Breaking the Single-Stage Barrier: Synergistic Data-Model Adaptation at Test-Time for Medical Image Segmentation

Wenjuan Zhou[1], Wei Chen[1†], Yulin He[1], Di Wu[1], and Chen Li[1]

[1]College of Computer Science and Technology, National University of Defense Technology, Changsha, China

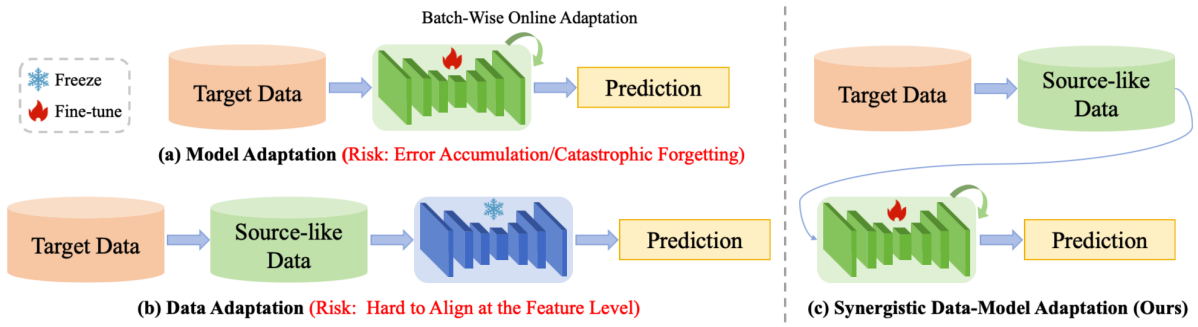

Figure 1: ***Comparison of different test-time adaptation (TTA) frameworks.*** *(a) Model adaptation adjusts the source model weights to fit the target data, but it is prone to error accumulation and catastrophic forgetting when the target domain shifts continuously. (b) Data adaptation freezes the source model and aligns distributions by transforming target data, yet achieves suboptimal feature-level matching. (c) We propose Synergistic Data-Model Adaptation (SDMA) that concurrently optimizes both model and data for superior domain alignment.*

Abstract

Domain shift, predominantly caused by variations in medical imaging across different institutions, often leads to a decline in the accuracy of medical image segmentation models. While Test-Time Adaptation (TTA) holds promise to address this issue, existing methods exhibit significant limitations: model adaptation is prone to error accumulation and catastrophic forgetting in continuous domain learning. Meanwhile, data adaptation struggles to achieve deep latent alignment due to the inaccessibility of source domain data. To address these challenges, we propose Synergistic Data-Model Adaptation (SDMA), which innovatively leverages Batch Normalization (BN) layers as a bidirectional bridge to enable a two-stage joint adaptation process. In the data adaptation stage, domain-aware prompts dynamically adjust the BN statistics of incoming test data, achieving low-level distribution alignment in the Fourier space. In the model adaptation stage, we dynamically optimize the BN affine parameters based on strong-weak data augmentation and entropy minimization, enabling adaptation to high-level semantic features. Experiments conducted on five retinal fundus image datasets from various medical institutions demonstrate that our method achieves an average Dice improvement of 1.23% over previous state-of-the-art (SOTA) methods, establishing a new SOTA performance.

CCS Concepts

• ***Computing methodologies*** → ***Image segmentation;***

1. Introduction

Medical image segmentation aims to automatically localize and recognize lesion areas, thereby improving diagnostic accuracy and efficiency while reducing human error [LTC*20]. With the success of deep learning, medical image segmentation has seen significant improvements in accuracy. However, deep networks are typically sensitive to domain shifts [HMW*23], which are common in medical imaging scenarios. Medical imaging exhibits intrinsic domain variability due to discrepancies in hardware scanner, protocol-specific parameter configurations, and temporal changes

† Corresponding author: Wei Chen, chenwei@nudt.edu.cn

DOI: 10.2312/pg.20251273

in signal-to-noise ratio. These variations can cause domain shift that degrade the accuracy of segmentation models in real-world applications. To address the domain shift problem, Unsupervised Domain Adaptation (UDA) [BGR*06, ZPIE17, HTP*18, SLL*21, GWD*23, ZWA*24, ZLSZ24] has been proposed, which leverages labeled source data and unlabeled target data during training with the goal of improving prediction accuracy on the target domain. However, in clinical practice, timely diagnosis and treatment often require immediate predictions on individual test samples or small batches, making it impractical to wait for a large volume of test images from the same domain for UDA. Besides, due to medical data privacy concerns, accessing source domain training data during testing is not always feasible. Therefore, Test-Time Adaptation (TTA) [WSL*20, WFVGD22, YCJ*22, LFW*24, CPY*24, BY24, CYPX25] emerges as a more promising solution in such scenarios.

Test-time adaptation relies solely on test data, adapting the model to incoming samples to improve performance. In real clinical environments, the target domain undergoes continuous changes. To address this, we focus on the continual test-time adaptation (CTTA) setup, which extends TTA to a sequence of distribution shifts. A common solution is model adaptation, as illustrated in Fig. 1(a), which continuously updates the weights of source model using the streaming target data. However, due to the absence of supervision, most TTA methods adopt self-supervised strategies, such as entropy minimization [WSL*20] and pseudo-labeling [LFW*24]. These methods are subject to noisy supervision, which can lead to error accumulation over time. Moreover, continuously training on new domains can cause catastrophic forgetting, resulting in notable accuracy degradation. To mitigate these issues, data adaptation strategies have recently emerged. Instead of updating model weights, data adaptation (Fig. 1(b)) transforms the target data distribution to resemble that of the source domain. For example, VPTTA [CPY*24] freezes the pre-trained model and learns low-frequency prompts for each image to conduct data adaptation. Basak et al. [BY24] propose a generative latent search paradigm to reconstruct the closest clone of each target image from the source latent space. Since the model remains frozen, it avoids problems like error accumulation and catastrophic forgetting. However, without access to source domain data, aligning the adapted target data with the source model becomes challenging. Despite apparent visual similarities, deep networks may still interpret them quite differently. Moreover, the target data itself contains valuable cues that can be exploited to further improve the model. A natural question arises: **can we combine the strengths of data adaptation and model adaptation**?

