Multi-View Structural Consistency Learning for Test-Time Adaptation in Medical Image Segmentation

Shaowei Fan
School of Artificial Intelligence and Data Science
Hebei University of Technology
Tianjin, China
gelbartschiff@gmail.com

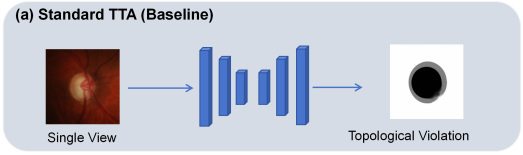
***Abstract*—Cross-domain distribution shifts substantially degrade the reliability of deep fundus image segmentation models in real clinical deployments. Test-time adaptation (TTA) offers an appealing solution by updating models online using unlabeled target data, yet existing segmentation TTA methods often rely on probability-only objectives that can over-smooth uncertain boundaries and may suffer from semantic drift under unconstrained updates. We propose Multi-View Structural Consistency Learning (MVSCL), a structure-aware and parameter-constrained TTA framework for joint optic disc and optic cup segmentation. MVSCL constructs multiple geometric views of each test image via flips and multi-scale resizing, and enforces consistency under a teacher–student formulation across output probabilities, Sobel-based boundary responses, and decoder feature representations, thereby preserving anatomically meaningful morphology under domain shift. To stabilize online adaptation, we further restrict optimization to the Batch Normalization affine parameters in the decoder and segmentation head, while keeping the encoder frozen to retain generalizable representations. Extensive experiments on five public fundus datasets under a single-source training and multi-target testing protocol demonstrate that MVSCL consistently improves robustness and achieves state-of-the-art performance on the evaluated cross-domain benchmarks without requiring source data.The code is available at https://github.com/gelbartschiff-cloud/MVSCL.git**

***Index Terms*—Test-Time Adaptation, Fundus Image Segmentation, Consistency Learning**

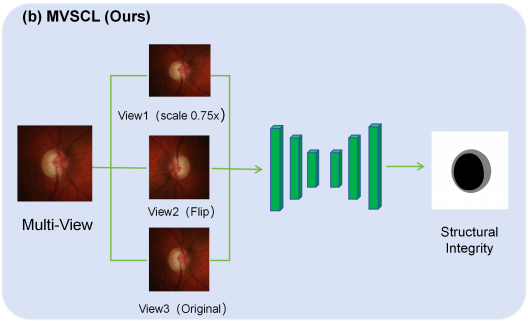
I. Introduction

While deep neural networks have achieved remarkable success in medical image segmentation—a critical prerequisite for clinical diagnosis and treatment planning [1], [2]—their deployment in real-world clinical environments remains challenging. In particular, variations in scanners/vendors, acquisition protocols, and preprocessing pipelines inevitably induce distribution shifts between the source and target domains, which can substantially degrade the performance of models trained on the source data when applied to unseen target data [3].

To address domain shift, test-time adaptation (TTA) has emerged as an efficient paradigm that adapts a pre-trained model at inference time using only unlabeled test (target-domain) data, thereby reducing performance degradation under distribution shifts without accessing source-domain samples.



(a) Standard TTA



(b) MVSCL (Ours)

Fig. 1: **Benefit of multi-view.** (a) Standard TTA suffers from **topological violations** due to lack of constraints. (b) MVSCL leverages diverse views to enforce geometric invariance, effectively preserving **structural integrity** and preventing boundary over-smoothing.

Early studies largely focused on correcting feature distribution mismatch by updating Batch Normalization (BN) statistics [4]. More recent approaches extend beyond BN-statistics updating by leveraging self-supervised objectives such as contrastive learning [5], as well as normalization variants designed to better balance or recalibrate feature distributions under shift [6]. In parallel, Yuan *et al.* proposed TEA to explicitly account for covariate shift that can otherwise lead to decreased generalizability during test-time adaptation [7].

However, in more complex settings such as medical image segmentation, prior TTA techniques developed primarily for image classification may not transfer reliably. Unlike classification, segmentation is a dense prediction problem that requires pixel-wise discrimination and spatially consistent

representations, making it more sensitive to subtle feature misalignment and local distribution shifts. Many existing TTA methods adapt models to the target domain via unsupervised objectives (e.g., entropy minimization or regularization) defined on unlabeled test samples, and update parameters online through backpropagation. This gradient-based update mechanism can introduce non-trivial computational overhead and optimization instability at inference time, and may also risk overfitting to individual test cases or degrading generalization.

