Feature-Space Consistency for Semantics-Preserving Test-Time Adaptation in Medical Segmentation

1st Jianyu Fang
School of Information and Communications Engineering
Xi'an Jiaotong University
Xi'an, China
3454414956@qq.com

2nd Chengyuan Deng
School of Future Technology
Xi'an Jiaotong University
Xi'an, China
cclanro@stu.xjtu.edu.cn

***Abstract*—Test-time adaptation (TTA) offers a practical solution for deploying medical image segmentation models under cross-center domain shifts, where source data are often inaccessible and target annotations are unavailable. However, existing TTA approaches can be unstable in safety-critical settings due to semantic drift in latent representations induced by strong augmentations and confirmation bias from unreliable pseudo supervision, which together lead to noisy updates and error accumulation under continual exposure. In this paper, we propose a reliability-aware and semantics-preserving TTA framework for medical image segmentation. Our method introduces a feature-space consistency constraint that regularizes representation deviation between an input and its strongly augmented view, improving anatomical semantic preservation beyond output-level consistency alone. To further stabilize optimization, we incorporate an entropy-guided gradient-alignment regularization. Moreover, we develop a confidence-aware two-stage adaptation strategy that performs trusted updates first and conditionally refines on hard samples with an early-termination safeguard when reliable evidence is insufficient. We evaluate our approach on cross-domain optic disc/cup segmentation across five public retinal fundus datasets under an online, batch-of-one adaptation protocol that updates only BatchNorm affine parameters. Extensive experiments over 20 cross-domain transfer scenarios demonstrate consistent improvements over prior TTA baselines, achieving state-of-the-art performance under realistic cross-center shifts. Codes available https://github.com/JyFang999/Feature-Space-Consistency.git**

***Index Terms*—Domain shift, Test-time adaptation,Medical Image Segmentation**

I. Introduction

Medical image segmentation [1, 2] aims to delineate anatomical structures and pathological regions (*e.g.* organs, tumors, and lesions) from clinical imaging modalities such as CT, MRI, ultrasound, and digital pathology. As a core component of computer-aided diagnosis and clinical decision-making, accurate segmentation underpins a wide range of downstream applications, including quantitative analysis, longitudinal disease monitoring, surgical navigation, and radiotherapy planning.

In recent years, data-driven approaches, particularly deep convolutional neural networks and their variants [3, 4, 5], have achieved remarkable performance gains by learning hierarchical representations directly from imaging data, substantially reducing reliance on handcrafted features and expert-designed pipelines. Despite these advances, the reliable deployment

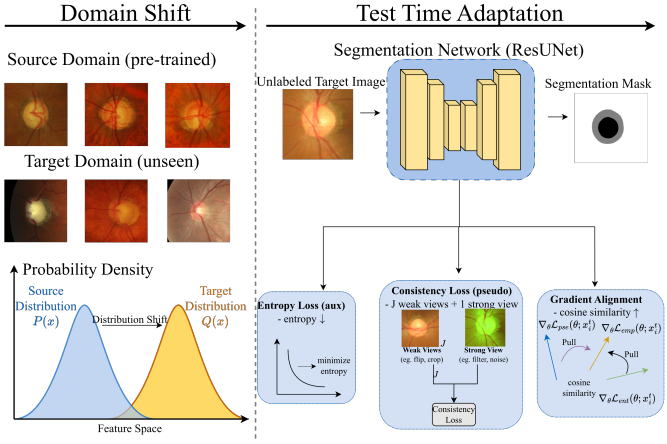

Fig. 1. The concept of domain shift in medical image segmentation and the most common TTA optimization strategies.

of segmentation models in real-world clinical environments remains challenging. A major obstacle arises from domain shifts induced by variations in imaging scanners, acquisition protocols, reconstruction algorithms, and institution-specific population characteristics. As a result, models trained on a source domain often suffer significant performance degradation when applied to unseen target domains, where the data distribution differs from that observed during training. This generalization gap not only limits the scalability of current segmentation systems but also raises critical concerns regarding prediction reliability and patient safety in clinical practice. Addressing domain shift in a deployment-compatible manner therefore remains a central challenge for advancing medical image segmentation toward real-world adoption.

To alleviate performance degradation caused by domain shifts, test-time adaptation (TTA) Fig. 1 has emerged as a practical and deployment-friendly paradigm for medical image segmentation [6, 7]. Unlike conventional domain adaptation methods [8, 9], which typically require access to source data or a pre-collected target dataset, TTA adapts a pre-trained model online during inference using only unlabeled test samples. This setting better aligns with privacy-preserving and real-world clinical scenarios, where data sharing is restricted and target distributions may evolve over time. In TTA, the model is continuously or episodically updated by optimizing self-

supervised objectives on incoming target samples, enabling on-the-fly mitigation of distribution mismatch without relying on explicit target annotations.

Recent advances in test-time adaptation (TTA) for medical image segmentation have increasingly emphasized deployment-oriented designs that prioritize efficiency and stability under clinical constraints. On-the-fly TTA [10] and VPTTA [11] avoid test-time backpropagation and mitigate catastrophic drift; however, their corrective capacity is largely limited to feature-statistic alignment and may degrade under severe or semantic domain shifts. DeTTA [12] introduces label-free objectives but typically assumes corruption-like discrepancies and may oversmooth fine anatomical boundaries. AdaAtlas [13] incorporates anatomical priors through atlas guidance, yet its effectiveness critically depends on robust registration and representative atlas construction. MED-TTT [14] further enables efficient global-context modeling via dynamic inference-time adjustment, at the cost of additional components and hyperparameters that may compromise adaptation stability.