Despite the clear motivation, the above question has been hardly explored in prior works, to the best of our knowledge. A major challenge is that data and model naturally serve varying roles in deep learning, making it hard to jointly optimize data and model adaptation. Nevertheless, we found that Batch Normalization (BN) layer is a subtle intermediate variable to bridge data and model adaptation. The BN layer consists of both statistical parameters (mean μ and variance σ) and affine parameters (scale factor γ and shift factor β). The normalization step first standardizes the input x using μ and σ to obtain $\bar{x} = (x-\mu)/\sigma$, then transforms it through affine parameters to produce the final output $x' = \gamma\bar{x} + \beta$. In which, the statistical parameters are non-trainable and capture the data distribution, while the affine parameters are trainable and enhance the model's adaptability to different domains. When domain shift occurs, discrepancies in statistical parameters between the target and source domains lead to a decline in model accuracy [WSL*20, LXY*21, NWZ*23].

To address this issue, we propose Synergistic Data-Model Adaptation (SDMA) (Fig. 1(c)) that aligns the BN layer's outputs between target and source domain data. Data adaptation modifies the mean and variance of statistical parameters to match low-level characteristics (style/texture), while model adaptation fine-tunes the affine parameters via gradient updates to handle high-level semantic shifts. These two strategies work synergistically to mitigate the impact of output mismatches between source and target domains in the BN layer. In the data adaptation stage, we optimize domain-aware prompts to adjust the low-frequency Fourier components of test images (encoding style/texture), aligning them with the source distribution. The key adaptation signal comes from minimizing BN statistics divergence (μ/σ) between source and prompted target features. In the model adaptation stage, we first leverage multi-head predictions on weakly augmented target samples to generate reliable pseudo-labels, which are then used to supervise the predictions of strongly augmented target images. In addition, we apply entropy minimization as a regularization strategy to update the affine parameters of the BN layers. After updating both the statistical parameters and affine parameters of the BN layers, we effectively align the data and latent feature distributions between source and target domains.

Our main contributions are as follows:

- We introduce a two-stage TTA framework that synergistically integrates data-level distribution alignment with model-level feature calibration, establishing a new paradigm beyond single-stage approaches.
- We incorporate a consistency loss with strong–weak data augmentations and investigate different update strategies to better align data adaptation with model adaptation. These carefully designed enhancements enable a synergistic effect.
- We reveal the bidirectional bridging role of Batch Normalization layers as distribution shift detectors (via μ and σ) for data adaptation and feature calibrators (via γ and β) for model adaptation, achieving error reduction through co-adaptation.
- Extensive validation on five clinical retinal fundus datasets shows our method achieves 1.23% average Dice gain over SOTA, proving the superiority of employing two-stage synergistic data-model adaptation.

2. Related Work

2.1. Unsupervised Domain Adaptation

Unsupervised Domain Adaptation (UDA) has been extensively explored as a means to address performance degradation caused by domain shifts between training (source) and testing (target) domains. Existing UDA approaches can generally be classified into three main categories. The first category leverages feature statistics alignment [BGR*06, SLL*21], aiming to reduce domain discrepancies by matching the statistical characteristics of feature

distributions. The second category is based on adversarial learning [ZPIE17, HTP*18, ZLSZ24], where domain discriminators are employed to extract domain-invariant features through adversarial optimization. The third category relies on self-training strategies [ZWA*24, GWD*23], which iteratively refine the model using pseudo-labels generated from its own predictions on unlabeled target data.

While effective, these methods usually require source data and abundant target samples, making them impractical when source data is unavailable and target data is scarce or arrives incrementally. This has motivated growing interest in test-time adaptation, which adapts models at deployment without source data. Building on the test-time adaptation paradigm, our method entirely eliminates the need for source data.

2.2. Test-Time Adaptation

Test-Time Adaptation (TTA) aims to adapt a source-pre-trained model to target domain data during inference in a source-free and online manner. In practice, we focus on the continual test-time adaptation (CTTA) setup. Most existing methods focus on model adaptation, updating the model's parameters to fit the test data. For example, TENT [WSL*20] minimizes the entropy of model predictions by updating the trainable parameters in Batch Normalization (BN) layers. CoTTA [WFVGD22] introduces a teacher-student framework, where the student model is adapted through weight-averaged parameters and augmentation-averaged predictions. DLTTA [YCJ*22] proposes a dynamic learning rate strategy to achieve efficient and stable adaptation.

However, the absence of labels and the continuously changing distribution of the target domain can lead to error accumulation and catastrophic forgetting in such model-centric methods. To overcome these issues, recent works have begun to explore data adaptation, which aims to transform test-time inputs to better match a fixed, source-trained model, thus avoiding direct parameter updates. For instance, VPTTA [CPY*24] freezes the pre-trained model and introduces low-frequency prompts generated per image to adjust the input in medical image segmentation. Similarly, Basak et al. [BY24] propose a variational sampling strategy in the source representation space to retrieve the closest "clone" of a target image and reconstruct it using the latent distribution of the source domain. By keeping the model fixed, these data-centric methods mitigate error propagation and forgetting. However, in the absence of source data, the transformed target inputs often fail to fully align with the source-trained model.

To address these challenges, we propose a Synergistic Data-Model Adaptation (SDMA) strategy that combines the strengths of both paradigms. Our method jointly aligns low-level image characteristics and high-level semantic features.

2.3. Batch Normalization-Based Test-Time Adaptation

Many existing TTA methods leverage the statistical and affine parameters of Batch Normalization (BN) layers to facilitate domain adaptation. For example, Adaptive Batch Normalization (AdaBN) [LWS*16] recomputes BN statistics on test data to enhance the generalization capability of deep neural networks. Liu et al. [SLL*21] categorized BN parameters into low-order components (mean and variance) and high-order components (scale and shift). During inference, they updated only the low-order statistics using test data while keeping high-order parameters fixed, achieving improved performance in medical image segmentation. SoTTA [GKL*23] introduced an exponential moving average strategy to update BN statistics dynamically. Similarly, several studies [SRE*20, NPS*20, ZMD*21] have shown that recomputing BN statistics with test data can effectively alleviate the domain shift between source and target. In addition, some works [GKL*23, NWZ*23] have explored updating only the affine parameters of BN layers via backpropagation, demonstrating that tuning these limited parameters can yield adaptation performance comparable to full model fine-tuning, while offering better stability and efficiency.

