

TANGO: Learning Distribution-wise Foundation Prior Consistency and Instance-wise Style Calibration for Medical Image Generalization

Chuang Liu¹ Yichao Cao² Xiu Su² * Haogang Zhu^{1,3} *

¹State Key Laboratory of Complex & Critical Software Environment, Beihang University

²Big Data Institute, Central South University

³Hangzhou International Innovation Institute, Beihang University

cliu_trans@buaa.edu.cn, caoyichao@csu.edu.cn, xiusu@csu.edu.cn,
 haogangzhu@buaa.edu.cn

Abstract

Test-time adaptation (TTA) has emerged as a promising solution to address real world domain shifts in medical image segmentation. Current approaches adapt by updating or regularizing a pre-trained source model. However, they face two major issues: (i) the source models on which they rely are prone to overfitting under domain shifts; (ii) in dynamic continual testing scenarios, error accumulation and class forgetting are further exacerbated. To overcome these limitations, we propose **TanGo**, a novel framework that combines **Training to adapt with Foundation Guidance and continual style calibration**. During training, TanGo learns generalization priors from vision foundation models (VFMs) through distribution-wise consistency learning. We incorporate stable low-frequency representations from a frozen encoder of VFMs as priors to guide the source model, constraining its output feature distribution to yield a more generalizable feature space. At test time, we propose an instance-wise style calibration method that uses a learnable data decorator to shift dynamic test images back toward an enhanced source-like distribution. Subsequently, a series of constraints is applied to draw the decorated test samples closer to the enhanced source distribution while maintaining their semantic integrity, thus boosting the continual adaptability of the source model. Extensive experiments on multiple medical image segmentation tasks demonstrate that TanGo achieves state-of-the-art performance.

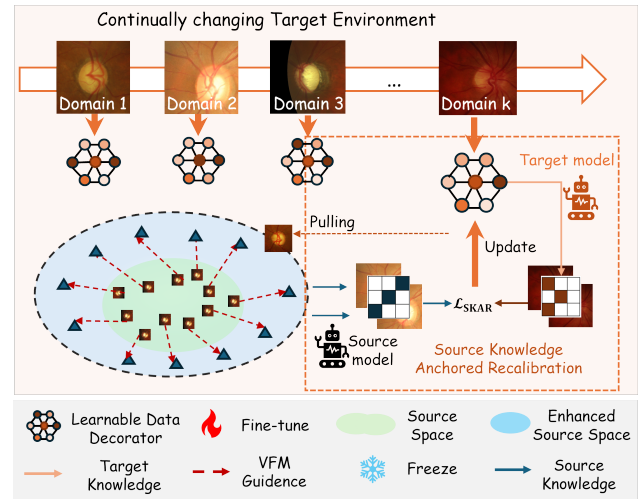


Figure 1. Our Training-to-Adapt CTTA paradigm. During training, we learn generalization priors consistency from a frozen foundation model to enhance the source model’s feature robustness. At test time, a learnable per-sample data decorator, together with Source-Knowledge Anchored Recalibration, continuously pulls the evolving test distribution toward the enhanced source feature space, enabling stable and effective online adaptation.

1. Introduction

Medical image segmentation is a cornerstone of computer-aided diagnosis, enabling clinicians to localize lesions and organs with high precision [34, 43, 64]. However, scanner, protocol, and site heterogeneity induce domain shifts that degrade the performance of segmentation models trained on a source domain when they are deployed on unseen target domains [4, 44, 63]. Domain generalization methods attempt to address this by learning domain-invariant representations [62] or by simulating unknown target domains via data [30] or feature [5, 28] perturbations. Yet, given the limited diversity of available source samples, these approaches struggle to cover the wide and continuously evolving distribution shifts

*Corresponding author. This work was supported in part by the National Science and Technology Major Project(2025ZD0123401), in part by the National Natural Science Foundation of China under Grant 62406014, in part by the Beijing Natural Science Foundation (IS24081,7262079), in part by the Zhejiang Natural Science Foundation(LZ26F020012) and in part by the Start-up Funds of Hangzhou International Innovation Institute of Beihang University under Grant No.2024KQ045 and No.2024KQ027.

encountered in the real world. To tackle this limitation, Test-Time Adaptation (TTA) [38, 53] and Continual Test-Time Adaptation (CTTA) [53, 54] have been proposed. These methods adapt parts of the source model online at test time using unlabeled target samples, aiming to counteract real-world domain drifts. They improve adaptation by designing better loss functions [36, 50, 54] and introducing regularizers to prevent excessive parameter updates. For example, DomainAdaptor [59] mixes statistical parameters from source and target models to mitigate shifts in normalization layers, and VPTTA [6] aligns BN affine parameters between the source and target models extracted from augmented samples to optimize visual prompts and alleviate domain shift.

However, most existing TTA and CTTA methods rest on two implicit yet impractical assumptions: (i) *the source model possesses strong generalization priors*, and (ii) *the target distribution remains relatively stable during testing*. In practice, first, the source model often overfits the limited labeled source data (see the t-SNE visualization in Fig. 3). This overfitting undermines the effectiveness of regularization strategies [6, 54, 59] based on source-domain predictions and can misguide the optimization of the target model. Secondly, as shown in the first and third rows of Fig. 4, domain shifts between a single source and multiple target domains are often substantial and continuously evolving. Such dynamics further amplify the risks of adaptation from an initialization of an overfitted source model, leading to pseudo-label error accumulation and category forgetting, which ultimately degrades continual adaptation performance.

To address both these limitations, we propose a unified paradigm named **TanGo** (Training to adapt with Foundation Guidance and Continual Style Calibration), as illustrated in Fig. 1. Specifically, in the training phase, we leverage the strong image content understanding and stable feature representation of vision foundation models [13, 26, 45] (e.g., SAM [25] and DINOv2 [40]). By introducing distribution-level consistency learning, we incorporate the stable low-frequency representations extracted from the frozen encoder of a foundational model as priors. These priors guide the source model by constraining its feature distribution at the output stage, leading to a more robust source-domain feature space. During testing, we introduced the *Instance-Wise Style Adaptive Calibration* method. Firstly, a lightweight instance-level decorator is adapted to transform each test image toward an enhanced source-domain space. Then, we employ the *Source Knowledge Anchored Recalibration* (SKAR) strategy, which optimizes adaptation through three objectives: (1) *Source-Anchored Distributional Loss*, which enforces global style consistency between the intermediate features of the adaptive and enhanced source models, driving test samples toward the enriched source-domain space; (2) *Source-Anchored Semantic Loss*, which constrains the segmentation outputs of the source and target models to pre-

serve the semantic integrity of augmented samples; and (3) *Entropy Minimization Loss*, which stabilizes the optimization of learnable batch normalization parameters in the target model, reduces prediction uncertainty, and accelerates convergence. Thus, our method establishes a unified adaptation framework that jointly addresses style alignment, semantic preservation, and uncertainty reduction. As a result, test samples with diverse styles are dynamically calibrated toward the enhanced source-domain space, enabling the target model to adapt effectively to continuously changing testing environments. The contributions are summarized as follows:

- We propose **TanGo**, a unified framework that bridges training and testing through collaborative continual test-time adaptation. It transfers the generalization learned during training to style calibration at test time, effectively mitigating domain shifts and enhancing cross-domain robustness.
- During training, we introduce distribution-wise consistency learning, which injects strong generalization priors from vision foundation models into the source model to create a more resilient source feature space.
- We develop an instance-wise style adaptive calibration method that jointly optimizes distributional, semantic, and uncertainty objectives, ensuring stable adaptation and semantic consistency under continuously testing conditions.
- Extensive experiments on multiple medical image segmentation benchmarks under both CTTA and mixed CTTA settings demonstrate the superior generalization and continual adaptation capabilities of TanGo.