To address these challenges in segmentation-oriented TTA, Valanarasu *et al.* proposed On-the-Fly Adaptation, which equips each convolutional block with an adaptive Batch Normalization module to modulate features conditioned on domain codes [8]. Wen *et al.* further introduced Denoising Test-Time Adaptation (DeTTA), adapting the model on a per-sample basis to better handle target-domain shift and noise-corrupted inputs [9]. More broadly, to further exploit model capacity under domain shift, the prevailing paradigm has shifted towards gradient-based optimization with self-supervised pretext tasks, which fine-tunes model parameters during inference to improve generalization without source data [10].

Despite these advances, existing segmentation TTA strategies often struggle to preserve fine-grained anatomical structures. In particular, standard consistency regularization defined on output probabilities can impose overly rigid pixel-wise alignment across perturbations or views. When applied to boundary regions with inherently high uncertainty, such constraints tend to suppress high-frequency boundary responses and induce over-smoothed predictions. This is especially detrimental for small and ambiguous structures such as the optic cup in fundus images, whose boundaries are frequently blurred and whose area is limited. As a result, critical morphological details may be attenuated during adaptation, undermining the clinical reliability of the final segmentation [11].

To address these issues, we propose **Multi-View Structural Consistency Learning (MVSCL)** for domain-shifted fundus segmentation. Within a teacher–student framework, MVSCL enforces complementary consistency across multiple views, including output probabilities, boundary responses, and decoder feature representations, thereby explicitly modeling the structural invariance of the optic disc and optic cup. Moreover, to mitigate semantic drift induced by unsupervised test-time updates, we adopt a parameter-efficient optimization strategy that restricts adaptation to only the affine parameters (γ, β) of Batch Normalization layers in the decoder and segmentation head. By jointly leveraging multi-view structural consistency and parameter-efficient fine-tuning, our method adapts effectively to the target distribution without additional supervision or costly retraining. Extensive experiments show that MVSCL achieves new state-of-the-art performance on cross-domain fundus segmentation benchmarks, delivering superior accuracy particularly in boundary-ambiguous regions.

II. Related Work

A. Image Segmentation

Image segmentation is a fundamental computer vision task that partitions an image into semantically meaningful regions by assigning a label to each pixel, enabling precise delineation of object boundaries and anatomical structures in complex scenes [12]. Classic deep segmentation frameworks include Fully Convolutional Networks (FCNs) and their variants [13], [14], as well as architectures that incorporate region-based modeling or iterative refinement mechanisms to improve localization and boundary quality [15]–[18].

Recent advances have further improved segmentation performance in medical imaging by introducing stronger backbone designs and more effective feature modeling. For example, hybrid CNN–Transformer architectures (e.g., TransUNet) enhance global context modeling for accurate delineation [19], while Mamba-style sequence modeling has also been explored to strengthen long-range dependency capture in segmentation pipelines (e.g., U-Mamba and Swin-UMamba) [20], [21].

Despite these advances, most segmentation models are trained under the closed-world assumption and remain fixed after training, making them vulnerable to domain shifts caused by changes in scanners/vendors, acquisition protocols, or preprocessing pipelines in real clinical deployments. This limitation motivates research on **Test-Time Adaptation (TTA)**, which aims to adapt a pre-trained model to the target (test) distribution during inference without requiring target-domain annotations.

B. Test-Time Adaptation

To mitigate these distribution shifts, **Test-Time Adaptation (TTA)** updates a pre-trained model using unlabeled target-domain data during inference [22], [23]. A representative line of work performs entropy minimization and adapts only a small subset of parameters (e.g., normalization statistics) to achieve stable and efficient test-time updates [23].

Recent studies have further explored efficiency and robustness in non-stationary settings. Karmanov *et al.* [24] proposed backpropagation-free adaptation using key–value caching to reduce online computation. Marsden *et al.* [25] investigated universal test-time adaptation and addressed bias and forgetting via mechanisms such as weight ensembling. Zhu *et al.* [26] proposed a shape-aware continual TTA framework that improves generalization while mitigating catastrophic forgetting through regularization and weight-reset strategies.

Owing to the strict privacy regulations and ethical constraints in medical data sharing, TTA provides an appealing solution for medical image segmentation, as it enables target-domain adaptation without accessing the original source data.

C. Test-Time Adaptation for Medical Image Segmentation

In medical image segmentation, domain shifts are pervasive due to heterogeneous scanners, protocols, and patient populations, which can significantly degrade model performance when deployed across centers. Accordingly, recent work has

developed segmentation-oriented TTA methods tailored to dense prediction and clinical noise characteristics.