In contrast to prior approaches, our method leverages BN layers as a bidirectional bridge between data adaptation and model adaptation. Through a two-stage adaptation framework, we jointly utilize BN statistics and affine parameters, enabling more effective and robust domain adaptation during inference.

3. Methodology

The pipeline of our proposed method is illustrated in Fig. 2. In this section, we first present the problem formulation, and then detail the data adaptation and model adaptation stage.

3.1. Problem Formulation

We consider a continual test-time adaptation setting. The source domain dataset is denoted as $\mathcal{D}^S = \{(x_i^S, y_i^S)\}_{i=1}^{N^S}$, consisting of N^S labeled medical images, where $x_i^S \in \mathcal{X}^S$ are the input images and $y_i^S \in \mathcal{Y}^S$ are the corresponding labels. A pre-trained model F_{θ_0} is obtained by training on $\mathcal{D}^S$. Our objective is to continuously adapt this model to a dynamically changing target domain in an online manner, without access to source domain data during inference.

The unlabeled target data $\mathcal{X}^T$ arrives sequentially as a time series, with the model only accessing data at the current timestep. We consider an extreme scenario where only a single image is available at each timestep. At time t, the model receives an input x_t^T and must predict its corresponding label $\hat{y} \in \mathcal{Y}^T$. While the source and target domains differ in distribution ($\mathcal{X}^S \neq \mathcal{X}^T$), they share the same label space ($\mathcal{Y}^S = \mathcal{Y}^T$). Notably, the distribution of x_t^T may shift over time, and the model is evaluated solely based on its online prediction performance.

3.2. Data Adaptation

Directly applying target domain test data to a model trained on the source domain often results in significant performance degradation due to domain shift. To alleviate this, we adapt the test data to better match the source domain distribution before inference. Inspired by previous work [GBL*23, CPY*24], we employ domain-aware prompts to adjust the appearance of test images, encouraging similarity to the source domain in style and texture.

As shown in Fig. 2, we first apply the Fast Fourier Transform (FFT) to the test image to decompose it into amplitude and

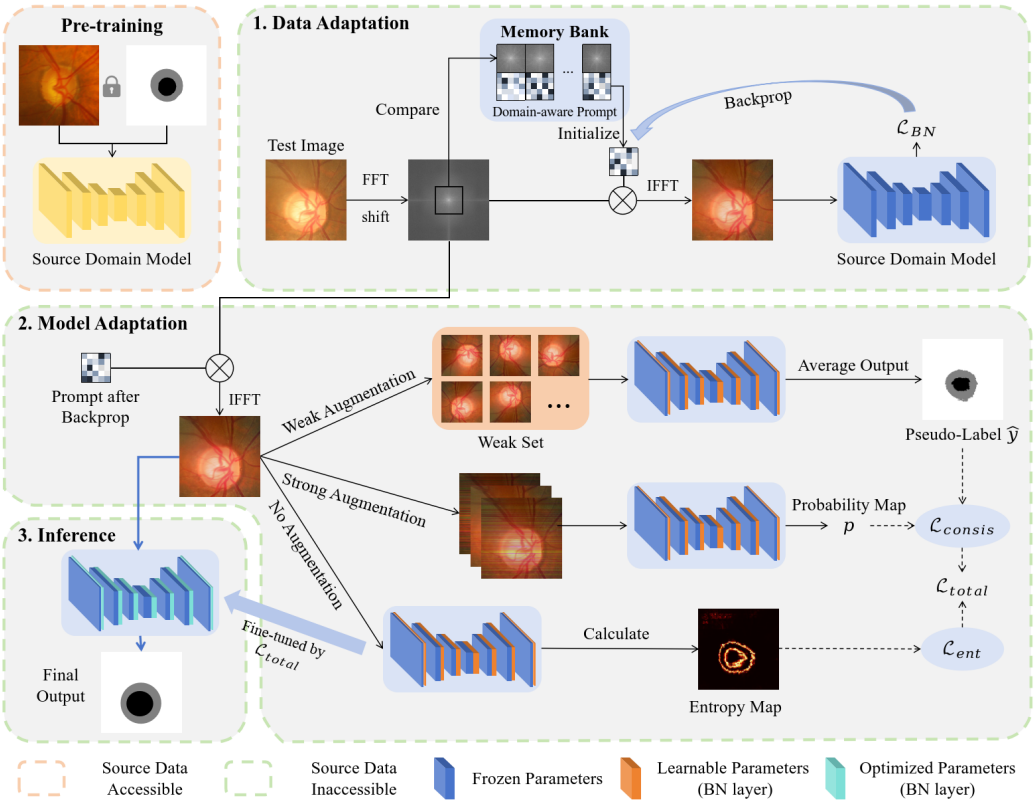

Figure 2: ***Overview of our proposed method.*** *After obtaining the source domain model during the pretraining stage, each test image undergoes two adaptation stages. In the data adaptation stage, the test image is transformed toward a source-like distribution using Fourier transformation and domain-aware prompt optimization. A memory bank is utilized to effectively initialize these prompts, while the absolute distance of BN statistical parameters guides the prompt optimization process. In the model adaptation stage, the affine parameters of the Batch Normalization (BN) layers are optimized through consistency learning between strong and weak data augmentations, along with entropy minimization. During inference, the adapted test image is passed through the adapted model to generate the final segmentation result. By jointly calibrating BN statistics and affine parameters, the process achieves synergistic adaptation.*

phase components. The frequency spectrum is centered so that low-frequency components, which are closely related to image style and texture [CPY*24], are positioned at the center to enable effective frequency-domain manipulation. A learnable domain-aware prompt is then used to modify the low-frequency region of the amplitude spectrum, adjusting the image's global appearance while preserving its semantic content.