Beyond these deployment-oriented designs, optimization-based approaches have also been explored to improve the stability of test-time updates. For example, GraTa [15] leverages gradient-consistency signals to modulate optimization dynamics, aiming to reduce the instability of naive test-time learning. Nevertheless, when deployed under realistic cross-center domain shifts, existing approaches still share several fundamental limitations. First, reliability control remains limited, as uncertain predictions or hard samples can trigger noisy updates or ineffective conditioning. Second, semantic preservation is often insufficient, where aggressive augmentations, denoising objectives, or statistic matching may distort anatomically meaningful representations. Third, continual exposure to target data can lead to error accumulation over time due to temporal correlation, imperfect priors, or miscalibrated adaptation dynamics. These challenges collectively motivate the need for reliability-aware and semantics-aware test-time adaptation strategies in safety-critical medical image segmentation.

To address these limitations, we revisit test-time adaptation for medical segmentation from the complementary perspectives of reliability and semantic preservation. We observe that performance degradation is often driven by two coupled factors: (i) semantic distortion in deep feature representations induced by strong augmentations under domain shift, and (ii) noisy gradient accumulation caused by unreliable pseudo-supervision during early adaptation stages. Motivated by these observations, we introduce two complementary improvements that explicitly target these failure modes, aiming to stabilize adaptation updates while preserving clinically plausible semantic representations.

In this work, we propose a reliability-aware and semantics-preserving test-time adaptation framework for medical image segmentation, designed to improve the robustness and stability of test-time optimization under realistic cross-center domain shifts. Specifically, we make the following contributions. First, we introduce a feature-space domain shift constraint that enforces semantic consistency between an input image and its augmented counterpart, explicitly regularizing excessive feature deviation during adaptation. By constraining representation drift at the feature level, this design preserves fine-grained anatomical semantics and prevents test-time optimization from converging to over-confident yet clinically implausible representations. Second, we develop a confidence-aware two-stage updating scheme that selectively filters unreliable high-entropy samples and incorporates an early-termination criterion when trustworthy adaptation signals are insufficient. This mechanism serves as a deployment-oriented safeguard against error accumulation and model drift under continual test-time exposure. Together, these designs improve the robustness and stability of test-time optimization, particularly in real-world cross-center scenarios where augmentation-induced representation distortion and pseudo-label noise can critically undermine adaptation performance. Extensive experiments across multiple medical segmentation benchmarks demonstrate that our approach consistently outperforms existing state-of-the-art TTA methods under realistic domain shift settings.

II. Related Work

A. Domain Adaptation and Source-Free Domain Adaptation

Domain adaptation (DA) [16, 17, 18, 19] aims to alleviate performance degradation under distribution shifts across domains. In medical image segmentation, unsupervised domain adaptation (UDA) typically assumes access to labeled source data and unlabeled target data, enabling offline alignment between domains. Representative UDA methods learn domain-invariant representations via joint domain–task objectives [20], leverage pseudo-label-based tri-training [21], or match distribution statistics using discrepancy measures such as central moment discrepancy [22]. However, the co-access assumption of source and target data is often violated in clinical settings due to strict privacy and governance constraints.

To relax this requirement, source-free domain adaptation (SFDA) [23, 24, 25] adapts a model using only a pretrained source model and unlabeled target data. Typical strategies include freezing the source classifier while optimizing information maximization with pseudo-labeling [26], or exploiting neighborhood structures to improve target clustering robustness [27]. While SFDA improves deployability, it is commonly formulated as an offline procedure that produces a fixed adapted model, which can be suboptimal under streaming and non-stationary clinical shifts where target data arrive sequentially across centers. In contrast, our method performs lightweight test-time adaptation without accessing source data and supports continual refinement on incoming unlabeled target samples, while explicitly stabilizing updates via reliability-aware mechanisms and feature-drift regularization.

B. Test-Time Adaptation

Test-time adaptation (TTA) [28, 29, 30] updates a pretrained model using unlabeled target samples at inference time, making it well suited to deployment scenarios where source data are inaccessible and domain shifts emerge after training. A