2. Related Work

2.1. Generalizable Medical Image Segmentation

Single-source domain generalization for medical image segmentation is challenging, as it must generalize from a single labeled source to unseen target domains that violate the i.i.d. assumption [4, 52] under limited supervision. To address these challenges, CSDG [42] enhances source diversity at the input level through causal intervention-based image augmentation, while TRID [5] perturbs feature-level statistics to expand the effective source distribution. CCSDG [18] learns domain-invariant representations via channel-decoupled contrastive learning. With the emergence of large-scale pretraining vision foundation models (VFMs) [65] such as DINO [40] and SAM [25], remarkable zero-shot generalization [29, 33] has been demonstrated. Furthermore, DAPSAM [56] adapts SAM to medical image segmentation through prototype-driven prompt generation without manual input. FADA [3] employs wavelet-based frequency adaptation to exploit the strong generalization priors of vision foundation models. However, these fine-tuning-based approaches often entail substantial computational overhead during inference and overlook the test-time adaptability of the source-domain model itself. In contrast, our method learns the founda-

tional priors consistency between large-scale VFMs with the lightweight source model (e.g., UNet), enabling it to dynamically adapt to unseen samples at test time.

2.2. Test-Time Adaptation

Test-time adaptation (TTA) aims to exploit the distributional cues of unlabeled test samples by updating part of the network to rapidly adapt a pretrained source model [27, 51, 54]. Recent methods achieve this through self-supervised optimization [16, 36, 53], normalization-layer alignment [6, 59], or adaptive input transformations [22, 58]. Tent [53] fine-tunes affine parameters of normalization layers via entropy minimization, while TASD [31] integrates dictionary learning and test-time regularization to capture domain-invariant shape priors. CTTA [54] extends TTA to dynamic and continuously changing environments, addressing challenges such as error accumulation and category forgetting during long-term inference. Some approaches [37, 54, 57] alleviate these issues by minimizing entropy losses or by combining data augmentation with model regularization. DLTTA [57] introduces a dynamic learning-rate strategy that assigns adaptive update rates to different test distributions, and [59] proposes adaptive fusion of source and test statistics to generalize entropy minimization. VPTTA [6] employs prompt learning to generate persistent low-frequency prompts for each test sample and refines them through statistical alignment to mitigate error accumulation. However, these methods rely on source models with limited out-of-domain generalization, making robust adaptation challenging. To overcome this, we embed strong generalization priors into the source model during training and leverage them as anchors at test time to jointly regularize the feature and output spaces, thereby ensuring stable test-time adaptation.

3. Methodology

3.1. Problem Formulation and Method Overview

Let the labeled source dataset be $\mathcal{S} = \{(\mathbf{x}_i^s, \mathbf{y}_i^s)\}_{i=1}^{n_s}$ and the incoming unlabeled target stream be $\mathcal{T} = \{\mathbf{x}_j^t\}_{j=1}^{n_t}$. Each input image $\mathbf{x} \in \mathbb{R}^{C \times H \times W}$ has a corresponding segmentation label $\mathbf{y} \in \{1, \dots, K\}^{H \times W}$, where K denotes the number of semantic classes. The joint distributions of the source and target differ, i.e., $P_S(\mathbf{x}, \mathbf{y}) \neq P_T(\mathbf{x}, \mathbf{y})$. Let $f_S : \mathcal{X} \rightarrow \mathcal{Y}$ denote a segmentation source model trained on $\mathcal{S}$, where $\mathcal{X} = \mathbb{R}^{C \times H \times W}$ and $\mathcal{Y} = \{1, \dots, K\}^{H \times W}$. *Continual test-time adaptation (CTTA)* aims to exploit the labeled knowledge encoded in f_S to generate reliable predictions for the evolving target stream under domain shift. In this setting, a target model f_T , initialized from f_S , is continuously optimized *online and progressively* during inference to maintain robustness against dynamically changing target distributions. To achieve this, as shown in Fig. 2, our **TanGo** framework follows a two-stage paradigm. During

training, the proposed *Distribution-wise Consistency Learning* distills frequency-domain generalization priors from the foundation model into the source model to form a unified, domain-harmonized feature space $\mathcal{H}_U$. At test time, the *Instance-wise Style Adaptive Calibration* mechanism adaptively aligns the distribution $\mathcal{H}_T^i$ of incoming target samples with the source knowledge-anchored space $\mathcal{H}_U$, ensuring efficient continual adaptation during inference.

3.2. Distribution-wise Consistency Learning

In this section, we leverage frequency domain priors as a bridge to transfer generalization capabilities from foundation to specialist models. Motivated by the fact that low-frequency components encode domain-specific styles while high-frequency signals retain anatomical structures [19, 23], we propose a *Distribution-wise Consistency Learning* method. It employs a two-branch design: a fused branch, formed by injecting low-frequency priors from the foundation model f_V into the specialist model f_S , guides the original source branch via consistency learning. This process cultivates a stable and broadly generalizable representation space $\mathcal{H}_U$, ensuring robust performance under continuously shifting test-time domains.

3.2.1. Frequency Aware Foundation Prior Injection

In this section, to adapt the Vision Foundation Model (VFM) to the medical domain, we first integrate Adapters [56] for parameter-efficient fine-tuning. It is composed of two fully-connected layers forming a bottleneck structure with a GELU activation in between. Details of Adapters are deferred to the Appendix. A. Subsequently, a *Frequency Aware Foundation Prior Injection* module (FAFPI) is introduced to inject generalizable priors from a frozen vision foundation model f_V into the specialist model f_S . Specifically, given the i -th intermediate feature map from the foundation encoder $F_i^V \in \mathbb{R}^{C_v \times H_v \times W_v}$ and the corresponding feature map from the specialist encoder $F_i^S \in \mathbb{R}^{C_s \times H_s \times W_s}$, We introduce the lightweight *Feature Refinement Module* Θ (FRM), composed of depthwise separable 3×3 convolutions, activation, and batch normalization layers, to obtain the spatially aligned feature $\tilde{F}_i^V = \Theta(F_i^V)$ that matches F_i^S in resolution and facilitates effective component-wise frequency injection. Secondly, we apply a 2D FFT [9] to obtain the corresponding frequency representations $\mathcal{F}_i^S$ and $\mathcal{F}_i^S$. As shown in the following equations, these frequency representations are decomposed into amplitude components:

$$[\mathcal{F}_i^{V,A}, \mathcal{F}_i^{S,A}] = \mathcal{O}\mathcal{A}([\mathcal{F}_i^V, \mathcal{F}_i^S]), \quad (1)$$

where $\mathcal{O}\mathcal{A}$ represents the frequency decomposition operation. To separate low- and high-frequency components of both $\mathcal{F}_i^{V,A}$ and $\mathcal{F}_i^{S,A}$, we introduce a binary mask $\mathcal{M} \in \mathbb{R}^{H_s \times W_s}$, as shown in Eq. (2), where the central region is