Formally, for a test image $x_t^T \in \mathbb{R}^{H\times W\times C}$ at timestep t, let $\mathcal{F}^A(x_t^T)$ and $\mathcal{F}^P(x_t^T)$ denote its amplitude and phase components, respectively. Given a prompt $\mathcal{P}_t \in \mathbb{R}^{(\alpha H)\times(\alpha W)\times C}$, the adapted image $\tilde{x}_t^T$ is computed as:

$$\tilde{x}_t^T = \mathcal{F}^{-1}\left(\left[\text{OnePadding}(\mathcal{P}_t)\odot\mathcal{F}^A(x_t^T),\mathcal{F}^P(x_t^T)\right]\right), \quad (1)$$

where $\mathcal{F}^{-1}$ is the inverse FFT, $\odot$ denotes element-wise multiplication, and OnePadding one-pads the prompt $\mathcal{P}_t$ to match the spatial resolution $H\times W$. The scaling factor $\alpha\in(0,1)$ controls the region of the amplitude spectrum modified by the prompt, encouraging it to focus on low-frequency components.

Prompt Initialization. To enable stable and effective prompt learning, we introduce a memory bank of length M for prompt initialization. This memory bank stores pairs of low-frequency amplitude components and their corresponding prompts, denoted as $\{k_m, v_m\}_{m=1}^M$, where k_m represents a stored low-frequency component and v_m is its corresponding prompt. All entries are initially set to ones and updated using a First-In-First-Out (FIFO) strategy: new entries are enqueued, and the oldest are removed.

The memory bank effectively captures the dynamically changing low-frequency styles encountered over time. To retrieve a relevant initialization for the current test image x_t^T, we compute the cosine similarity between its low-frequency amplitude component $\mathcal{F}^A_{\text{low}}(x_t^T)$ and each stored component k_m in the memory:

$$\text{Cos}(\mathcal{F}^A_{\text{low}}(x_t^T),k_m) = \frac{\langle\mathcal{F}^A_{\text{low}}(x_t^T),k_m\rangle}{\|\mathcal{F}^A_{\text{low}}(x_t^T)\|\cdot\|k_m\|}, \quad (2)$$

where $\langle\cdot,\cdot\rangle$ denotes the dot product, and $\|\cdot\|$ is the Euclidean norm.

We sort the similarity scores and retrieve the top-N nearest

neighbors to construct a support set $R_t = \{k_n, v_n\}_{n=1}^N$ for the current image. Each corresponding prompt v_n is assigned a weight w_n based on its similarity score, such that prompts corresponding to more similar low-frequency components receive higher weights. The weights are normalized to ensure $\sum_{n=1}^N w_n = 1$. The initialized prompt $\mathcal{P}_t$ is then computed as a weighted combination:

$$\mathcal{P}_t = \sum_{n=1}^{N} w_n v_n. \tag{3}$$

Statistical Alignment. Batch Normalization (BN) statistics, specifically the mean μ and variance σ, capture domain-specific characteristics. A major cause of performance degradation under domain shift is the mismatch between BN statistics from the source and target domains [NWZ*23]. To mitigate this, we use BN statistics to guide low-level distribution alignment of the target data.

After passing the adapted image $\tilde{x}_t^T$ through the source model F_{θ_0}, we update the prompt by minimizing the discrepancy in BN statistics. The loss function $\mathcal{L}_{BN}$ is defined as the average absolute difference between source and target BN statistics:

$$\mathcal{L}_{BN} = \frac{1}{J}\sum_{j=1}^{J}\left(|\mu_s^j - \mu_t^j| + |\sigma_s^j - \sigma_t^j|\right), \tag{4}$$

where J is the total number of BN layers in F_{θ_0}, and $\mu_s^j, \sigma_s^j, \mu_t^j, \sigma_t^j$ are the mean and variance from the j-th BN layer under source and target inputs.

During backpropagation, only the prompt $\mathcal{P}_t$ is updated to minimize $\mathcal{L}_{BN}$, while the parameters of F_{θ_0} remain frozen. The updated prompt is denoted as $\mathcal{P}_t'$. This enables lightweight and stable test-time adaptation by aligning the low-level style and texture of target data with the source domain.

3.3. Model Adaptation

During the model adaptation stage, we dynamically update the affine parameters (γ and β) of the Batch Normalization (BN) layers to better align deep semantic features. Although the domain-aware prompt narrows the visual gap between target and source data, discrepancies remain in their deep feature representations. To bridge this, we refine BN's affine parameters, adjusting γ for precise feature scaling and β for translation, aligning the target BN output $\gamma_t(x_t - \mu_t)/\sigma_t + \beta_t$, with the source BN output $\gamma_s(x_s - \mu_s)/\sigma_s + \beta_s$, thus achieving effective semantic alignment in feature space.

As illustrated in Fig. 2, at each timestep t, we use the prompt $\mathcal{P}_t'$ from the data adaptation stage to generate the adapted image $\hat{x}_t^T$:

$$\hat{x}_t^T = \mathcal{F}^{-1}\left(\left[\text{OnePadding}(\mathcal{P}_t') \odot \mathcal{F}^A(x_t^T), \mathcal{F}^P(x_t^T)\right]\right), \tag{5}$$

We apply multiple weak augmentations to this image, including rotation, horizontal flipping, and vertical flipping. These weakly augmented samples are fed into the model from the previous timestep, $F_{\theta_{t-1}}$, to obtain pseudo-labels $\hat{y}$ via a multi-head ensemble of predictions. This ensemble strategy leverages the fact that weight-averaged models tend to produce more reliable predictions, and that averaging over multiple augmentations can further enhance pseudo-label quality [WFVGD22].