Valanarasu *et al.* [8] proposed an on-the-fly adaptation scheme that enables fast per-image adaptation by modulating adaptive Batch Normalization layers using pre-trained domain codes, avoiding gradient-based backpropagation during inference. Wen *et al.* [9] introduced Denoising Test-Time Adaptation (DeTTA), which leverages denoising-driven objectives with an auxiliary decoder to improve robustness to noise-corrupted target inputs and support sample-level adaptation. Li *et al.* [27] developed a multi-view TTA approach based on multi-view knowledge distillation to facilitate target-domain updates without data centralization.

Despite their progress in reducing distribution mismatch, many existing segmentation TTA methods primarily focus on aligning output distributions or minimizing prediction uncertainty, and may be insufficient to preserve fine-grained anatomical structures (e.g., small regions and ambiguous boundaries) under substantial domain shift. This observation motivates structure-aware adaptation strategies that explicitly regularize morphological consistency while restricting the update scope to improve robustness on unseen target domains.

III. Method

To address cross-domain distribution shifts in fundus image segmentation, we propose a structure-aware TTA framework. Our approach synergizes Multi-View Structural Consistency Learning (MVSCL) with a Parameter-Constrained Adaptation strategy. MVSCL exploits intrinsic structural stability to promote consistency across output probabilities, boundaries, and features, while the parameter constraint prevents semantic drift by restricting updates to specific affine parameters. Together, these designs enable robust adaptation without external supervision.

A. Preliminaries and Problem Definition

Let f_θ denote a segmentation model pre-trained on a labeled source domain $\mathcal{D}_s = \{(x_s, y_s)\}$. In the Test-Time Adaptation (TTA) setting, the model adapts to an unlabeled target stream $\mathcal{D}_t = \{x_t\}$ in a fully online manner. For each sequentially arriving sample x_t, TTA aims to update parameters $\theta \to \theta_t$ in real-time, strictly without accessing $\mathcal{D}_s$, by minimizing prediction error solely via self-supervision.

State-of-the-art TTA methods, such as GraTa [10], primarily focus on resolving the gradient conflicts between the unsupervised entropy minimization objective $\mathcal{L}_{\text{ent}}$ and the consistency regularization objective $\mathcal{L}_{\text{con}}$ through gradient alignment mechanisms. GraTa introduces a lookahead mechanism for dynamic gradient alignment. Specifically, given x_t, it first computes $\mathcal{L}_{ent}$ and performs a virtual gradient step to obtain a temporary state θ':

$$\theta' = \theta - \nabla_\theta \mathcal{L}ent(\theta; x_t) \tag{1}$$

Subsequently, the consistency loss $\mathcal{L}con(\theta'; x_t)$ is computed on θ'. By Taylor expansion, minimizing $\mathcal{L}_{con}$ at this virtual state implicitly maximizes the gradient inner product of both tasks, enforcing direction convergence:

$$\begin{aligned} \min_\theta \mathcal{L}con(\theta - \nabla\mathcal{L}ent) \approx \\ \min_\theta \left(\mathcal{L}con(\theta) - \nabla\mathcal{L}con \cdot \nabla\mathcal{L}ent\right) \end{aligned} \tag{2}$$

Finally, the parameters are updated using an adaptive learning rate η based on gradient cosine similarity:

$$\theta \leftarrow \theta - \eta \nabla\theta \mathcal{L}_{con}(\theta'; x_t) \tag{3}$$

While GraTa effectively mitigates gradient conflicts, its reliance on pixel-level consistency is insufficient for handling severe structural shifts in fundus imagery, and its unconstrained updates risk semantic drift. To overcome these limitations, we introduce MVSCL to enforce structural invariance and a strict parameter update strategy to preserve pre-trained knowledge.

B. Multi-View Structural Consistency Learning

Given an unlabeled target-domain test image x_t, we construct a set of augmented views to simulate geometric variations that commonly occur in fundus imaging. Concretely, we consider a set of augmentation operators $\mathcal{T}$ composed of horizontal/vertical flips and multi-scale resizing with scale factors $s \in \{0.75, 1.0, 1.25\}$. For each operator $t \in \mathcal{T}$, we generate an augmented view $x_t^{(t)} = t(x_t)$. These transformations mimic changes in camera viewpoint and imaging distance.

For each view, the model produces a segmentation probability map $\hat{y}^{(t)} \in [0,1]^{H\times W\times C}$ as well as a decoder feature map $f^{(t)} \in \mathbb{R}^{H\times W\times D}$. We adopt a teacher–student formulation where the prediction and feature of the base view provide teacher signals through a stop-gradient operation, meaning that $(\hat{y}^{(\text{base})}, f^{(\text{base})})$ are treated as fixed targets. Based on these teacher signals, we enforce structural consistency across output probabilities, boundary responses, and decoder feature representations.