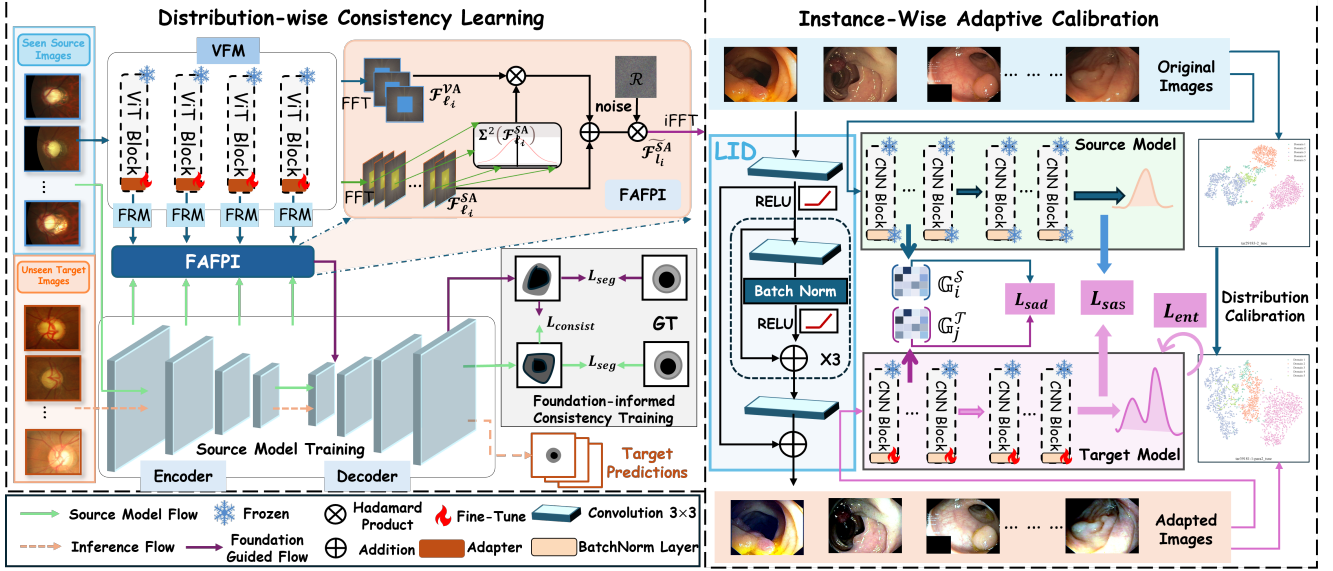


Figure 2. Overview of our TanGo framework. During training, the proposed *Distribution-wise Style Consistency Learning* transfers frequency-domain generalization priors from the foundation model into the source model to form a unified, domain-harmonized feature space $\mathcal{H}_U$. At test time, the *Instance-wise Adaptive Calibration* mechanism adaptively aligns the distribution $\mathcal{H}_T$ of incoming target samples with the source knowledge-anchored space $\mathcal{H}_U$, ensuring efficient continual adaptation during inference.

controlled by a hyperparameter r , which is set to 0.4.

$$\mathcal{M} = \Pi \left[\max \left(\left| u - \frac{H}{2} \right|, \left| v - \frac{W}{2} \right| \right) \leq \frac{r \cdot \min(H, W)}{2} \right], \quad (2)$$

where $\Pi(\cdot)$ is the indicator function that equals 1 if the condition is satisfied and 0 otherwise. The low- and high-pass filtered amplitudes are obtained as

$$\mathcal{F}_i^{SA} = \mathcal{M} \odot \mathcal{F}_i^{SA}, \quad \mathcal{F}_i^{VA} = \mathcal{M} \odot \mathcal{F}_i^{VA}, \quad (3)$$

where $\odot$ denotes element-wise multiplication. Finally, to incorporate the foundation prior into the specialist model, we integrate their low-frequency distributions—denoted as $\mathcal{F}_i^{SA}$ and $\mathcal{F}_i^{VA}$. Specifically, we compute the variance as: $\Sigma^2(\mathcal{F}_i^{SA}) = \frac{1}{B} \sum_{b=1}^B [\mathcal{F}_i^{SA} - \mathbb{E}(\mathcal{F}_i^{SA})]^2$, which quantifies the amplitude variation of each frequency component and thereby regulates the injection strength of the foundational low-frequency signal $\mathcal{F}_i^{VA}$. The injection of the low-frequency prior is thus formulated as:

$$\tilde{\mathcal{F}}_i^{SA} = \mathcal{F}_i^{SA} + \Sigma(\mathcal{F}_i^{SA}) \odot \mathcal{F}_i^{VA} \odot \mathcal{R}, \quad \mathcal{R} \sim \mathcal{U}(0, 1). \quad (4)$$

Here, $\mathcal{R} \in \mathbb{R}^{C \times H \times W}$ denotes random noise sampled from a uniform distribution $\mathcal{U}(0, 1)$ to perturb the fused amplitudes, encouraging f_S to focus on the structural cues of phase components for robustness. Based on the injection operation in Eq. (4), an inverse Fourier transform is subsequently applied to reconstruct the fused feature $\tilde{\mathcal{F}}_i^S$, which integrates the low-frequency priors from the foundation model f_V , enabling the source model f_S to emulate the latent distribution shifts encountered during test-time adaptation.

3.2.2. Foundation-informed Consistency Training

After integrating the low-frequency priors from the foundation model, the source model f_S produces two prediction branches during training: the original source branch $f_S(\mathcal{F}_i^S)$ and the foundation-informed branch $f_S(\tilde{\mathcal{F}}_i^S)$. To enforce consistency between these two branches, we define the foundation-informed consistency loss as:

$$\mathcal{L}_{FCL} = \frac{1}{2N} \sum_{i=1}^N [\text{KL}(\tilde{q}_i^S \| q_i^S) + \text{KL}(q_i^S \| \tilde{q}_i^S)], \quad (5)$$

where $\tilde{q}_i^S$ and q_i^S denote the softened prediction distributions of the foundation-informed and original branches, respectively. This bidirectional consistency learning compels the source model to produce consistent segmentation predictions from $f_S(\tilde{\mathcal{F}}_i^S)$ and $f_S(\mathcal{F}_i^S)$, thereby transferring the generalizable priors of the foundation model f_V into the source model f_S . By aligning the distributional behavior across both branches, this process strengthens the source model's representation space $\mathcal{H}_U$ and enables it to better accommodate the style variations encountered in continuously changing test environments.

Training Objective: The overall training objective integrates the supervised segmentation losses from both branches together with the proposed consistency regularization. The segmentation loss $\mathcal{L}_{\text{seg}}$ for each branch comprises a combination of Dice loss and Cross-Entropy loss. Accordingly, the total training loss is formulated as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{seg}}(\tilde{q}_i^S, y) + \mathcal{L}_{\text{seg}}(q_i^S, y) + \lambda_1 \mathcal{L}_{FCL}, \quad (6)$$

where the coefficients λ_1 is empirically set to 0.5. This joint optimization enables the source model f_S to absorb the generalizable priors from the foundation model while maintaining strong discriminative segmentation ability.

3.3. Instance-wise Style Adaptive Calibration

3.3.1. Learnable Instance Aware Decorator

To enhance the model’s continual adaptability in test time, we introduce a *Learnable Instance Aware Decorator* (LID), denoted as $f_D(\cdot; \phi)$. Unlike conventional visual prompt learning methods [15, 22] that adopt a fixed prompt shared across all test samples and thus overlook intra-domain variability, LID dynamically generates an *instance-specific* decoration for each incoming test image $\mathbf{x}_i^t$ as shown in Fig. 4, adapting its style distribution $\mathcal{H}_i^T$ to better align with the feature distribution $\mathcal{H}^U$ learned by the source model f_S . Concretely, for a forthcoming test sample $\mathbf{x}_i^t$, the decorator outputs an augmented counterpart as follows:

$$\tilde{\mathbf{x}}_i^t = \mathbf{x}_i^t + f_d(\mathbf{x}_i^t; \phi), \quad (7)$$

where ϕ are the lightweight decorator parameters. Specifically, LID is designed as a lightweight, learnable network composed of an input convolutional layer, a sequence of enhancement blocks, and an output convolutional layer. Each enhancement block consists of compact convolutional operations interleaved with nonlinear activations, while residual connections are introduced between blocks to preserve semantic integrity throughout the enhancement process.