Next, we apply strong augmentations to $\hat{x}_t^T$ and input the result into $F_{\theta_{t-1}}$ to obtain the predicted probability map p_t. The pseudo-label $\hat{y}$ from the weakly augmented samples is then used to supervise the prediction from the strongly augmented image, leading to the consistency loss $\mathcal{L}_{\text{consis}}$:

$$\mathcal{L}_{\text{consis}} = -\sum \hat{y} \log p_t. \tag{6}$$

In addition to consistency training, we also perform entropy minimization on the prediction $\tilde{y}$ from $F_{\theta_{t-1}}$ when fed with the adapted image $\hat{x}_t^T$. The entropy loss $\mathcal{L}_{\text{ent}}$ is defined as:

$$\mathcal{L}_{\text{ent}} = -\sum \tilde{y} \log \tilde{y}. \tag{7}$$

The final loss combines both terms as a weighted sum:

$$\mathcal{L}_{\text{total}} = \rho \mathcal{L}_{\text{consis}} + (1-\rho)\mathcal{L}_{\text{ent}}, \tag{8}$$

where $\rho \in (0,1)$ controls the trade-off between consistency and entropy minimization.

We apply $\mathcal{L}_{\text{total}}$ to update only the BN affine parameters in $F_{\theta_{t-1}}$, yielding the adapted model F_{θ_t} after just one iteration. By enforcing consistency between strongly and weakly augmented samples and minimizing the entropy of predictions, our method encourages the model to preserve augmentation-invariant features while enhancing prediction confidence.

3.4. Inference

After completing the data and model adaptation stages at each timestep t, the adapted image $\hat{x}_t^T$ is fed into the adapted model F_{θ_t} to produce the final segmentation output y_t:

$$y_t = F_{\theta_t}(\hat{x}_t^T). \tag{9}$$

Here, F_{θ_t} represents the model whose Batch Normalization affine parameters have been fine-tuned online using the latest target domain data, allowing it to better capture the changing distribution features. Consequently, the segmentation result y_t is expected to be more accurate and reliable compared to directly using the pre-trained source model without adaptation. This continuous adaptation enables the model to maintain robust performance despite ongoing domain shifts.

4. Experiments

4.1. Experimental Setup

Datasets and Evaluation Metrics. We evaluate our method on five publicly available retinal fundus image datasets, all focusing on

Table 1: *Overview of five retinal fundus image datasets from diverse medical centers for cross-domain experiments.*

Domain ID	Dataset	Sample size
Domain A	RIM-ONE-r3 [FAS*11]	159
Domain B	REFUGE [OFB*20]	400
Domain C	ORIGA [ZYL*10]	650
Domain D	REFUGE-Validation/Test [OFB*20]	800
Domain E	Drishti-GS [SKJ*14]	101

Table 2: *Quantitative evaluation results of the segmentation on the retinal fundus datasets. The ↑ sign indicates a higher score is better. The best results are in boldface.*

Method	Domain A	Domain B	Domain C	Domain D	Domain E	Average
	DSC	DSC	DSC	DSC	DSC	DSC↑
Source Only	64.53	76.06	71.18	52.67	64.87	65.86
TENT-Continual [WSL*20]	73.07	78.66	71.94	46.81	70.20	68.13
CoTTA [WFVGD22]	75.39	75.98	69.14	53.99	70.40	68.98
DLTTA [YCJ*22]	75.11	78.85	73.89	51.64	69.71	69.84
DUA [MMPB22]	72.28	76.59	70.13	56.17	71.38	69.31
SAR [NWZ*23]	74.55	77.71	70.78	55.40	71.72	70.03
DomainAdaptor [ZQSG23]	74.50	76.39	71.81	56.78	70.55	70.01
VPTTA [CPY*24]	73.91	**79.36**	**74.51**	56.51	**75.35**	71.93
SDMA (Ours)	**76.07**	78.87	71.18	**71.78**	67.91	**73.16**

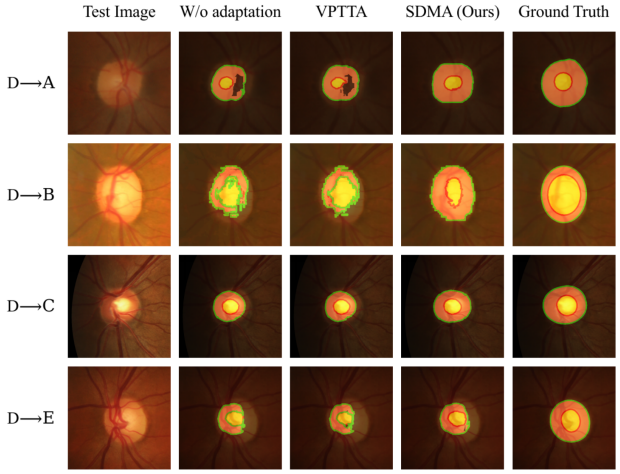

Figure 3: *Qualitative comparison of segmentation results on retinal fundus images among the W/o Adaptation baseline, VPTTA, and our proposed method. "D → A" indicates that the model is trained on source domain D and evaluated on target domain A. The optic cup (OC) and optic disc (OD) are displayed in red and green contours respectively.*

the optic cup (OC) and optic disc (OD) segmentation task. These datasets, collected from different medical centers, are designated to as domain A to domain E, which are described in Table 1.

For each image, we crop a region of interest (ROI) centered on the OD with a fixed size of 800×800 pixels, following the protocol in [HLX22]. Each ROI is then resized to 512×512 pixels and normalized using min-max normalization. For evaluation, we adopt the Dice Similarity Coefficient (DSC) as the metric, consistent with prior studies [CPY*24].

Implementation Details. We follow a cross-domain evaluation protocol in each experiment: the source model is trained on one domain and tested on the remaining target domains. We report the average performance across all target domains to evaluate generalization under diverse domain shifts.

We use ResUNet-34 [HZRS16] as the segmentation backbone. During test-time adaptation, all methods including ours and the baselines perform one adaptation iteration per incoming test sample (batch size = 1) to ensure fair comparison. In the data adaptation stage, we use the Adam optimizer [Kin14] with a learning rate of 0.05. The prompt size ratio α is set to 0.01. The memory bank size M is 40, and the support set size N is 16, following [CPY*24]. For model adaptation, we again use the Adam optimizer with a learning rate of 1×10^{-5}. The balancing factor ρ in the total loss is set to 0.5. Batch Normalization statistics are updated online using the incoming test data. Each adaptation step consists of one cycle combining data-level transformation and model-level parameter refinement.

