] A Tuan Nguyen, Thanh Nguyen-Tang, Ser-Nam Lim, and Philip HS Torr. Tipi: Test time adaptation with transformation invariance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24162–24171, 2023. 1, 2
- [40] Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Minghui Tan. Efficient test-time model adaptation without forgetting. In *International conference on machine learning*, pages 16888–16905. PMLR, 2022. 1, 2
- [41] Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Zhiqian Wen, Yaofo Chen, Peilin Zhao, and Minghui Tan. Towards stable test-time adaptation in dynamic wild world. *arXiv preprint arXiv:2302.12400*, 2023. 1, 3, 6
- [42] José Ignacio Orlando, Huazhu Fu, João Barbosa Breda, Karel Van Keer, Deepti R Bathula, Andrés Diaz-Pinto, Ruogu Fang, Pheng-Ann Heng, Jeyoung Kim, JoonHo Lee, et al. Refuge challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. *Medical image analysis*, 59:101570, 2020. 5
- [43] Hyejin Park, Jeongyeon Hwang, Sunung Mun, Sangdon Park, and Jungseul Ok. Medbn: Robust test-time adaptation against malicious test samples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5997–6007, 2024. 1
- [44] Junyoung Park, Jin Kim, Hyeonjun Kwon, Ilhoon Yoon, and Kwanghoon Sohn. Layer-wise auto-weighting for

- non-stationary test-time adaptation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1414–1423, 2024. [3](#)
- [45] Swami Sankaranarayanan, Yogesh Balaji, Arpit Jain, Ser Nam Lim, and Rama Chellappa. Learning from synthetic data: Addressing domain shift for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3752–3761, 2018. [1](#)
- [46] Neeraj Sharma and Lalit M Aggarwal. Automated medical image segmentation techniques. *Journal of medical physics*, 35(1):3–14, 2010. [1](#)
- [47] Shufan Shen, Junshu Sun, Xiangyang Ji, Qingming Huang, and Shuhui Wang. Expanding sparse tuning for low memory usage. *Advances in Neural Information Processing Systems*, 37:76616–76642, 2024. [3](#)
- [48] Juan Silva, Aymeric Histace, Olivier Romain, Xavier Dray, and Bertrand Granado. Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer. *International journal of computer assisted radiology and surgery*, 9(2):283–293, 2014. [5](#)
- [49] Samarth Sinha, Peter Gehler, Francesco Locatello, and Bernt Schiele. Test: Test-time self-training under distribution shift. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2759–2769, 2023. [2](#)
- [50] Jayanthi Sivaswamy, SR Krishnadas, Gopal Datt Joshi, Madhulika Jain, and A Ujjwaft Syed Tabish. Drishti-gs: Retinal image dataset for optic nerve head (onh) segmentation. In *2014 IEEE 11th international symposium on biomedical imaging (ISBI)*, pages 53–56. IEEE, 2014. [5](#)
- [51] Junha Song, Jungsoo Lee, In So Kweon, and Sungha Choi. Ecotta: Memory-efficient continual test-time adaptation via self-distilled regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11920–11929, 2023. [3](#)
- [52] Yusheng Su, Xiaozhi Wang, Yujia Qin, Chi-Min Chan, Yankai Lin, Huadong Wang, Kaiyue Wen, Zhiyuan Liu, Peng Li, Juanzi Li, et al. On transferability of prompt tuning for natural language processing. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3949–3969, 2022. [3](#)
- [53] Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *International conference on machine learning*, pages 9229–9248. PMLR, 2020. [2](#)
- [54] Yi-Lin Sung, Varun Nair, and Colin A Raffel. Training neural networks with fixed sparse masks. *Advances in Neural Information Processing Systems*, 34:24193–24205, 2021. [3](#)
- [55] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*, 2020. [1](#), [2](#), [3](#), [6](#)
- [56] Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7201–7211, 2022. [1](#), [2](#), [6](#)
- [57] Ran Wang, Hua Zuo, Zhen Fang, and Jie Lu. Multiple teacher model for continual test-time domain adaptation. In *Australasian Joint Conference on Artificial Intelligence*, pages 304–314. Springer, 2023. [1](#)
- [58] Shujun Wang, Lequan Yu, Kang Li, Xin Yang, Chi-Wing Fu, and Pheng-Ann Heng. Boundary and entropy-driven adversarial learning for fundus image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 102–110. Springer, 2019. [1](#)
- [59] Yanchao Yang and Stefano Soatto. Fda: Fourier domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4085–4095, 2020. [1](#)
- [60] Chenyu You, Weicheng Dai, Yifei Min, Fenglin Liu, David Clifton, S Kevin Zhou, Lawrence Staib, and James Duncan. Rethinking semi-supervised medical image segmentation: A variance-reduction perspective. *Advances in neural information processing systems*, 36:9984–10021, 2023. [1](#)
- [61] Guoning Zhang, Xiaoran Qi, Bo Yan, and Guotai Wang. Iplc: iterative pseudo label correction guided by sam for source-free domain adaptation in medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 351–360. Springer, 2024. [1](#)
- [62] Jian Zhang, Lei Qi, Yinghuan Shi, and Yang Gao. Domainadaptor: A novel approach to test-time adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18971–18981, 2023. [1](#), [2](#)
- [63] Marvin Zhang, Sergey Levine, and Chelsea Finn. Memo: Test time robustness via adaptation and augmentation. *Advances in neural information processing systems*, 35:38629–38642, 2022. [1](#), [2](#)
- [64] Zhuo Zhang, Feng Shou Yin, Jiang Liu, Wing Kee Wong, Ngan Meng Tan, Beng Hai Lee, Jun Cheng, and Tien Yin Wong. Origa-light: An online retinal fundus image database for glaucoma analysis and research. In *2010 Annual international conference of the IEEE engineering in medicine and biology*, pages 3065–3068. IEEE, 2010. [5](#)