1) Pixel-level Output Consistency: To ensure basic prediction stability, we impose constraints in the probability space. Since augmented views undergo geometric transformations, we align the prediction $\hat{y}_a$ of the augmented view back to the original space via an inverse transformation, denoted as $\hat{y}_a^{\uparrow}$, before loss calculation. The output consistency loss is defined as the Mean Squared Error (MSE) between the aligned prediction and the teacher prediction:

$$\mathcal{L}_{\text{prob}} = \sum_{a\in\mathcal{A}} \left\| \hat{y}_a^{\uparrow} - \text{stopgrad}(\hat{y}_{\text{base}}) \right\|_2^2 \tag{4}$$

This constraint forces the model to maintain the alignment of pixel-level classification probability distributions under geometric perturbations, thereby suppressing prediction noise caused by domain shifts.

2) Boundary-aware Structural Consistency: Addressing the boundary blurring issue caused by traditional constraints, we introduce a differentiable edge extraction operator. Specifically, we use the Sobel operator $E(\cdot)$ to calculate the gradient magnitude maps of the aligned student prediction $\hat{y}_a^{\uparrow}$ and

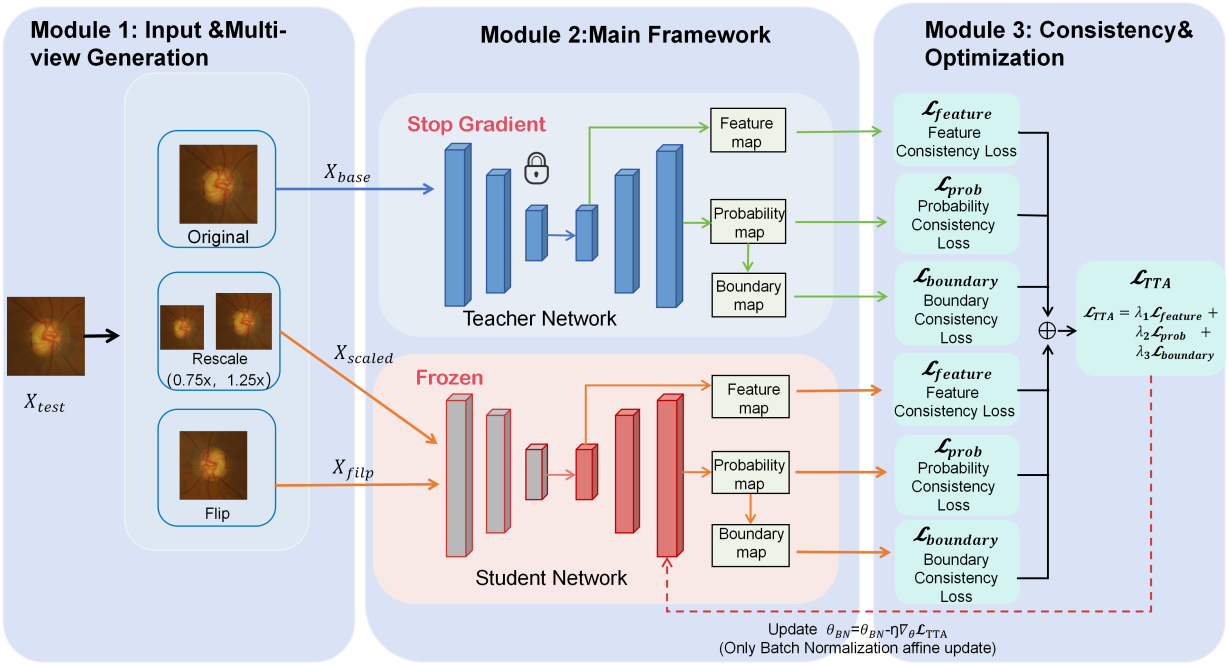


Fig. 2: **Overview of the proposed Structure-Aware TTA framework.** Given a test image, the framework generates diverse views (flip, multi-scale) fed into a Teacher-Student architecture. We enforce structural invariance via multi-level consistency constraints ($\mathcal{L}_{feature}, \mathcal{L}_{prob}, \mathcal{L}_{boundary}$) aligned between the frozen Teacher and the active Student. Adaptation is performed by minimizing $\mathcal{L}_{TTA}$ to exclusively update the affine parameters of specific Batch Normalization layers.

the teacher prediction $\hat{y}_{\text{base}}$, respectively. By constraining the consistency of gradient maps, we shift the optimization focus from pure pixel classification to structural edge fitting. The boundary consistency loss is defined as follows:

$$\mathcal{L}_{\text{boundary}} = \sum_{a \in \mathcal{A}} \left\| E(\hat{y}_a^{\uparrow}) - \text{stopgrad}(E(\hat{y}_{\text{base}})) \right\|_1 \quad (5)$$

Here, we utilize the L_1 norm because it induces sparsity more effectively than the L_2 norm, helping to sharpen boundary predictions, which is critical for the low-contrast and small Optic Cup structure.