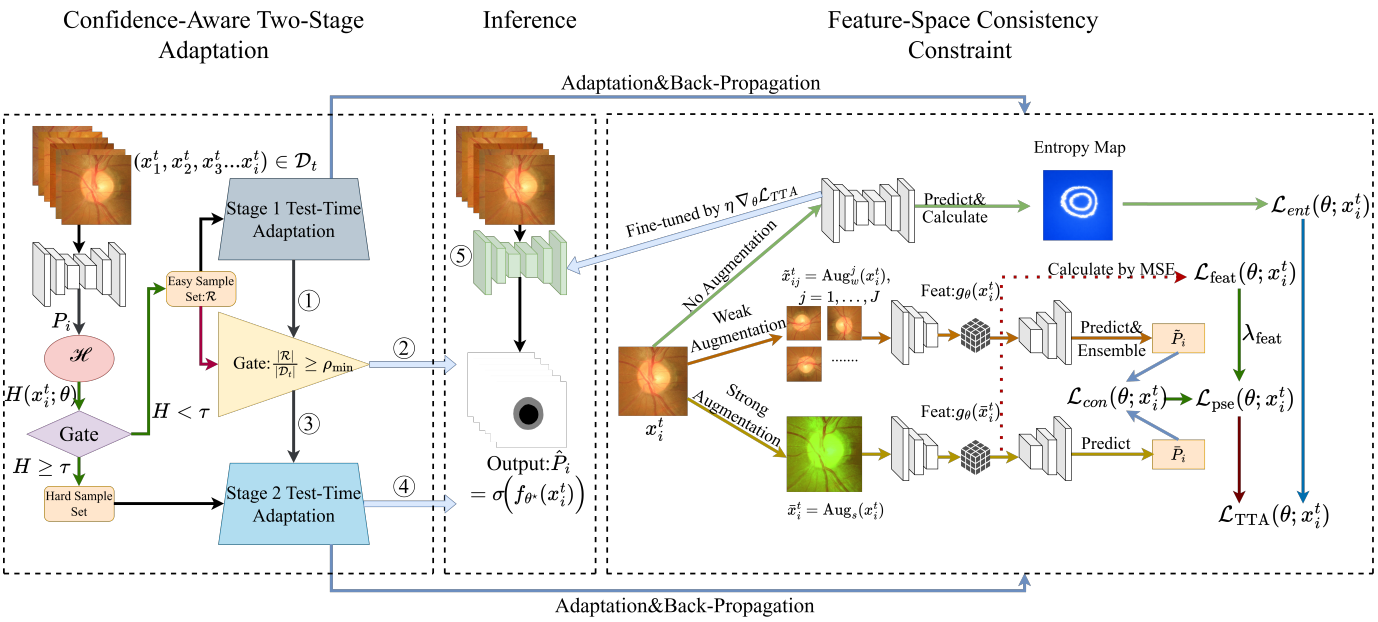

Fig. 2. **Overview.** In Stage 1, we iterate once over $\mathcal{D}_t$, compute an entropy score per sample, and adapt only on confident samples ($H < \tau$) via a single back-propagation step with a feature-space consistency constraint, while skipping the rest. ① After stage 1 all samples finish, if the trusted ratio $|\mathcal{R}|/|\mathcal{D}_t| < \rho_{\min}$, ② we early-stop and directly infer; otherwise, ③ we enter Stage 2 and perform a second pass that adapts only on uncertain samples ($H \geq \tau$) using the same one-step update rule.After ④ stage 2 all samples finish,we proceed to inference.Finally, we ⑤ freeze the model and use the resulting parameters $\theta^\star$ for inference on all test samples.

widely studied line of work adapts normalization statistics or affine parameters at test time (*e.g.* entropy minimization and test-time normalization), which offers an efficient mechanism to reduce distribution mismatch without target annotations.

Recent studies have further considered robustness under non-stationary test streams. For instance, some methods address non-i.i.d. inputs via instance-aware normalization and reservoir sampling [31], while others adopt adaptive test-time normalization to better accommodate varying batch sizes and shifting distributions [32]. To mitigate forgetting and unstable updates in continual settings, reliability-aware strategies such as pseudo-label calibration, neighbor-based filtration, and memory replay have also been explored [33]. Overall, TTA can be viewed as a label-free (zero-shot) adaptation paradigm [34, 35, 36] that improves deployment-time generalization without requiring target annotations.

C. Test-Time Adaptation in Medical Segmentation

Cross-center domain shifts in medical image segmentation have motivated TTA methods that emphasize deployment efficiency and stability. Deployment-oriented designs such as On-the-fly TTA [10] and prompt-based adaptation (VPTTA) [11] avoid test-time backpropagation to mitigate catastrophic drift, but their corrective capacity often centers on feature-statistic alignment and may degrade under severe or semantic shifts. Other approaches incorporate denoising objectives (DeTTA) [12] or anatomical priors (AdaAtlas) [13]; however, their performance can be sensitive to corruption assumptions, registration quality, and the representativeness of priors, potentially affecting fine-structure delineation. Backbone-integrated test-time training (Med-TTT) [14] enhances global-context modeling via dynamic inference-time adjustment, at the cost of additional components and hyperparameters that may complicate stable deployment.

Beyond architecture- and prior-driven designs, optimization-based approaches have also been explored for medical TTA. For example, GraTa [15] improves the stability of test-time optimization by encouraging agreement between pseudo and auxiliary gradients and adapting step sizes based on their cosine similarity. Nevertheless, under continual exposure in realistic clinical streams, optimization-based updates can still be challenged by unreliable pseudo supervision and augmentation-induced representation drift, leading to noisy gradients and semantic distortion. Motivated by these limitations, we introduce a confidence-aware two-stage update scheme with early termination and a feature-drift regularization term to strengthen reliability control and semantic preservation under challenging cross-center shifts.