All experiments are implemented in PyTorch and conducted on a single NVIDIA Quadro RTX 6000 GPU with 24 GB memory. The strong data augmentation strategies include brightness adjustment, contrast variation, Gamma transformation, Gaussian noise, and Gaussian blur.

4.2. Comparison with State-of-the-Arts

We begin by outlining the comparative methods. As shown in Table 2, "Source Only" serves as the lower bound, where the model is trained solely on the source domain and evaluated on the target domain without any adaptation. Our proposed method is compared against a range of continual test-time adaptation (CTTA) methods, including both model adaptation and data adaptation methods. Among model adaptation methods, CoTTA [WFVGD22] is based on pseudo-labeling, while TENT-continual [WSL*20] and SAR [NWZ*23] employ entropy minimization to guide adaptation. DLTTA [YCJ*22] introduces dynamic learning rate adjustment, DomainAdaptor [ZQSG23] combines entropy loss with Batch Normalization (BN) statistics fusion, and DUA [MMPB22] focuses on modifying BN statistics. In terms of data adaptation, VPTTA [CPY*24] adjusts input data distributions via visual prompts. The performance results of these methods are taken from [CPY*24] for consistency and fair comparison.

Our proposed approach achieves state-of-the-art performance on the optic cup and optic disc segmentation task, with an average Dice Similarity Coefficient of 73.16%, outperforming all baseline methods. Notably, on Domain D, our method achieves a DSC of 71.78%, surpassing the second-best method, DomainAdaptor, which scores 56.78%, by more than 15 percentage points.

Table 3: *Ablation Study on different components of our method.*

Components		Domain A	Domain B	Domain C	Domain D	Domain E	Average
Data Adaptation	Model Adaptation	DSC	DSC	DSC	DSC	DSC	DSC↑
		64.53	76.06	71.18	52.67	64.87	65.86
✓		73.90	79.35	74.51	56.53	75.35	71.93
	✓	74.49	77.47	70.73	58.75	73.14	70.92
✓	✓	76.07	78.87	71.18	71.78	67.91	73.16

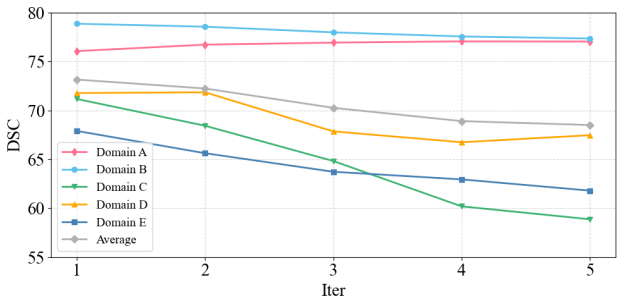

Figure 4: *The change of DSC score on OC/OD segmentation with different training iterations in the model adaptation stage.*

This highlights the strong cross-domain generalization capability of our method. While VPTTA performs well on Domain B with 79.36%, Domain C with 74.51%, and Domain E with 75.35%, our method demonstrates more balanced performance across all domains. Specifically, SDMA achieves the highest DSC on Domain A with 76.06%, outperforming VPTTA's 73.91%. On Domain B, our result of 78.87% is nearly on par with VPTTA, with only a 0.49% difference. Most notably, on Domain D, our method achieves 71.63%, substantially outperforming VPTTA's 56.51%, demonstrating superior robustness.

This consistent performance advantage across domains suggests that our method effectively mitigates domain shifts, especially in challenging scenarios such as Domain D, where other methods experience severe degradation exceeding 10%. The strength of our approach lies in its synergistic integration of data and model adaptation, enabling more comprehensive domain adaptation.

Finally, Fig. 3 provides a qualitative comparison of segmentation results on Domain D. The W/o Adaptation baseline exhibits significant prediction bias, including structural distortion and poor boundary delineation. While VPTTA shows moderate improvements, it still suffers from segmentation artifacts and incomplete delineation, particularly within the optic cup. In contrast, our method produces more accurate and complete segmentation, demonstrating superior spatial precision and boundary localization.

4.3. Ablation Study

To evaluate the contribution of each component within our proposed framework, we perform an ablation study on Data Adaptation and Model Adaptation. The results across five source domains (A–E) are presented in Table 3.

Applying Data Adaptation alone boosts the average DSC from 65.86% (no adaptation) to 71.93%, with consistent gains across all domains. This underscores the importance of input-level distribution alignment for mitigating domain shifts. With Model Adaptation alone, the average DSC reaches 70.92%. Although slightly less effective than Data Adaptation in some domains (e.g., A and B), it still provides substantial improvement by refining semantic features via adaptive feature normalization. The best performance (73.16%) occurs when both components are enabled, confirming their complementarity: Data Adaptation reduces distribution gaps at the visual level, while Model Adaptation enforces feature-level consistency and prediction robustness.

Overall, these results clearly demonstrate the effectiveness of each component and validate the design of our synergistic data-model adaptation strategy. The integration of both adaptation stages provides superior generalization across diverse domains.

4.4. Discussions

Analysis of Training Iterations in the Model Adaptation Stage. The trend of DSC over adaptation iterations in the modal adaptation stage is presented in Fig. 4 for five source domains (A–E). In most domains, performance peaks after the first iteration and then gradually declines, which indicates that a single iteration is sufficient for optimal adaptation. This degradation arises because only one image is available at each test step, making excessive iterations prone to overfitting and consequently causing catastrophic forgetting. Notably, Domain A behaves differently, where DSC continues to improve with more iterations, suggesting that extended adaptation can sometimes be advantageous.

Effectiveness of Affine Parameter Update. Fig. 5 presents a comparison of segmentation performance under different parameter update strategies on five source domains (A–E). The results clearly show that our adaptation method, which updates only the affine parameters of Batch Normalization layers, consistently outperforms full-model fine-tuning in most domains. This demonstrates the effectiveness and efficiency of focusing on BN parameters for domain adaptation. Specifically, BN-based adaptation achieves higher DSC scores in four out of five domains. The only exception is Domain E, where performance is slightly lower, possibly due to its greater intra-domain variability.