3.3.2. Source Knowledge Anchored Recalibration

Despite the flexibility of the LID module, without proper constraints, the LID module could generate fictitious test samples deviating from the source feature manifold $\mathcal{H}^S$. To mitigate this, we propose *Source Knowledge Anchored Recalibration* (SKAR) method, which anchors the adapted test samples to the source space through two complementary objectives: a *Source-Anchored Distributional Loss* $\mathcal{L}_{sad}$ that enforces global distribution alignment and a *Source-Anchored Semantic Loss* $\mathcal{L}_{sas}$ that preserves semantic consistency. They serve two main purposes: (1) $\mathcal{L}_{sad}$ aligns the features of LID-transformed test samples $\tilde{\mathbf{x}}_i^t$ with their original counterparts at all encoder layers of f_T and f_S , reducing the distribution gap between source and target domains; (2) $\mathcal{L}_{sas}$ enforces the global semantic consistency in segmentation outputs before and after transformation, ensuring that the LID module generates semantically reliable test samples.

During adaptation, we first update the target model f_T using a self-supervised entropy loss $\mathcal{L}_{ent}$ applied to the prediction $\tilde{P}_i^T$ of LID-transformed test samples $\tilde{\mathbf{x}}_i^t$ as:

$$\mathcal{L}_{ent}(\theta; \tilde{\mathbf{x}}_i^t) = -\tilde{P}_i^T \log \tilde{P}_i^T, \quad \tilde{P}_i^T = \text{Sigmoid}(f_T(\tilde{\mathbf{x}}_i^t)) \quad (8)$$

Here, θ denotes all learnable parameters in f_T , including the LID parameters as well as the transformation parameters

γ and β of the batch normalization layers. Secondly, *the Gram matrix*, widely adopted in style transfer [17] for capturing channel-wise correlations that characterize feature style, motivates us to align the intermediate Gram representations between the adapted and source domains through the style alignment loss $\mathcal{L}_{sas}$. Specifically, let $\mathbf{F}_{ij}^T \in \mathbb{R}^{C \times HW}$ and $\mathbf{F}_{ij}^S \in \mathbb{R}^{C \times HW}$ denote the feature maps of the j -th block from $f_T(\tilde{\mathbf{x}}_i^t)$ and $f_S(\mathbf{x}_i^t)$ for the i -th sample (with the index i omitted in later formulas for brevity). We first obtain their Gram matrices of as follows:

$$\mathbb{G}_j^T = \mathbf{F}_j^T \mathbf{F}_j^{T\top}, \quad \mathbb{G}_j^S = \mathbf{F}_j^S \mathbf{F}_j^{S\top} \quad (9)$$

where the $\mathbb{G}_j^T, \mathbb{G}_j^S \in \mathbb{R}^{C \times C}$. We then apply row-wise $\mathcal{L}_2$ normalization to focus on relational structures:

$$\tilde{\mathbb{G}}_j^T = \frac{\mathbb{G}_j^{T[i,:]} }{\|\mathbb{G}_j^{T[i,:]} \|_2}, \quad \tilde{\mathbb{G}}_j^S = \frac{\mathbb{G}_j^{S[i,:]} }{\|\mathbb{G}_j^{S[i,:]} \|_2} \quad (10)$$

Subsequently, the distributional loss $\mathcal{L}_{sad}$ is formulated as the mean squared error between the normalized Gram matrices of the features of the target and source model:

$$\mathcal{L}_{sad} = \frac{1}{C} \left\| \tilde{\mathbb{G}}_j^T - \tilde{\mathbb{G}}_j^S \right\|_2^2, \quad (11)$$

where C represents the channels of feature maps. Additionally, our source anchored semantic loss $\mathcal{L}_{sas}$ aligns the output logits of f_T and f_S using the mean squared error as:

$$\mathcal{L}_{sas} = \frac{1}{B} \sum_{i=1}^B \|f_T(\tilde{\mathbf{x}}_i^t) - f_S(\mathbf{x}_i^t)\|_2^2, \quad (12)$$

where B is the batch size. This objective in Eq. (12) ensures the semantic consistency between the adapted and original samples at the prediction level. Therefore, the overall test-time objective of *Tango* framework is formulated as:

$$\mathcal{L}_{SKAR} = \lambda_2 \cdot \mathcal{L}_{sad} + \lambda_3 \cdot \mathcal{L}_{sas} + \mathcal{L}_{ent} \quad (13)$$

Empirically, we set the coefficients as $\lambda_2 = 0.5$, and $\lambda_3 = 0.6$ based on experimental validation in Appendix. C.4.

4. Experiments

4.1. Datasets and Evaluation Metrics

We evaluate TanGo on multiple medical segmentation benchmarks under TTA, CTTA, and mixed-CTTA settings to validate its effectiveness. **The fundus (OD/OC) segmentation task:** The first benchmark is constructed from five public datasets collected from different medical centers, denoted as domain A (RIM-ONE-r3 [14]), B (REFUGE [41]), C (ORIGA [60]), D (REFUGE-Validation/Test [41]), and E (Drishti-GS [49]), containing 159, 400, 650, 800, and 101 images, respectively. Following [30], we cropped a region of interest (ROI) of size 800×800 centered on the

Table 1. Comparison with ‘No Adapt’ and nine competing methods under the CTTA setting on the OD/OC segmentation task. Top three results are highlighted as **best**, **second** and **third**, respectively.

Methods	Domain A	Domain B	Domain C	Domain D	Domain E	Average $\uparrow$
	<i>DSC</i>	<i>DSC</i>	<i>DSC</i>	<i>DSC</i>	<i>DSC</i>	<i>DSC</i>
No Adapt [ResUNet-34]	64.53	76.06	71.18	52.67	64.87	65.86
TENT [53] (ICLR 2021)	73.07	78.66	71.94	46.81	70.20	68.13
CoTTA [54] (CVPR 2022)	75.39	75.98	69.14	53.99	70.40	68.98
DLTTA [57] (TMI 2022)	75.11	78.85	73.89	51.64	69.71	69.84
DUA [35] (CVPR 2022)	72.28	76.59	70.13	56.17	71.38	69.31
SAR [39] (ICLR 2023)	74.55	77.71	70.78	55.40	71.72	70.03
DomainAdaptor [59] (CVPR 2023)	74.50	76.39	71.81	56.78	70.55	70.01
VPTTA [6] (CVPR 2024)	73.91	79.36	74.51	56.51	75.35	71.93
CertainTTA [10] (IF 2025)	74.06	79.63	73.78	56.47	75.28	71.84
GraTa [7] (AAAI 2025)	76.58	78.72	76.27	67.15	72.88	74.32
TanGo (Ours)	87.76	85.96	84.49	84.41	83.52	85.23

Table 2. Comparison with ‘No Adapt’ and eight competing methods under the CTTA setting on the polyp segmentation task. Top three results are highlighted as **best**, **second** and **third**, respectively.