III. Method

A. Preliminaries

Given an unlabeled test sample $x_i^t \in \mathcal{D}_t$, we consider a standard test-time adaptation (TTA) setup that updates the model on target samples using an entropy-based objective and an augmentation-based consistency objective. We follow a gradient-stabilized test-time optimization formulation below. Let $f_\theta(\cdot)$ denote the network output (logit map) and $\sigma(\cdot)$ the element-wise sigmoid. The predicted probability map for the original input is

$$P_i = \sigma\left(f_\theta(x_i^t)\right). \tag{1}$$

Entropy objective. The entropy objective $\mathcal{L}_{ent}$ is adopted as the auxiliary objective to align with the psedo loss $\mathcal{L}_{pse}$:

$$\mathcal{L}_{ent}(\theta; x_i^t) = -P_i \log\left(P_i + \epsilon\right) \tag{2}$$

Algorithm 1 Confidence-Aware Two-Stage Test-Time Adaptation

Require: Unlabeled test set $\mathcal{D}_t = \{x_i^t\}_{i=1}^N$; pretrained model f_{θ_0}.
Require: Confidence threshold τ; minimum trusted ratio $\rho_{\min}$; learning rate η; numerical constant ϵ.
Require: Objective $\mathcal{L}_{TTA}(\theta; x_i^t)$ (Eq. 8).
Initialize $\theta \leftarrow \theta_0$, trusted set $\mathcal{R} \leftarrow \emptyset$.
Stage 1: trusted online updates
for $i = 1$ to N **do**
  $P \leftarrow \sigma\big(f_\theta(x_i^t)\big)$
  $H \leftarrow -P_i \log\big(P_i + \epsilon\big)$
  if $H < \tau$ **then**
    $\mathcal{R} \leftarrow \mathcal{R} \cup \{x_i^t\}$
    $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}_{TTA}(\theta; x_i^t)$
  end if
end for
$\rho \leftarrow \frac{|\mathcal{R}|}{|\mathcal{D}_t|}$.
Stage 2: hard-sample refinement (if $\rho \geq \rho_{\min}$)
if $\rho \geq \rho_{\min}$ **then**
  for $i = 1$ to N **do**
    $P \leftarrow \sigma\big(f_\theta(x_i^t)\big)$
    $H \leftarrow -P_i \log\big(P_i + \epsilon\big)$
    if $H \geq \tau$ **then**
      $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}_{TTA}(\theta; x_i^t)$
    end if
  end for
end if
Final inference
for $i = 1$ to N **do**
  $\hat{P}_i \leftarrow \sigma\big(f_{\theta^\star}(x_i^t)\big)$
end for
return $\{\hat{P}_i\}_{i=1}^N$

where ϵ is a small constant for numerical stability.

Consistency objective. We construct J weakly augmented views(Aug_w^j denotes one of the following six transformations respectively: identity mapping, horizontal flipping, vertical flipping, and rotations of 90°, 180°, and 270°.) and one strongly augmented view(Aug_s incorporates brightness adjustment, contrast adjustment, gamma transformation, Gaussian noise, and Gaussian blur.):

$$\begin{aligned} \tilde{x}_{ij}^t &= \mathrm{Aug}_w^j(x_i^t), j = 1, \ldots, J, \\ \bar{x}_i^t &= \mathrm{Aug}_s(x_i^t), \end{aligned} \tag{3}$$

and compute the aggregated weak prediction and the strong-view prediction:

$$\begin{aligned} \tilde{P}_i &= \frac{1}{J} \sum_{j=1}^{J} \sigma\big(f_\theta(\tilde{x}_{ij}^t)\big), \\ \bar{P}_i &= \sigma\big(f_\theta(\bar{x}_i^t)\big). \end{aligned} \tag{4}$$

The consistency loss is defined as the pixel-wise cross-entropy between $\bar{P}_i$ and $\tilde{P}_i$:

$$\mathcal{L}_{con}(\theta; x_i^t) = \ell_{ce}\left(\bar{P}_i, \tilde{P}_i\right). \tag{5}$$

This formulation is commonly stabilized by leveraging gradient-consistency signals to modulate the effective update magnitude during test-time optimization [15]. In the following, we introduce two complementary component,the feature-space consistency regularizer and the confidence-aware two-stage update rule.

B. Feature-Space Consistency Constraint

While $\mathcal{L}_{con}$ enforces output-level invariance, strong augmentations may still induce semantic drift in latent representations. To better preserve anatomical semantics during adaptation, we regularize the encoder features between an input sample and its strongly augmented view.

Let $g_\theta(\cdot)$ denote the encoder feature map. Given a test sample x_i^t and its strong augmentation $\bar{x}_i^t = \mathrm{Aug}_s(x_i^t)$ (Eq. 3), we define the feature-space consistency loss as

$$\mathcal{L}_{feat}(\theta; x_i^t) = \left\| g_\theta(x_i^t) - g_\theta(\bar{x}_i^t) \right\|_2^2, \tag{6}$$

Where the squared ℓ_2 distance is computed element-wise on the feature maps and averaged over spatial locations (and channels, if applicable).

We then augment the pseudo objective as

$$\mathcal{L}_{pse}(\theta; x_i^t) = \mathcal{L}_{con}(\theta; x_i^t) + \lambda_{feat}\, \mathcal{L}_{feat}(\theta; x_i^t), \tag{7}$$

Where λ_{feat} controls the strength of feature regularization.