3) Semantic Feature Consistency: Beyond the output space, the stability of deep semantic features is equally crucial. While shallow visual features are susceptible to domain shifts (e.g., illumination), high-level decoder features should encode domain-agnostic semantic information. Therefore, we align representations across different views in the feature space. To eliminate the influence of numerical scales, feature vectors are normalized via $\text{Norm}(\cdot)$ before distance calculation. For notational brevity, we denote this semantic feature consistency loss as $\mathcal{L}_{\mathcal{F}}$:

$$\mathcal{L}_{\mathcal{F}} = \sum_{a \in \mathcal{A}} \left\| \text{Norm}(f_a^{\uparrow}) - \text{Norm}(\text{stopgrad}(f_{\text{base}})) \right\|_2^2 \quad (6)$$

This encourages the model to extract more compact and discriminative feature representations.

Algorithm 1: Optimization Procedure of MVSCL

Initialize: Pre-trained parameters θ_0, active set $\mathcal{S}_{active}$ (BN affine weights in Decoder/Head);
for *each incoming test image* x_t **do**
  for *step* $n \leftarrow 1$ **to** N **do**
    Generate multi-view set $\mathcal{A}$ and teacher targets;
    Compute $\mathcal{L}_{TTA}$ (Prob + Boundary + Feature);
    Update: $\theta \leftarrow \theta - \eta \cdot \nabla_\theta \mathcal{L}_{TTA} \cdot \mathbb{I}(\theta \in \mathcal{S}_{active})$;
  end
  Output: Final prediction $\mathbb{P}_t = \text{Sigmoid}(f_\theta(x_t))$;
end

Finally, to achieve rapid adaptation while avoiding *Catastrophic Forgetting*, we implement a parameter-constrained update strategy. We strictly update only the affine parameters (i.e., γ, β) of the Batch Normalization layers. The final TTA optimization objective is a weighted sum of the three consistency losses:

$$\mathcal{L}_{\text{TTA}} = \lambda_{\text{prob}}\mathcal{L}_{\text{prob}} + \lambda_{\text{boundary}}\mathcal{L}_{\text{boundary}} + \lambda_{\mathcal{F}}\mathcal{L}_{\mathcal{F}} \quad (7)$$

where $\lambda_{\text{prob}}, \lambda_{\text{boundary}}, \lambda_{\mathcal{F}}$ are hyperparameters balancing the loss terms.

C. Parameter-Constrained Adaptation

To prevent semantic drift while enabling rapid adaptation, we implement a module-wise parameter update strategy. We partition the network parameters Θ into encoder Θ_{enc}, decoder

TABLE I: The values represent the Dice Similarity Coefficient (DSC, %). We compare our **MVSCL** with the baseline GRATA and other methods. The best results are **bolded**

Method	Domain A	Domain B	Domain C	Domain D	Domain 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 (Base)	76.63	78.84	76.58	67.10	73.13	74.45
MVSCL (Ours)	**77.60**	**79.64**	**82.34**	65.34	**73.70**	**75.72**

TABLE II: Component analysis of the proposed method. We report the average DSC (%) across five domains. **Bold** indicates the best results.

Method	RIM-ONE r3 *DSC*	REFUGE *DSC*	ORIGA *DSC*	REFUGE Valid *DSC*	Drishti-GS *DSC*	Average *DSC* ↑
GraTa	76.61	78.81	76.52	66.84	72.94	74.34
+ Structural Consistency	**77.85**	**79.92**	81.40	66.10	72.20	75.49
+ Affine-Only Update	77.10	79.20	78.60	**67.30**	73.40	75.12
Ours (Full)	77.60	79.64	**82.34**	65.34	**73.70**	**75.72**

Θ_{dec}, and head Θ_{head}. To preserve general visual representations captured during pre-training, we **freeze the encoder** (Θ_{enc}) entirely. Updates are strictly limited to the affine parameters (γ, β) of the Batch Normalization (BN) layers within the decoder and segmentation head ($\Theta_{\text{dec}} \cup \Theta_{\text{head}}$).

Formally, at the t-th test iteration, the parameters are updated as follows:

$$\theta_{t+1} = \theta_t - \eta \nabla_\theta \mathcal{L}_{\text{TTA}}(x_t) \cdot \mathbb{I}(\theta \in \mathcal{S}_{\text{active}}) \tag{8}$$

where η is the learning rate, $\mathbb{I}(\cdot)$ is the indicator function, and $\mathcal{S}_{\text{active}}$ denotes the set containing only the BN affine parameters of the decoder and segmentation head.