Methods	Domain A			Domain B			Domain C			Domain D			Average $\uparrow$		
	<i>DSC</i>	$E_{\phi}^{\max}$	S_{α}	<i>DSC</i>	$E_{\phi}^{\max}$	S_{α}	<i>DSC</i>	$E_{\phi}^{\max}$	S_{α}	<i>DSC</i>	$E_{\phi}^{\max}$	S_{α}	<i>DSC</i>	$E_{\phi}^{\max}$	S_{α}
No Adapt (PraNet)	79.90	87.97	84.66	66.33	78.51	76.72	73.89	84.64	81.28	82.95	90.84	88.08	75.77	85.49	82.69
TENT [53] (ICLR 2021)	74.86	84.58	80.52	67.51	78.66	78.05	17.79	40.04	53.30	73.55	83.38	82.41	58.43	71.67	73.57
CoTTA [54] (CVPR 2022)	76.46	85.37	82.56	66.77	76.75	79.17	71.39	83.42	80.18	70.71	79.81	82.54	71.33	81.34	81.11
DLTTA [57] (TMI 2022)	76.27	85.23	82.41	66.58	77.00	79.24	63.72	78.23	75.56	71.20	81.32	83.47	69.44	80.45	80.17
DUA [35] (CVPR 2022)	78.93	87.37	83.96	66.84	78.52	77.51	76.53	86.45	83.05	86.24	93.23	89.82	77.13	86.39	83.58
SAR [39] (ICLR 2023)	76.48	85.89	81.49	66.45	77.35	78.05	71.46	83.23	79.40	70.41	80.11	81.07	71.20	81.65	80.00
DomainAdaptor [59] (CVPR 2023)	77.48	86.31	82.40	70.82	81.76	80.88	71.96	83.06	79.97	76.89	85.89	84.45	74.29	84.26	81.93
VPTTA [6] (CVPR 2024)	81.00	88.91	84.91	76.87	87.31	84.08	77.58	87.48	83.64	86.39	93.47	89.87	80.46	89.29	85.62
CertainTTA [10] (IF 2025)	81.55	88.97	84.96	75.73	86.86	83.46	73.94	86.76	82.92	87.11	93.88	90.27	79.58	89.12	85.40
TanGo (Ours)	83.95	91.33	86.55	82.44	91.45	86.42	81.11	89.81	85.76	87.29	94.80	89.97	83.70	91.85	87.14

optic disk in each image, which was subsequently resized to 256×256 to reduce computational overhead. All images were normalized using min–max scaling. The Dice similarity coefficient (DSC) is employed as the evaluation metric for this task. **The polyp segmentation task:** The second benchmark consists of four widely used datasets, denoted as domain A (BKAI-IGH-NeoPolyp [2]), B (CVCCLinDB [1]), C (ETIS-LaribPolypDB [47]), and D (Kvasir-Seg [12]), comprising 1000, 612, 196, and 1000 images, respectively. Following the preprocessing in [12], each image was resized to 352×352 and normalized with the mean and variance computed from ImageNet. For evaluation, we adopt three complementary metrics: DSC, the enhanced-alignment measure ($E_{\phi}^{\max}$) [11], and the structure-aware similarity measure (S_{α}) [8]. **Implementation Details:** For all experiments, we train the source model f_S on a single source domain and directly evaluate it on the remaining target domains. Following VPTTA [6], f_S is implemented as a ResUNet-34 for the OC/OD segmentation task and as a PraNet with Res2Net backbone for the polyp segmentation task. During training, the ViT-B SAM encoder is used as the foundation model f_V and fine-tuned with adapters [56]. For comprehensive implementation details, please refer to Appendix. B.

4.2. Experiment Results

We compare TanGo with the ‘No Adapt’ (training only and testing without adaptation) baseline and nine representative methods under TTA and CTTA settings, including entropy-

based approaches (TENT-Continual [53], SAR [39]), pseudo-labeling and uncertainty-driven methods (CoTTA [54], CertainTTA [10]), BN-statistics simulation methods (DUA [35], DomainAdaptor [59]), a dynamic learning rate strategy (DLTTA [57]), a visual prompt learning based method (VPTTA [6]), a single-domain generalization method (MGMDG [32]), and a gradient-alignment method (GraTa [7]), which aligns pseudo and auxiliary gradients with a dynamic learning rate. All methods are re-implemented with the same backbone as VPTTA for a fair and consistent comparison. Quantitative segmentation results under the CTTA setting for the fundus and polyp tasks are reported in Tab. 1 and Tab. 2. For each experiment, we treat one domain as the source and evaluate the pretrained model on the remaining target domains. All TTA-based methods outperform the ‘No Adapt’ baseline across domains, and our method, TanGo, attains the highest average performance in every domain. In particular, TanGo exceeds the second-best method by 10.91% and improves over no-adapt by 19.37% in the fundus task. A consistent trend is observed in the polyp task: TanGo outperforms the second-best VPTTA and the ‘No Adapt’ baseline by 3.24% and 7.93%, respectively. These gains stem from TanGo’s combination of distribution-level consistency learning during training and instance-level distribution correction at test time, which together yield more stable adaptation under continuously evolving test distributions. We provide qualitative segmentation comparisons in Fig. 6 of the Ap-

Table 3. Comparison with ‘No Adapt’ and nine competing TTA methods on the polyp segmentation task, reporting only DSC (%). Top three results are highlighted as **best**, **second** and **third**, respectively.

Methods	A→B	A→D	B→A	B→D	C→A	C→B	C→D	D→A	D→B	A→rest	B→rest	D→rest	Average ↑
No Adapt	66.49	70.84	76.21	70.03	72.87	69.13	70.44	79.78	72.74	67.70	62.80	74.19	72.86
TENT [53] (ICLR 2021)	68.13	66.11	69.20	74.06	73.21	65.04	72.35	77.88	70.01	66.92	59.87	73.12	68.28
DLTTA [57] (TMI 2022)	72.23	74.56	64.87	78.47	71.22	64.78	71.00	79.89	79.10	68.59	61.12	73.22	71.87
DUA [35] (CVPR 2022)	72.17	72.53	74.93	78.82	75.14	70.33	75.67	80.26	76.81	70.32	67.89	78.59	75.73
SAR [39] (ICLR 2023)	70.34	79.45	74.98	80.21	76.82	71.27	77.40	82.65	78.55	77.69	71.87	80.78	77.64
DIGA [55] (CVPR 2023)	76.18	76.25	78.25	79.60	78.05	72.33	78.14	83.47	79.26	76.58	71.86	81.12	77.69
DomainAdaptor [59] (CVPR 2023)	78.39	79.45	74.98	80.21	76.02	73.40	81.33	86.10	80.56	76.92	71.12	77.26	77.96
VPTTA [6] (CVPR 2024)	78.18	82.73	73.64	82.86	77.90	70.64	77.22	87.19	84.97	80.98	78.84	82.32	79.84
CertainTTA [10] (IF 2025)	79.10	83.26	78.15	81.36	78.10	72.33	78.52	88.15	81.27	78.32	78.76	82.78	80.00
MGMDG [32] (CVPR 2025)	80.47	83.18	78.33	83.58	78.86	76.49	80.20	89.17	81.68	82.98	79.36	83.51	81.48
TanGo (Ours)	83.01	86.89	81.06	85.95	80.24	80.56	83.65	89.88	87.80	86.80	80.22	87.44	84.46

Table 4. Ablation study of DSCL in the polyp tasks. The best results in each column are highlighted in **bold**.