To stabilize optimization at test time, we employ a gradient-alignment regularization between the pseudo objective and the entropy objective. Specifically, we penalize the angular discrepancy between their gradients:

$$\begin{aligned} \mathcal{L}_{TTA}(\theta; x_i^t) = \mathcal{L}_{pse}(\theta; x_i^t) \\ + \angle\left(\nabla_\theta \mathcal{L}_{pse}(\theta; x_i^t),\, \nabla_\theta \mathcal{L}_{ent}(\theta; x_i^t)\right), \end{aligned} \tag{8}$$

Where $\angle(\mathbf{a}, \mathbf{b})$ denotes the angle between vectors $\mathbf{a}$ and $\mathbf{b}$ in parameter space (smaller values indicate higher directional consistency). In practice, we implement $\angle(\mathbf{a}, \mathbf{b})$ via a cosine-based surrogate, which is $1 - \frac{\mathbf{a}^\top \mathbf{b}}{\|\mathbf{a}\|\|\mathbf{b}\| + \epsilon}$, to ensure differentiability and numerical stability.Through gradient alignment, the model extracts shared task-relevant information from different gradients, thereby better approximating the empirical gradient.

C. Confidence-Aware Two-Stage Adaptation

Directly adapting on all test samples can suffer from confirmation bias when predictions are highly uncertain. We therefore adopt a confidence-driven curriculum: Stage 1 updates the model only on trusted samples, and Stage 2 is enabled only when Stage 1 has accumulated sufficient trusted evidence.

Uncertainty score. Given the probability map $P_i = \sigma\big(f_\theta(x_i^t)\big)$, we measure prediction uncertainty using the entropy:

$$H(x_i^t; \theta) = -P_i \log\big(P_i + \epsilon\big) \tag{9}$$

where ϵ is a small constant for numerical stability. A smaller H indicates higher confidence.

Stage 1: trusted online updates. Let $\mathcal{D}_t = \{x_i^t\}_{i=1}^N$ be the unlabeled test stream. We perform a single online pass and update the model only for confident samples. For each x_i^t, we compute $H(x_i^t; \theta)$ and accept x_i^t as a trusted sample

TABLE I
PERFORMANCE (DSC, %) OF STATE-OF-THE-ART TTA METHODS ON FIVE DOMAINS (A–E). BEST IN EACH COLUMN IS IN BOLD.

Method	A	B	C	D	E	Avg.
No Adapt	67.96	76.71	74.71	54.07	68.88	68.47
DUA	73.10	77.10	75.34	58.62	72.99	71.43
DIGA	76.57	77.10	73.64	62.04	71.54	72.18
MedBN	75.14	76.77	74.46	59.91	72.00	71.65
TENT	74.06	78.55	74.56	51.45	69.66	69.65
DLTTA	75.19	78.48	76.37	57.50	71.35	71.78
DomainAdaptor	75.88	77.43	75.66	58.64	72.38	72.00
SAR	75.26	78.07	75.07	61.24	**73.69**	72.67
DeTTA	75.15	78.17	75.27	61.26	73.61	72.69
GraTa	76.61	78.81	76.52	66.84	72.94	74.34
Ours	**78.19**	**79.36**	**76.92**	**68.84**	72.29	**75.12**

if $H(x_i^t;\theta) < \tau$, where τ is a fixed confidence threshold. Trusted samples are collected into $\mathcal{R}$, and we perform an online gradient update by minimizing $\mathcal{L}_{TTA}$ (Eq. 8):

$$\begin{aligned} &\mathcal{R} \leftarrow \mathcal{R} \cup \{x_i^t\}, \\ &\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}_{TTA}(\theta; x_i^t), \quad \text{if } H(x_i^t;\theta) < \tau, \end{aligned} \tag{10}$$

where η denotes the test-time learning rate. Otherwise, the update is skipped. All decisions are made sequentially under the continuously updated model.

Stage gating and early termination. After Stage 1, we compute the trusted ratio $\rho = \frac{|\mathcal{R}|}{|\mathcal{D}_t|}$. If $\rho < \rho_{\min}$, we terminate adaptation early to avoid unreliable updates under severe shifts; otherwise, Stage 2 is activated.

Stage 2: hard-sample online refinement. When $\rho \geq \rho_{\min}$, we perform another online pass over $\mathcal{D}_t$ and update the model only on hard samples ($H(x_i^t;\theta) \geq \tau$):

$$\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}_{TTA}(\theta; x_i^t), \text{if } H(x_i^t;\theta) \geq \tau, \forall\, x_i^t \in \mathcal{D}_t. \tag{11}$$

This guarded refinement focuses computation on uncertain cases while mitigating drift when confidence is insufficient.

Final inference. Let $\theta^\star$ be the final parameters after adaptation. We produce the final probability maps as

$$\hat{P}_i = \sigma\big(f_{\theta^\star}(x_i^t)\big), \qquad \forall\, x_i^t \in \mathcal{D}_t, \tag{12}$$

followed by standard post-processing (*e.g.* thresholding for binary segmentation) to obtain the final masks.

D. Framework Overview

Fig. 2 provides an overview of our test-time adaptation pipeline. For each unlabeled test sample, we form the pseudo objective by combining output-level consistency with the proposed feature-space regularization (Eqs. 5 and 6), and stabilize optimization by coupling it with the entropy objective through gradient-alignment regularization (Eq. 8). Model updates are then executed sequentially using the confidence-aware two-stage strategy, which updates on trusted samples and conditionally refines on hard samples with an early-termination safeguard (Eqs. 10 and 11). The complete adaptation procedure is summarized in Alg. 1, and the final predictions are produced using the adapted parameters $\theta^\star$ (Eq. 12).