IV. Experiments

A. Datasets and Evaluation Metrics

To comprehensively evaluate the generalization capability and robustness of the model across data distributions from different medical centers, we conducted extensive experiments on the joint Optic Disc (OD) and Optic Cup (OC) segmentation task. We established an experimental benchmark using five representative public fundus image datasets, denoted as Domains A through E: RIM-ONE-r3 (Domain A) [28], REFUGE (Domain B) [29], ORIGA (Domain C) [30], REFUGE-Validation/Test (Domain D) [29], and Drishti-GS (Domain E) [31]. These datasets contain 159, 400, 650, 800, and 101 images, respectively.

In the data preprocessing phase, to eliminate irrelevant background noise and focus on key anatomical structures, we followed the standardized preprocessing pipeline of Chen *et al.* [10]. Specifically, a Region of Interest (ROI) of 800×800 pixels centered on the OD was cropped for each image. Subsequently, all ROIs were resized to a resolution of 512×512 and normalized using min-max normalization. For quantitative evaluation, we adopted the Dice Similarity Coefficient (DSC), a standard metric in medical segmentation, to measure the overlap between the predicted masks and the ground truth.

B. Implementation Details

In all experimental settings, consistent with Chen *et al.* [10] for fair comparison, we employed ResUNet-34 as the backbone network. We adopted a "single-source training, multi-target testing" cross-domain evaluation protocol. Specifically, a model trained on a specific domain (source) performs Test-time Adaptation (TTA) directly on the remaining four domains (target). We traversed all possible source-target combinations, resulting in 4×5 cross-domain scenarios, and reported the mean performance metrics.

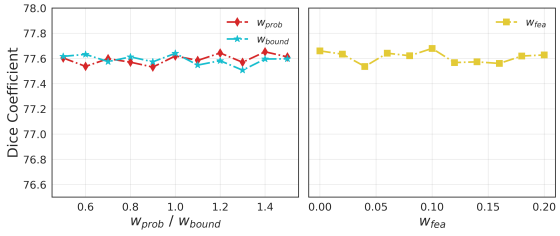


Fig. 4: Hyper-parameter sensitivity analysis using the control variable method. We assessed the Dice coefficient by varying w_{prob} and w_{bound} in $[0.5, 2.5]$ (step 0.2), and w_{fea} in $[0, 0.5]$ (step 0.05).

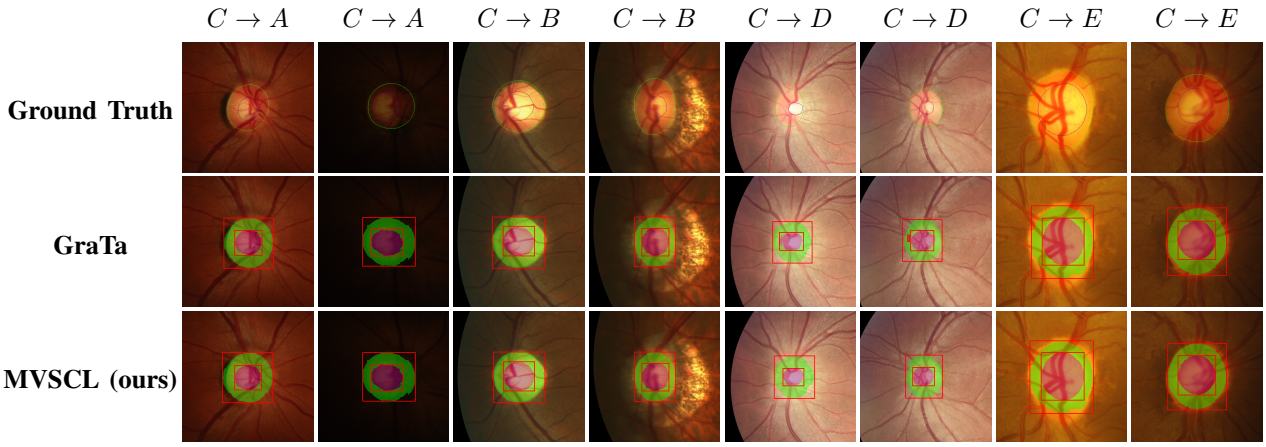


Fig. 3: **Visual comparison of segmentation results on unseen target domains.** The model is trained on the source domain ORIGA (Domain C). Rows from top to bottom represent: Ground Truth, the baseline GraTa, and our proposed MVSCL. The Optic Disc (OD) and Optic Cup (OC) are highlighted in green and purple, respectively. Red bounding boxes indicate the Ground Truth regions for facilitating boundary comparison.