Methods	Domain A	Domain B	Domain C	Domain D	Average ↑
	Dice Score Metric (DSC)				
No Adapt	79.90	66.33	73.89	82.95	75.77
w/ FAFPI	80.53	74.68	74.73	83.42	78.34
w/ FCL	81.89	78.89	76.88	85.98	80.91
Ours	83.95	82.44	81.11	87.29	83.70

pendix. Visualizations of both datasets show that TanGo produces clearer and more accurate organ boundaries than competing methods, further corroborating its effectiveness.

Quantitative results of static TTA on the polyp segmentation task are summarized in Tab. 3. Our approach achieves the highest average DSC among nine competitors, surpassing the second-best by 2.98%, demonstrating its strong segmentation capability even under static adaptation. Furthermore, we conduct three extended static TTA evaluations—A→rest, B→rest, and D→rest (last three columns)—to assess long-term adaptation performance. The results show that our method consistently maintains superior accuracy and robustness, achieving a clear margin over the second-best method. This indicates that our approach benefits particularly from sufficient test samples, showcasing its stable and effective adaptation behavior in realistic TTA scenarios. See Appendix C.2 for additional experiments of mixing distribution shifts in the CTTA setting.

4.3. Ablation Analysis of Our Method

Effectiveness of the DSCL module. In this section, we conduct ablation studies to evaluate the effectiveness of the proposed Distribution-wise Consistency Learning (DSCL) during training. As shown in Tab. 4, three configurations are compared: (1) With our FAFPI (denoted as *w/FAFPI*), (2) with FAFPI and FCL during training (denoted as *w/FCL*); (3) with DSCL during training and Instance-Wise Style Adaptive Calibration during test time (denoted as *Ours*). As summarized in Tab. 4, the progressive performance gains from (1) to (2) indicate that the generalization priors from the foundation model are effectively transferred

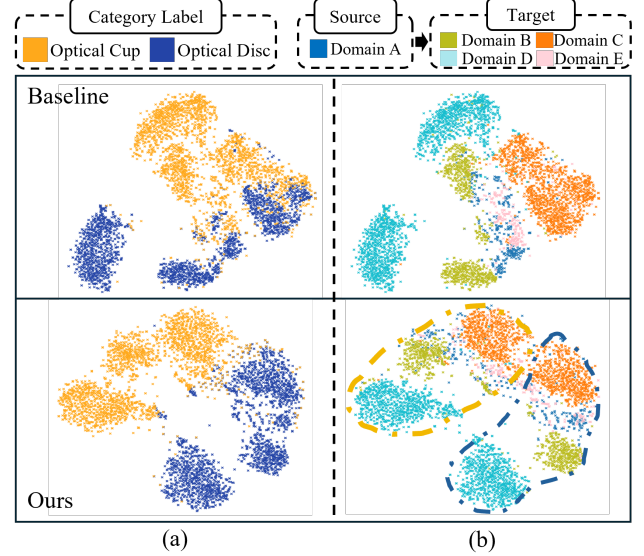


Figure 3. The t-SNE visualization uses colors to distinguish categories in the first column and domains in the second, with the **orange** and **blue** dashed lines marking the distribution contours of the two categories; our method (second row) demonstrates superior discriminative ability and more uniform distribution coherence.

to the source model, leading to consistent improvements in segmentation performance. The further average gain of 2.79 DSC in (3) demonstrates the complementary effect between the enhanced source model from training and the test-time calibration module, which jointly strengthen the model’s adaptability to unseen domains. This integrated paradigm of training-time transfer and test time adaptation effectively enhances the model’s adaptability during test time.

Visualization analysis of DSCL. As illustrated by the t-SNE results in Fig. 3, the source model enhanced by the foundation prior exhibits a more compact and discriminative feature distribution compared with the baseline. The inter-domain gaps are significantly reduced, providing a more uniform and generalizable representation space that facilitates stable test-time adaptation to unseen domains.

Comparison DSCL to distillation methods. We compare Distribution-wise Consistency Learning (DSCL) against several feature distillation methods CWD [46], AttnKD [24]

Table 5. Comparison of Different Distillation Methods in polyp segmentation tasks. The best results are highlighted in **bold**.

	No Adapt	SpaKD	CWD	AttnKD	Ours
polyp $\uparrow$	75.77	76.73	78.69	79.02	80.91
fundus $\uparrow$	68.92	72.62	77.56	81.32	83.05

Table 6. Ablation Study of Different Constraints in Instance-wise Style Adaptive Calibration on the polyp tasks.

Components	Domain A	Domain B	Domain C	Domain D	Average
$\mathcal{L}_{sas}$ $\mathcal{L}_{sad}$ $\mathcal{L}_{ent}$	DSC	DSC	DSC	DSC	
	79.90	66.33	73.89	82.95	75.77
	79.87	73.78	74.49	84.87	78.25
	81.26	79.76	76.21	86.34	80.89
$\checkmark$	83.04	81.32	80.45	86.51	82.83
$\checkmark$ $\checkmark$	83.95	82.44	81.11	87.29	83.70

and SpaKD [21]. As shown in Tab. 5, while all distillation variants enhance the source model’s generalization by leveraging the strong priors of the foundation model, DSCL achieves the best performance. This superiority stems from its fundamental design: DSCL operates at a global distribution level to integrate the foundation model’s generalization prior, learning to form a more robust and transferable feature space. In contrast, methods like channel or attention distillation only perform local-level imitation of specific features or spatial maps, which offers a weaker form of guidance.

Effectiveness of ISAC. We conduct ablation studies on the effectiveness of the Instance-wise Style Adaptive Calibration (ISAC) in Tab. 6, focusing on the key components: the Source-Anchored Distributional Loss ($\mathcal{L}_{sad}$), the Source-Anchored Semantic Loss ($\mathcal{L}_{sas}$), and the entropy loss $\mathcal{L}_{ent}$. The results in Tab. 6 reveal three important observations: (1) $\mathcal{L}_{sas}$ effectively constrains the Learnable Instance Aware Decorator module (LID) to align the distribution of test samples toward the enhanced source space, thereby narrowing the domain gap and improving performance. (2) By enforcing semantic consistency between the outputs for each test sample before and after LID augmentation, $\mathcal{L}_{sad}$ enhances model robustness and performance in dynamic test environments. (3) The best results are achieved when $\mathcal{L}_{sas}$, $\mathcal{L}_{sad}$, and $\mathcal{L}_{ent}$ work jointly, enforcing consistency in both the feature distribution and semantic outputs.

Visualization analysis of ISAC. As illustrated in Fig. 4, under the dual constraints of the Source Knowledge Anchored Recalibration method, the pseudo test samples generated by the Learnable Instance aware Decorator (LID) exhibit source-like styles that are clearly separated from the original test domain, demonstrating the effectiveness of $\mathcal{L}_{sas}$ and $\mathcal{L}_{sad}$. Similarly, the t-SNE visualization in Fig. 7 (supported by additional results in Appendix. C.3) shows that ISAC achieves a more uniform feature distribution, effectively mitigates domain shifts and error accumulations during continual test-time adaptation.