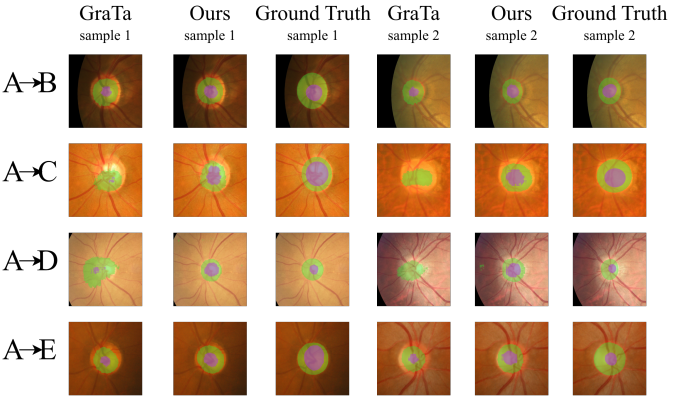

Fig. 3. This figure present the qualitative analysis results comparing GraTa, our method, and the ground truth.

IV. EXPERIMENTS

A. Dataset and Evaluation Metrics

We evaluate cross-domain optic disc (OD) and optic cup (OC) segmentation on five public retinal fundus datasets: RIM-ONE-r3, REFUGE, ORIGA, REFUGE-Validation/Test, and Drishti-GS, denoted as domains A–E, respectively. The datasets contain 159 (A), 400 (B), 650 (C), 800 (D), and 101 (E) images collected from different medical centers, resulting in substantial domain shifts. For preprocessing, we crop an 800×800 region of interest centered on the optic disc, resize it to 512×512, and apply min–max intensity normalization. Performance is measured by the Dice Similarity Coefficient (DSC; higher is better).Dice Similarity Coefficient (DSC) is used to measure the overlap between the predicted region P and the ground-truth region G, which is defined as $\text{DSC} = \frac{2|P \cap G|}{|P|+|G|}$.

B. Implementation Details

Our segmentation backbone is ResUNet-34 (ResNet-34 encoder + UNet decoder). We train five source-specific models with supervised segmentation loss, and evaluate cross-domain generalization under a 4×5 protocol (each source adapted/tested on the other four targets, then averaged). Test-time adaptation is performed in a single-step, online manner (batch size $= 1$), updating only affine-transformation parameters while recomputing BN statistics on-the-fly. Each test image triggers a lightweight update driven by an entropy objective on the original image and an augmentation-based consistency objective, augmented with a feature-space regularizer $\mathcal{L}_{feat}$ to constrain representation drift. We set the learning rate η to 0.0001 empirically, and employ a confidence-aware two-stage schedule that prioritizes low-entropy samples and enables hard sample updates only when the easy sample ratio exceeds $\rho_{\min}$.

C. Comparison with State-of-the-Art

We compare our approach against the other state-of-the-art TTA methods on the cross-domain OD/OC segmentation benchmark. Table I summarizes the mean DSC achieved by each method on domains A–E, where the result of each domain is averaged over four transfer scenarios (the four transfer

TABLE II
ABLATION STUDY (DSC, %) OF OUR PROPOSED COMPONENTS. BEST IN EACH COLUMN IS IN BOLD.

Method	A	B	C	D	E	Avg.
No Adapt	67.96	76.71	74.71	54.07	68.88	68.47
Method1 (feature-space constraint only)	77.51	79.16	76.36	67.43	72.01	74.49
Method2 (two-stage adaptation only)	77.00	78.94	76.53	67.07	**73.07**	74.52
Ours	**78.19**	**79.36**	**76.92**	**68.84**	72.29	**75.12**

scenarios are target domain corresponding to current domain). The compared methods include: No Adapt, the baseline model evaluated without any test-time adaptation; three BN statistics adaptation methods (DUA, DIGA, MedBN); and six fine-tuning based methods (TENT, DLTTA, DomainAdaptor, SAR, DeTTA,GraTa) that optimize model parameters via self-supervised losses. All methods use the same pre-trained source models for fairness.Table I shows that our method achieves the best overall performance, reaching 75.12 average DSC,*i.e.* , +6.65 over No Adapt (68.47), with the largest gain on the most challenging domain D (+14.77, 54.07→68.84). Ours consistently outperforms prior normalization/statistics-based adaptation (71.43–72.18 avg.) by +2.94–3.69 and optimization-based TTA method (69.65–72.69 avg.) by +2.43–5.47, indicating stronger robustness under cross-center shifts. Compared to the strongest method in the table GraTa (74.34 avg.), Ours further improves by +0.78 on average, driven primarily by a notable boost on domain D (+2.00), while maintaining competitive performance on the remaining domains.

Qualitative analysis. GraTa achieves the best average performance among the evaluated TTA baselines; thus,We adopt GraTa as a representative state-of-the-art method for visualization. We fix Domain A as the current domain and use the remaining four domains as targets for qualitative analysis. For each target domain, we visualize two representative samples and compare GraTa, our method, and the ground truth Fig. 3.

D. Ablation Study

To quantify the contributions of our proposed enhancements, we conduct an ablation study with four variants: the No Adapt,feature-space constraint only, two-stage adaptation only, and the full proposed method(Ours). Table II validates the effectiveness and complementarity of our two components. Both single modules already yield large gains over No Adapt (68.47 avg.): the feature-space constraint alone reaches 74.49 (+6.02), and the two-stage adaptation alone achieves 74.52 (+6.05), demonstrating that each independently mitigates domain-shift degradation. Combining them further improves performance to 75.12 (+6.65 over No Adapt), with consistent best results on the most challenging domains, especially D (54.07→68.84, +14.77). Overall, the full method provides the strongest average accuracy, indicating that constraining representation drift and scheduling updates by confidence jointly enhance test-time adaptation stability and effectiveness.