Following chen*et al.*, we simulated batch 1 online processing with an adaptive iteration for each test sample. We use Adam β =0.0001 to update BN statistics on test data and optimize affine parameters.

C. Results

We compare MVSCL with GraTa and additional representative TTA baselines. Table I presents the quantitative comparison between MVSCL and the state-of-the-art TTA method, GraTa, across five cross-domain test subsets. In terms of overall performance, MVSCL demonstrates superior generalization capabilities, achieving a mean Dice coefficient of 75.72%. Notably, on the ORIGA dataset (Domain C), MVSCL yields breakthrough performance, significantly surpassing other methods by elevating the Dice score from 76.58% to 82.34%. This substantial margin underscores the necessity of incorporating structure-aware constraints. Compared with other methods, which primarily relies on pixel-level objectives, MVSCL regularizes anatomical stability through multi-view structural consistency, leading to more reliable segmentation under blurred boundaries and drastic intensity variations.

Ablation study. To evaluate the contributions of the proposed Multi-View Structural Consistency Learning (MVSCL) and Selective BN strategy, we conducted a series of ablation experiments, as reported in Table II. The results demonstrate that: (1) introducing MVSCL alone explicitly models structural invariance, improving the average DSC from 74.34% to 75.49%; (2) employing the Selective BN strategy limits parameter updates to the decoder, effectively mitigating semantic drift and achieving an average DSC of 75.12%; and (3) the optimal performance is achieved when both components are utilized jointly (i.e., Ours), reaching an average DSC of 75.72% and verifying the complementarity between structural constraints and parameter control.

Visualization Analysis: Fig. 3 illustrates the segmentation results transferred from Domain C (ORIGA) to four target domains. While the baseline GraTa suffers from severe boundary blurring and shape distortion, particularly under low contrast or vessel interference (e.g., $C \to B, E$), our MVSCL preserves the morphological integrity of the Optic Disc and Cup. The superior alignment with the Ground Truth validates MVSCL's efficacy in mitigating distribution shifts and alleviating over-smoothing issues in challenging anatomical regions.

Hyper-parameter Sensitivity Analysis: The sensitivity analysis on Domain A (Fig. 4) demonstrates hyper-parameter impacts. For w_{prob}, narrow fluctuations indicate robustness to weight variations. For w_{bound}, 1.0 achieves a peak, balancing geometric integrity without the instability observed at 1.3. Regarding w_{fea}, Dice peaks at 0.10; higher weights (> 0.10) reduce performance, likely due to over-regularization. Consequently, we adopted $w_{prob} = 1.0$, $w_{bound} = 1.0$, and $w_{fea} = 0.1$ for the optimal accuracy-stability trade-off.

V. Conclusion

In this paper, we propose **MVSCL**, a structure-aware and parameter-constrained test-time adaptation (TTA) framework for cross-domain fundus image segmentation. MVSCL enforces multi-view consistency on output probabilities, boundary responses, and decoder features under a teacher–student formulation, which helps preserve anatomical morphology under domain shifts. To mitigate semantic drift during unsupervised test-time updates, we restrict adaptation to the Batch Normalization affine parameters in the decoder and segmentation head. Extensive experiments across diverse cross-domain scenarios demonstrate that MVSCL improves robustness and achieves state-of-the-art performance on the evaluated benchmarks. Future work will extend MVSCL to other modalities and continual TTA under non-stationary shifts.

REFERENCES

[1] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image segmentation using deep learning: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 7, pp. 3523–3542, 2021.

[2] H. Guan and M. Liu, "Domain adaptation for medical image analysis: a survey," *IEEE Transactions on Biomedical Engineering*, vol. 69, no. 3, pp. 1173–1185, 2021.

[3] K. Zhou, Z. Liu, Y. Qiao, T. Xiang, and C. C. Loy, "Domain generalization: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 4, pp. 4396–4415, 2022.

[4] S. Schneider, E. Rusak, L. Eck, O. Bringmann, W. Brendel, and M. Bethge, "Improving robustness against common corruptions by covariate shift adaptation," *Advances in neural information processing systems*, vol. 33, pp. 11 539–11 551, 2020.

[5] H. Su, B. Wang, D. Liu, J. Li, C.-B. Feng, and C.-M. Vong, "Towards fully test-time adaptation via variance balancing and semantic augmentation," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.

[6] Y. Su, X. Xu, and K. Jia, "Towards real-world test-time adaptation: Tri-net self-training with balanced normalization," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 13, 2024, pp. 15 126–15 135.