Effectiveness of Continually Updating Model. To validate the effectiveness of continuous model updating, we compare two distinct adaptation strategies. The first is the reset-to-source approach, which reinitializes model parameters to their source-domain configuration after each batch adaptation. The second is the

Table 4: *Performance of two strategies: (1) Reset-to-source model - where parameters are reinitialized to the source model after each batch, and (2) Continuously updated model - where parameters are iteratively refined.*

Strategy	Domain A	Domain B	Domain C	Domain D	Domain E	Average
	DSC	DSC	DSC	DSC	DSC	DSC↑
(1) Reset	74.80	78.42	71.56	58.96	73.00	71.35
(2) Continuously	76.07	78.87	71.18	71.78	67.91	73.16

Table 5: *Performance of two strategies: (1) performing secondary data-model adaptation after the initial data-model adaptation cycle on a batch, and (2) only perform data-model adaptation once.*

Strategy	Domain A	Domain B	Domain C	Domain D	Domain E	Average
	DSC	DSC	DSC	DSC	DSC	DSC↑
(1) Two Cycles	74.44	77.87	71.21	59.17	72.10	70.96
(2) One Cycle	76.07	78.87	71.18	71.78	67.91	73.16

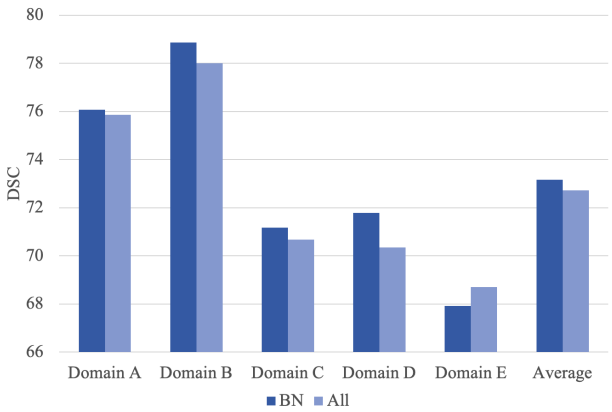

Figure 5: *Segmentation performance between BN-layer-only updating and full-model fine-tuning.*

continuous-update paradigm, where model parameters are progressively updated across iterations, with each subsequent batch processed using the model adapted from the previous one. As shown in Table 4, the continuous-update strategy achieves superior overall results, with an average Dice Similarity Coefficient of 73.16 and an improvement of 1.81 percentage points. It offers substantial gains in challenging domains such as Domain D while maintaining competitive performance across others. Despite the performance drop in Domain E, the overall results highlight the effectiveness of continuous adaptation in accumulating and transferring knowledge across sequential batches, enabling later samples to benefit from earlier adjustments. These findings provide strong empirical support for the importance of maintaining model adaptation continuity to achieve robust cross-domain performance.

Cycles of Data-Model Adaptation. Table 5 compares the segmentation performance of single versus double cycles of data-model adaptation. The results indicate that a single adaptation cycle achieves superior overall performance, with an average DSC of 73.16%, notably outperforming the 70.96% achieved after a second adaptation cycle. The improvement is particularly pronounced in Domain D, suggesting that repeated adaptation may lead to overfitting in certain domains. However, performance in Domain E slightly declines under the single-cycle setting. Overall, although multiple adaptation cycles may benefit specific domains, a carefully optimized single-cycle adaptation demonstrates more robust performance across multiple domains. This may be attributed to its ability to preserve the core knowledge of the source model while still achieving effective domain alignment. In addition to an average DSC improvement of 2.2 percentage points, the single-cycle approach also offers greater computational efficiency by reducing adaptation steps by 50%, making it a more practical and preferable choice for real-world deployment.

5. Conclusion

In this paper, we propose Synergistic Data-Model Adaptation (SDMA) for test-time adaptive medical image segmentation. We introduce a novel use of Batch Normalization layers as a bidirectional bridge to facilitate a two-stage joint adaptation process. The data adaptation operates by learning domain-aware prompts that transform test images in Fourier space to match source-domain style/texture distributions, measured through BN statistical parameter alignment. Concurrently, the model adaptation refines affine parameters through onsistency learning between strong and weak data augmentations, along with entropy minimization. The data and model adaptation align both low-level data characteristics and high-level semantic features through Batch Normalization layer optimization. Extensive experiments on five public retinal fundus datasets demonstrate the effectiveness of our method, which outperforms several state-of-the-art TTA methods.

References

[BGR*06] BORGWARDT K. M., GRETTON A., RASCH M. J., KRIEGEL H.-P., SCHÖLKOPF B., SMOLA A. J.: Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics 22*, 14 (2006), e49–e57. 2

[BY24] BASAK H., YIN Z.: Quest for clone: Test-time domain adaptation for medical image segmentation by searching the closest clone in latent space. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (2024), Springer, pp. 555–566. 2, 3

[CPY*24] CHEN Z., PAN Y., YE Y., LU M., XIA Y.: 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* (2024), pp. 11184–11193. 2, 3, 4, 6

[CYPX25] CHEN Z., YE Y., PAN Y., XIA Y.: Gradient alignment improves test-time adaptation for medical image segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence* (2025), vol. 39, pp. 2429–2437. 2

[FAS*11] FUMERO F., ALAYÓN S., SANCHEZ J. L., SIGUT J., GONZALEZ-HERNANDEZ M.: Rim-one: An open retinal image database for optic nerve evaluation. In *2011 24th international symposium on computer-based medical systems (CBMS)* (2011), IEEE, pp. 1–6. 5

[GBL*23] GAN Y., BAI Y., LOU Y., MA X., ZHANG R., SHI N., LUO L.: Decorate the newcomers: Visual domain prompt for continual test time adaptation. In *Proceedings of the AAAI conference on artificial intelligence* (2023), vol. 37, pp. 7595–7603. 3

[GKL*23] GONG T., KIM Y., LEE T., CHOTTANANURAK S., LEE S.-J.: Sotta: Robust test-time adaptation on noisy data streams. *Advances in Neural Information Processing Systems 36* (2023), 14070–14093. 3