E. Parameter Analysis

We further analyze key hyperparameters and design choices in our approach.

Feature-space constraint loss weight λ_{feat}. We vary λ_{feat} and observe that $\lambda_{feat} = 0.2$ provides the best trade-off between effective adaptation and preserving anatomically meaningful representations. Notably, $\lambda_{feat} = 0.2$ suggests that, in medical scenarios, a lightweight intervention is often optimal: it allows necessary feature refinement while retaining the majority ($\sim$80%) of the original feature distribution, thereby maintaining robustness under cross-domain shifts. This term acts as a feature-space anchor that constrains the encoder to adapt to appearance (*i.e.* style) variations without drifting away from core structural semantics. Empirically, larger weights tend to over-regularize and limit target adaptation, while smaller weights offer insufficient anchoring and lead to less stable updates.

Confidence threshold τ and minimum easy-sample ratio $\rho_{\min}$. Unless otherwise specified, we set $\tau = 0.2$ and $\rho_{\min} = 70\%$, which already yield strong and stable performance across domains. These two parameters implement an adaptive trust boundary and a safety gate: τ controls the reliability–coverage trade-off when selecting low-entropy easy samples for early adaptation, while $\rho_{\min}$ prevents premature exposure to hard and noisy samples by enabling Stage 2 only after sufficient confident coverage is achieved. Although these defaults work well in our setting, more exhaustive sweeps (or domain-adaptive calibration) of τ and $\rho_{\min}$ are promising to further improve robustness under varying shift severity.

V. CONCLUSION

We proposed a reliability-aware and semantics-preserving test-time adaptation framework for medical image segmentation under cross-center domain shifts. Our approach combines feature-space consistency regularization with a confidence-aware two-stage update rule to reduce representation drift and mitigate unreliable updates during continual adaptation. Experiments on cross-domain optic disc/cup segmentation across five public fundus datasets demonstrate that the proposed method consistently improves robustness in an online, batch-of-one setting while maintaining deployment efficiency. Future work will extend the framework to broader modalities and multi-class tasks and explore more principled uncertainty modeling for test-time update control.

REFERENCES

[1] R. Azad, E. K. Aghdam, A. Rauland, Y. Jia, A. H. Avval, A. Bozorgpour, S. Karimijafarbigloo, J. P. Cohen, E. Adeli, and D. Merhof, "Medical image segmentation review: The success of u-net," *IEEE TPAMI*, vol. 46, no. 12, pp. 10 076–10 095, 2024.

[2] H. Xiao, L. Li, Q. Liu, X. Zhu, and Q. Zhang, "Transformers in medical image segmentation: A review," *Biomed. Signal Process. Control*, vol. 84, p. 104791, 2023.

[3] P.-H. Conze, G. Andrade-Miranda, V. K. Singh, V. Jaouen, and D. Visvikis, "Current and emerging trends in medical image segmentation with deep learning," *IEEE Trans. Radiat. Plasma Med. Sci.*, vol. 7, no. 6, pp. 545–569, 2023.

[4] D. Jha, M. A. Riegler, D. Johansen, P. Halvorsen, and H. D. Johansen, "Doubleu-net: A deep convolutional

neural network for medical image segmentation," in *Int. Symp. Comput.-Based Med. Syst.*, 2020, pp. 558–564.

[5] A. Khan, Z. Rauf, A. Rehman Khan, S. Rathore, S. Hussain Khan, N. Saher Shah, U. Farooq, H. Asif, A. Asif, U. Zahoora, R. Ullah Khalil, S. Qamar, U. Hani Tayyab, F. Babar Khan, A. Majid, and J. Gwak, "A recent survey of vision transformers for medical image segmentation," *IEEE Access*, vol. 13, pp. 191 824–191 849, 2025.

[6] M. Hu, T. Song, Y. Gu, X. Luo, J. Chen, Y. Chen, Y. Zhang, and S. Zhang, "Fully test-time adaptation for image segmentation," in *MICCAI*, 2021, pp. 251–260.

[7] Y. He, A. Carass, L. Zuo, B. E. Dewey, and J. L. Prince, "Autoencoder based self-supervised test-time adaptation for medical image analysis," *Med. Image Anal.*, vol. 72, p. 102136, 2021.

[8] M. Bateson, J. Dolz, H. Kervadec, H. Lombaert, and I. B. Ayed, "Constrained domain adaptation for image segmentation," *IEEE Trans. Med. Imag.*, vol. 40, no. 7, pp. 1875–1887, 2021.

[9] P. Zhang, J. Li, Y. Wang, and J. Pan, "Domain adaptation for medical image segmentation: A meta-learning method," *J. Imag.*, vol. 7, no. 2, 2021.

[10] J. M. J. Valanarasu, P. Guo, V. VS, and V. M. Patel, "On-the-fly test-time adaptation for medical image segmentation," in *Med. Imaging Deep Learn.*, vol. 227, Jul. 2024, pp. 586–598.