[7] Y. Yuan, B. Xu, L. Hou, F. Sun, H. Shen, and X. Cheng, "Tea: Test-time energy adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 23 901–23 911.

[8] J. M. J. Valanarasu, P. Guo, V. M. Patel *et al.*, "On-the-fly test-time adaptation for medical image segmentation," in *Medical Imaging with Deep Learning*. PMLR, 2024, pp. 586–598.

[9] 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 *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 464–474.

[10] Z. Chen, Y. Ye, Y. Pan, and Y. Xia, "Gradient alignment improves test-time adaptation for medical image segmentation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 3, 2025, pp. 2429–2437.

[11] H. Fu, J. Cheng, Y. Xu, D. W. K. Wong, J. Liu, and X. Cao, "Joint optic disc and cup segmentation based on multi-label deep network and polar transformation," *IEEE transactions on medical imaging*, vol. 37, no. 7, pp. 1597–1605, 2018.

[12] S. Yuheng and Y. Hao, "Image segmentation algorithms overview," *arXiv preprint arXiv:1707.02051*, 2017.

[13] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[14] J. Dai, Y. Li, K. He, and J. Sun, "R-fcn: Object detection via region-based fully convolutional networks," *Advances in neural information processing systems*, vol. 29, 2016.

[15] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, 2015.

[16] B. Li, J. Yan, W. Wu, Z. Zhu, and X. Hu, "High performance visual tracking with siamese region proposal network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8971–8980.

[17] A. Sherstinsky, "Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network," *Physica D: Nonlinear Phenomena*, vol. 404, p. 132306, 2020.

[18] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille, "Deep captioning with multimodal recurrent neural networks (m-rnn)," *arXiv preprint arXiv:1412.6632*, 2014.

[19] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "Transunet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021.

[20] J. Ma, F. Li, and B. Wang, "U-mamba: Enhancing long-range dependency for biomedical image segmentation," *arXiv preprint arXiv:2401.04722*, 2024.

[21] J. Liu, H. Yang, H.-Y. Zhou, Y. Xi, L. Yu, C. Li, Y. Liang, G. Shi, Y. Yu, S. Zhang *et al.*, "Swin-umamba: Mamba-based unet with imagenet-based pretraining," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2024, pp. 615–625.

[22] Y. Sun, X. Wang, Z. Liu, J. Miller, A. Efros, and M. Hardt, "Test-time training with self-supervision for generalization under distribution shifts," in *International conference on machine learning*. PMLR, 2020, pp. 9229–9248.

[23] D. Wang, E. Shelhamer, S. Liu, B. Olshausen, and T. Darrell, "Tent: Fully test-time adaptation by entropy minimization," *arXiv preprint arXiv:2006.10726*, 2020.

[24] A. Karmanov, D. Guan, S. Lu, A. El Saddik, and E. Xing, "Efficient test-time adaptation of vision-language models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 14 162–14 171.

[25] R. A. Marsden, M. Döbler, and B. Yang, "Universal test-time adaptation through weight ensembling, diversity weighting, and prior correction," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 2555–2565.

[26] J. Zhu, B. Bolsterlee, Y. Song, and E. Meijering, "Improving cross-domain generalizability of medical image segmentation using uncertainty and shape-aware continual test-time domain adaptation," *Medical Image Analysis*, vol. 101, p. 103422, 2025.

[27] H. Li, M. Ou, H. Li, Z. Qiu, K. Niu, H. Fu, and J. Liu, "Multi-view test-time adaptation for semantic segmentation in clinical cataract surgery," *IEEE Transactions on Medical Imaging*, 2025.

[28] F. Fumero, S. Alayón, J. L. Sanchez, J. 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)*. IEEE, 2011, pp. 1–6.

[29] J. I. Orlando, H. Fu, J. B. Breda, K. Van Keer, D. R. Bathula, A. Diaz-Pinto, R. Fang, P.-A. Heng, J. Kim, J. Lee *et al.*, "Refuge challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs," *Medical image analysis*, vol. 59, p. 101570, 2020.

[30] Z. Zhang, F. S. Yin, J. Liu, W. K. Wong, N. M. Tan, B. H. Lee, J. Cheng, and T. Y. 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*. IEEE, 2010, pp. 3065–3068.

[31] J. Sivaswamy, S. Krishnadas, G. D. Joshi, M. Jain, and A. U. S. Tabish, "Drishti-gs: Retinal image dataset for optic nerve head (onh) segmentation," in *2014 IEEE 11th international symposium on biomedical imaging (ISBI)*. IEEE, 2014, pp. 53–56.