[GWD*23] GONG R., WANG Q., DANELLJAN M., DAI D., VAN GOOL L.: Continuous pseudo-label rectified domain adaptive semantic segmentation with implicit neural representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 7225–7235. 2, 3

[HLX22] HU S., LIAO Z., XIA Y.: Prosfda: Prompt learning based source-free domain adaptation for medical image segmentation. *arXiv preprint arXiv:2211.11514* (2022). 6

[HMW*23] HONG T., MA X., WANG X., CHE R., HU C., FENG T., ZHANG W.: Mapman: Multi-stage u-shaped adaptive pattern matching network for semantic segmentation of remote sensing images. In *Computer Graphics Forum* (2023), vol. 42, Wiley Online Library, p. e14978. 1

[HTP*18] HOFFMAN J., TZENG E., PARK T., ZHU J.-Y., ISOLA P., SAENKO K., EFROS A., DARRELL T.: Cycada: Cycle-consistent adversarial domain adaptation. In *International conference on machine learning* (2018), Pmlr, pp. 1989–1998. 2, 3

[HZRS16] HE K., ZHANG X., REN S., SUN J.: Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016), pp. 770–778. 6

[Kin14] KINGMA D. P.: Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014). 6

[LFW*24] LI X., FANG H., WANG C., LIU M., DUAN L., XU Y.: Cache-driven spatial test-time adaptation for cross-modality medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (2024), Springer, pp. 146–156. 2

[LTC*20] LI C., TAN Y., CHEN W., LUO X., HE Y., GAO Y., LI F.: Anu-net: Attention-based nested u-net to exploit full resolution features for medical image segmentation. *Computers & Graphics 90* (2020), 11–20. 1

[LWS*16] LI Y., WANG N., SHI J., LIU J., HOU X.: Revisiting batch normalization for practical domain adaptation. *arXiv preprint arXiv:1603.04779* (2016). 3

[LXY*21] LIU X., XING F., YANG C., EL FAKHRI G., WOO J.: Adapting off-the-shelf source segmenter for target medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part II 24* (2021), Springer, pp. 549–559. 2

[MMPB22] MIRZA M. J., MICOREK J., POSSEGGER H., BISCHOF H.: The norm must go on: Dynamic unsupervised domain adaptation by normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2022), pp. 14765–14775. 6

[NPS*20] NADO Z., PADHY S., SCULLEY D., D'AMOUR A., LAKSHMINARAYANAN B., SNOEK J.: Evaluating prediction-time batch normalization for robustness under covariate shift. *arXiv preprint arXiv:2006.10963* (2020). 3

[NWZ*23] NIU S., WU J., ZHANG Y., WEN Z., CHEN Y., ZHAO P., TAN M.: Towards stable test-time adaptation in dynamic wild world. *arXiv preprint arXiv:2302.12400* (2023). 2, 3, 5, 6

[OFB*20] ORLANDO J. I., FU H., BREDA J. B., VAN KEER K., BATHULA D. R., DIAZ-PINTO A., FANG R., HENG P.-A., KIM J., LEE J., ET AL.: Refuge challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. *Medical image analysis 59* (2020), 101570. 5

[SKJ*14] SIVASWAMY J., KRISHNADAS S., JOSHI G. D., JAIN M., TABISH A. U. S.: Drishti-gs: Retinal image dataset for optic nerve head (onh) segmentation. In *2014 IEEE 11th international symposium on biomedical imaging (ISBI)* (2014), IEEE, pp. 53–56. 5

[SLL*21] SHUI C., LI Z., LI J., GAGNÉ C., LING C. X., WANG B.: Aggregating from multiple target-shifted sources. In *International conference on machine learning* (2021), PMLR, pp. 9638–9648. 2, 3

[SRE*20] SCHNEIDER S., RUSAK E., ECK L., BRINGMANN O., BRENDEL W., BETHGE M.: Improving robustness against common corruptions by covariate shift adaptation. *Advances in neural information processing systems 33* (2020), 11539–11551. 3

[WFVGD22] WANG Q., FINK O., VAN GOOL L., DAI D.: Continual test-time domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 7201–7211. 2, 3, 5, 6

[WSL*20] WANG D., SHELHAMER E., LIU S., OLSHAUSEN B., DARRELL T.: Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726* (2020). 2, 3, 6

[YCJ*22] YANG H., CHEN C., JIANG M., LIU Q., CAO J., HENG P. A., DOU Q.: Dltta: Dynamic learning rate for test-time adaptation on cross-domain medical images. *IEEE Transactions on Medical Imaging 41*, 12 (2022), 3575–3586. 2, 3, 6

[ZLSZ24] ZHANG X., LI Y., SHENG H., ZHANG X.: Adversarial unsupervised domain adaptation for 3d semantic segmentation with 2d image fusion of dense depth. In *Computer Graphics Forum* (2024), vol. 43, Wiley Online Library, p. e15250. 2, 3

[ZMD*21] ZHANG M., MARKLUND H., DHAWAN N., GUPTA A., LEVINE S., FINN C.: Adaptive risk minimization: Learning to adapt to domain shift. *Advances in Neural Information Processing Systems 34* (2021), 23664–23678. 3

[ZPIE17] ZHU J.-Y., PARK T., ISOLA P., EFROS A. A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision* (2017), pp. 2223–2232. 2, 3

[ZQSG23] ZHANG J., QI L., SHI Y., GAO Y.: Domainadaptor: A novel approach to test-time adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2023), pp. 18971–18981. 6

[ZWA*24] ZHANG X., WU Y., ANGELINI E., LI A., GUO J., RASMUSSEN J. M., O'CONNOR T. G., WADHWA P. D., JACKOWSKI A. P., LI H., ET AL.: Mapseg: Unified unsupervised domain adaptation for heterogeneous medical image segmentation based on 3d masked autoencoding and pseudo-labeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2024), pp. 5851–5862. 2, 3

[ZYL*10] ZHANG Z., YIN F. S., LIU J., WONG W. K., TAN N. M., LEE B. H., CHENG J., WONG T. Y.: 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* (2010), IEEE, pp. 3065–3068. 5