[11] Z. Chen, Y. Pan, Y. Ye, M. Lu, and Y. Xia, "Each test image deserves a specific prompt: Continual test-time adaptation for 2d medical image segmentation," in *CVPR*, Jun. 2024, pp. 11 184–11 193.

[12] R. Wen, H. Yuan, D. Ni, W. Xiao, and Y. Wu, "From denoising training to test-time adaptation: Enhancing domain generalization for medical image segmentation," in *WACV*, Jan. 2024, pp. 464–474.

[13] J. Guo, W. Zhang, M. Sinclair, D. Rueckert, and C. Chen, "Pay attention to the atlas: Atlas-guided test-time adaptation method for robust 3d medical image segmentation," *arXiv*, 2023.

[14] J. Xu, "Med-ttt: Vision test-time training model for medical image segmentation," *arXiv*, 2024.

[15] Z. Chen, Y. Ye, Y. Pan, and Y. Xia, "Gradient alignment improves test-time adaptation for medical image segmentation," *AAAI*, vol. 39, no. 3, pp. 2429–2437, Apr. 2025.

[16] J. Hu, J. Lu, and Y.-P. Tan, "Deep transfer metric learning," in *CVPR*, Jun. 2015.

[17] S. Motiian, M. Piccirilli, D. A. Adjeroh, and G. Doretto, "Unified deep supervised domain adaptation and generalization," in *ICCV*, Oct. 2017.

[18] T. Gebru, J. Hoffman, and L. Fei-Fei, "Fine-grained recognition in the wild: A multi-task domain adaptation approach," in *ICCV*, Oct. 2017.

[19] H. Yan, Y. Ding, P. Li, Q. Wang, Y. Xu, and W. Zuo, "Mind the class weight bias: Weighted maximum mean discrepancy for unsupervised domain adaptation," in *CVPR*, Jul. 2017.

[20] E. Tzeng, J. Hoffman, T. Darrell, and K. Saenko, "Simultaneous deep transfer across domains and tasks," in *ICCV*, Dec. 2015.

[21] K. Saito, Y. Ushiku, and T. Harada, "Asymmetric tri-training for unsupervised domain adaptation," in *ICML*, vol. 70, Aug. 2017, pp. 2988–2997.

[22] W. Zellinger, T. Grubinger, E. Lughofer, T. Natschläger, and S. Saminger-Platz, "Central moment discrepancy (CMD) for domain-invariant representation learning," in *ICLR*, 2017.

[23] Q. Zang, S. Wang, D. Zhao, Y. Hu, D. Quan, J. Li, N. Sebe, and Z. Zhong, "Generalized source-free domain-adaptive segmentation via reliable knowledge propagation," in *ACM MM*, 2024, pp. 5967–5976.

[24] Y. Mitsuzumi, A. Kimura, and H. Kashima, "Understanding and improving source-free domain adaptation from a theoretical perspective," in *CVPR*, Jun. 2024, pp. 28 515–28 524.

[25] J. Li, Z. Yu, Z. Du, L. Zhu, and H. T. Shen, "A comprehensive survey on source-free domain adaptation," *IEEE TPAMI*, vol. 46, no. 8, pp. 5743–5762, 2024.

[26] J. Liang, D. Hu, and J. Feng, "Do we really need to access the source data? Source hypothesis transfer for unsupervised domain adaptation," in *ICML*, vol. 119, Jul. 2020, pp. 6028–6039.

[27] S. Yang, y. wang, J. van de Weijer, L. Herranz, and S. Jui, "Exploiting the intrinsic neighborhood structure for source-free domain adaptation," in *NeurIPS*, vol. 34, 2021, pp. 29 393–29 405.

[28] D. Wang, E. Shelhamer, S. Liu, B. Olshausen, and T. Darrell, "Tent: Fully test-time adaptation by entropy minimization," in *ICLR*, 2021.

[29] D. Chen, D. Wang, T. Darrell, and S. Ebrahimi, "Contrastive test-time adaptation," in *CVPR*, Jun. 2022, pp. 295–305.

[30] Q. Wang, O. Fink, L. Van Gool, and D. Dai, "Continual test-time domain adaptation," in *CVPR*, Jun. 2022, pp. 7201–7211.

[31] T. Gong, J. Jeong, T. Kim, Y. Kim, J. Shin, and S.-J. Lee, "Note: Robust continual test-time adaptation against temporal correlation," in *NeurIPS*, vol. 35, 2022, pp. 27 253–27 266.

[32] H. Lim, B. Kim, J. Choo, and S. Choi, "TTN: A domain-shift aware batch normalization in test-time adaptation," in *ICLR*, 2023.

[33] T. H. Rafi, A. Agarwal, H. L. Patel, and D.-K. Chae, "Towards robust continual test-time adaptation via neighbor filtration," in *CIKM*, 2025, pp. 5161–5165.

[34] S. Niu, C. Miao, G. Chen, P. Wu, and P. Zhao, "Test-time model adaptation with only forward passes," in *ICML*, vol. 235, Jul. 2024, pp. 38 298–38 315.

[35] S. Niu, J. Wu, Y. Zhang, Z. Wen, Y. Chen, P. Zhao, and M. Tan, "Towards stable test-time adaptation in dynamic wild world," in *ICLR*, 2023.

[36] H. Li, P. Hu, Q. Zhang, X. Peng, XitingLiu, and M. Yang, "Test-time adaptation for cross-modal retrieval with query shift," in *ICLR*, 2025.