Ablation study on hyperparameter r . Our analysis of hy-

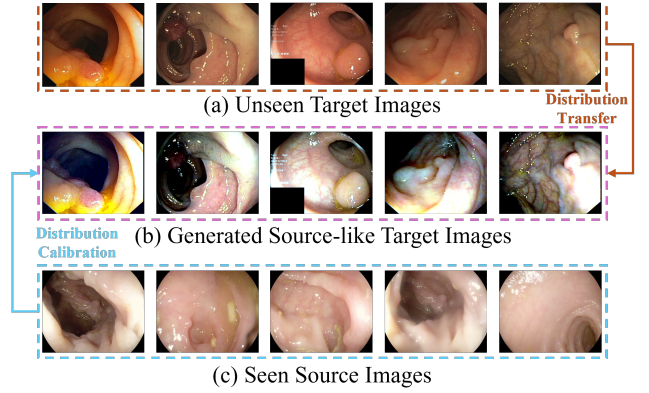


Figure 4. The visualization of (a) unseen multi-target domains, (b) LID generated images, and (c) the single source domain demonstrates that our method effectively transforms stylistically diverse test samples into source-like distributions while preserving their semantic integrity.

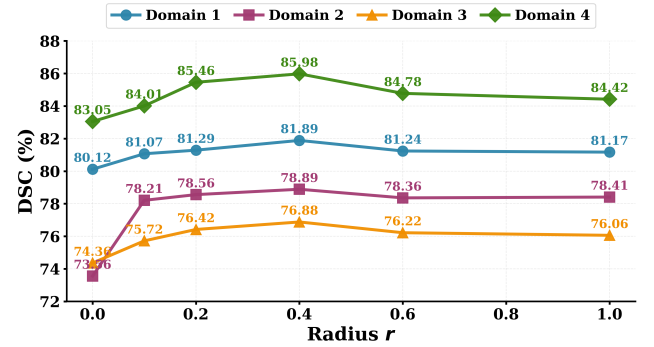


Figure 5. Performance of VPTTA with different values of r across the four polyp segmentation domains.

perparameter r in Eq. (2) in the polyp tasks (Fig. 5) reveals that a value of 0.4 yields the best performance. This value strikes a balance: lower values (e.g., 0) lead to insufficient VFM guidance, while higher ones degrade generalization by distorting high-frequency features from the source model.

5. Conclusion

We propose *TanGo*, a continual test time adaptation framework that unifies distribution-level consistency learning at training time with instance-wise style adaptive calibration at test time. During training, *TanGo* builds a stable and more generalizable source feature space by injecting robust low-frequency priors from vision foundation models, thereby mitigating source overfitting. At test time, our instance-wise style calibration adaptively transforms each test sample toward this enhanced source space under a set of source-anchored recalibration constraints, which suppress error accumulation and reduce catastrophic forgetting during continual adaptation. Extensive experiments on multiple medical image segmentation benchmarks show that *TanGo* consistently outperforms state-of-the-art baselines. We hope that our work offers a fresh and different perspective for advancing continual domain adaptation in medical imaging.

References

- [1] Jorge Bernal, Javier Sánchez, and Fernando Vilarino. Towards automatic polyp detection with a polyp appearance model. *Pattern Recognition*, 45(9):3166–3182, 2012. 6
- [2] Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, Debora Gil, Cristina Rodríguez, and Fernando Vilariño. Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized medical imaging and graphics*, 43:99–111, 2015. 6
- [3] Qi Bi, Jingjun Yi, Hao Zheng, Haolan Zhan, Yawen Huang, Wei Ji, Yuexiang Li, and Yefeng Zheng. Learning frequency-adapted vision foundation model for domain generalized semantic segmentation. *Advances in Neural Information Processing Systems*, 37:94047–94072, 2024. 2
- [4] Fabio M Carlucci, Antonio D’Innocente, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. Domain generalization by solving jigsaw puzzles. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2229–2238, 2019. 1, 2
- [5] Ziyang Chen, Yongsheng Pan, Yiwen Ye, Hengfei Cui, and Yong Xia. Treasure in distribution: a domain randomization based multi-source domain generalization for 2d medical image segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 89–99. Springer, 2023. 1, 2
- [6] Ziyang Chen, Yongsheng Pan, Yiwen Ye, Mengkang Lu, and Yong Xia. Each test image deserves a specific prompt: Continual test-time adaptation for 2d medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11184–11193, 2024. 2, 3, 6, 7, 1, 4
- [7] Ziyang Chen, Yiwen Ye, Yongsheng Pan, and Yong Xia. Gradient alignment improves test-time adaptation for medical image segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2429–2437, 2025. 6, 3
- [8] Ming-Ming Cheng and Deng-Ping Fan. Structure-measure: A new way to evaluate foreground maps. *International Journal of Computer Vision*, 129:2622–2638, 2021. 6
- [9] Lu Chi, Borui Jiang, and Yadong Mu. Fast fourier convolution. *Neural Information Processing Systems, Neural Information Processing Systems*, 2020. 3
- [10] Xingbo Dong, Liwen Wang, Xingguo Lv, Xiaoyan Zhang, Hui Zhang, Bin Pu, Zhan Gao, Iman Yi Liao, and Zhe Jin. Certaintta: Estimating uncertainty for test-time adaptation on medical image segmentation. *Information Fusion*, page 103300, 2025. 6, 7, 2, 3
- [11] Deng-Ping Fan, Cheng Gong, Yang Cao, Bo Ren, Ming-Ming Cheng, and Ali Borji. Enhanced-alignment measure for binary foreground map evaluation. *arXiv preprint arXiv:1805.10421*, 2018. 6
- [12] Deng-Ping Fan, Ge-Peng Ji, Tao Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao. Pranet: Parallel reverse attention network for polyp segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 263–273. Springer, 2020. 6, 1
- [13] Shenghao Fu, Junkai Yan, Qize Yang, Xihan Wei, Xiaohua Xie, and Wei-Shi Zheng. Frozen-detr: Enhancing detr with image understanding from frozen foundation models. *Advances in Neural Information Processing Systems*, 37:105949–105971, 2024. 2
- [14] Francisco Fumero, Silvia Alayón, José L Sanchez, Jose Sigut, and M Gonzalez-Hernandez. Rim-one: An open retinal image database for optic nerve evaluation. In *2011 24th international symposium on computer-based medical systems (CBMS)*, pages 1–6. IEEE, 2011. 5
- [15] Yulu Gan, Yan Bai, Yihang Lou, Xianzheng Ma, Renrui Zhang, Nian Shi, and Lin Luo. Decorate the newcomers: Visual domain prompt for continual test time adaptation. In *Proceedings of the AAAI conference on artificial intelligence*, pages 7595–7603, 2023. 5
- [16] Yossi Gandelsman, Yu Sun, Xinlei Chen, and Alexei Efros. Test-time training with masked autoencoders. *Advances in Neural Information Processing Systems*, 35:29374–29385, 2022. 3
- [17] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Texture synthesis using convolutional neural networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, page 262–270, 2015. 5
- [18] Jintao Guo, Lei Qi, and Yinghuan Shi. Domaindrop: Suppressing domain-sensitive channels for domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19114–19124, 2023. 2
- [19] Jintao Guo, Na Wang, Lei Qi, and Yinghuan Shi. Aloft: A lightweight mlp-like architecture with dynamic low-frequency transform for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24132–24141, 2023. 3
- [20] Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. Towards a unified view of parameter-efficient transfer learning. *arXiv preprint arXiv:2110.04366*, 2021. 1
- [21] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 8
- [22] Shishuai Hu, Zehui Liao, and Yong Xia. Source-free domain adaptation using prompt learning for medical image segmentation. *Pattern Recognition*, page 112290, 2025. 3, 5
- [23] Jiaying Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. Fsdrr: Frequency space domain randomization for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6891–6902, 2021. 3
- [24] Mingi Ji, Byeongho Heo, and Sungrae Park. Show, attend and distill: Knowledge distillation via attention-based feature matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7945–7952, 2021. 7
- [25] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4015–4026, 2023. 2, 1
- [26] Junzhe Li, Xuerui Qiu, Linrui Xu, Liya Guo, Delin Qu, Tingting Long, Chun Fan, and Ming Li. Unif2ace: Fine-grained

- face understanding and generation with unified multimodal models. *arXiv preprint arXiv:2503.08120*, 3, 2025. 2
- [27] Ming Li, Pan Zhou, Jia-Wei Liu, Jussi Keppo, Min Lin, Shuicheng Yan, and Xiangyu Xu. Instant3d: Instant text-to-3d generation. *International Journal of Computer Vision*, 132(10):4456–4472, 2024. 3
- [28] Chuang Liu, Yichao Cao, Xiu Su, and Haogang Zhu. Universal frequency domain perturbation for single-source domain generalization. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 6250–6259, 2024. 1
- [29] Chuang Liu, Hongyan Xu, Yichao Cao, Xiu Su, Zhe Qu, Tianfa Li, Shan An, and Haogang Zhu. Tinyimg: Transferring generalization from vision foundation models to single-domain medical imaging. In *International Conference on Machine Learning*, pages 39898–39914. PMLR, 2025. 2
- [30] Quande Liu, Cheng Chen, Jing Qin, Qi Dou, and Pheng-Ann Heng. Feddg: Federated domain generalization on medical image segmentation via episodic learning in continuous frequency space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1013–1023, 2021. 1, 5
- [31] Quande Liu, Cheng Chen, Qi Dou, and Pheng-Ann Heng. Single-domain generalization in medical image segmentation via test-time adaptation from shape dictionary. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1756–1764, 2022. 3
- [32] Xingguo Lv, Xingbo Dong, Liwen Wang, Jiewen Yang, Lei Zhao, Bin Pu, Zhe Jin, and Xuejun Li. Test-time domain generalization via universe learning: A multi-graph matching approach for medical image segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 15621–15631, 2025. 6, 7
- [33] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1):654, 2024. 2
- [34] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. Ieee, 2016. 1
- [35] M Jehanzeb Mirza, Jakub Micorek, Horst Possegger, and Horst Bischof. The norm must go on: Dynamic unsupervised domain adaptation by normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14765–14775, 2022. 6, 7, 3, 4
- [36] A Tuan Nguyen, Thanh Nguyen-Tang, Ser-Nam Lim, and Philip HS Torr. Tipi: Test time adaptation with transformation invariance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24162–24171, 2023. 2, 3
- [37] Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In *International conference on machine learning*, pages 16888–16905. PMLR, 2022. 3
- [38] Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In *International conference on machine learning*, pages 16888–16905. PMLR, 2022. 2
- [39] Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Zhiqian Wen, Yaofo Chen, Peilin Zhao, and Mingkui Tan. Towards stable test-time adaptation in dynamic wild world. *arXiv preprint arXiv:2302.12400*, 2023. 6, 7, 2, 3, 4
- [40] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 2, 4
- [41] José Ignacio Orlando, Huazhu Fu, João Barbosa Breda, Karel Van Keer, Deepti R Bathula, Andrés Diaz-Pinto, Ruogu Fang, Pheng-Ann Heng, Jeyoung Kim, JoonHo Lee, et al. Refuge challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. *Medical image analysis*, 59:101570, 2020. 5
- [42] Cheng Ouyang, Chen Chen, Surui Li, Zeju Li, Chen Qin, Wenjia Bai, and Daniel Rueckert. Causality-inspired single-source domain generalization for medical image segmentation. *IEEE Transactions on Medical Imaging*, 42(4):1095–1106, 2022. 2
- [43] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 1
- [44] Swami Sankaranarayanan, Yogesh Balaji, Arpit Jain, Ser Nam Lim, and Rama Chellappa. Learning from synthetic data: Addressing domain shift for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3752–3761, 2018. 1
- [45] Yufei Shi, Weilong Yan, Gang Xu, Yumeng Li, Yucheng Chen, Zhenxi Li, Fei Yu, Ming Li, and Si Yong Yeo. Pvchat: Personalized video chat with one-shot learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 23321–23331, 2025. 2
- [46] Changyong Shu, Yifan Liu, Jianfei Gao, Zheng Yan, and Chunhua Shen. Channel-wise knowledge distillation for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5311–5320, 2021. 7
- [47] Juan Silva, Aymeric Histace, Olivier Romain, Xavier Dray, and Bertrand Granado. Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer. *International journal of computer assisted radiology and surgery*, 9:283–293, 2014. 6
- [48] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025. 4
- [49] Jayanthi Sivaswamy, SR Krishnadas, Gopal Datt Joshi, Madhulika Jain, and A Ujjwaft Syed Tabish. Drishti-gs: Retinal image dataset for optic nerve head (onh) segmentation. In *2014 IEEE 11th international symposium on biomedical imaging (ISBI)*, pages 53–56. IEEE, 2014. 5
- [50] Junha Song, Jungsoo Lee, In So Kweon, and Sungha Choi. Ecotta: Memory-efficient continual test-time adaptation via

- self-distilled regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11920–11929, 2023. [2](#)
- [51] Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *International conference on machine learning*, pages 9229–9248. PMLR, 2020. [3](#)
- [52] Lulu Tian, Hongxun Yao, and Ming Li. Fakepoi: A large-scale fake person of interest video detection benchmark and a strong baseline. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(11):6819–6831, 2023. [2](#)
- [53] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*, 2020. [2](#), [3](#), [6](#), [7](#), [4](#)
- [54] Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7201–7211, 2022. [2](#), [3](#), [6](#), [4](#)
- [55] Wei Wang, Zhun Zhong, Weijie Wang, Xi Chen, Charles Ling, Boyu Wang, and Nicu Sebe. Dynamically instance-guided adaptation: A backward-free approach for test-time domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24090–24099, 2023. [7](#)
- [56] Zhikai Wei, Wenhui Dong, Peilin Zhou, Yulian Gu, Zhou Zhao, and Yongchao Xu. Prompting segment anything model with domain-adaptive prototype for generalizable medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 533–543. Springer, 2024. [2](#), [3](#), [6](#), [1](#)
- [57] Hongzheng Yang, Cheng Chen, Meirui Jiang, Quande Liu, Jianfeng Cao, Pheng Ann Heng, and Qi Dou. Dltta: Dynamic learning rate for test-time adaptation on cross-domain medical images. *IEEE Transactions on Medical Imaging*, 41(12): 3575–3586, 2022. [3](#), [6](#), [7](#), [2](#), [4](#)
- [58] Chuyan Zhang, Hao Zheng, Xin You, Yefeng Zheng, and Yun Gu. Pass: test-time prompting to adapt styles and semantic shapes in medical image segmentation. *IEEE Transactions on Medical Imaging*, 2024. [3](#)
- [59] Jian Zhang, Lei Qi, Yinghuan Shi, and Yang Gao. Domainadaptor: A novel approach to test-time adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18971–18981, 2023. [2](#), [3](#), [6](#), [7](#), [4](#)
- [60] Zhuo Zhang, Feng Shou Yin, Jiang Liu, Wing Kee Wong, Ngan Meng Tan, Beng Hai Lee, Jun Cheng, and Tien Yin Wong. Origa-light: An online retinal fundus image database for glaucoma analysis and research. In *2010 Annual international conference of the IEEE engineering in medicine and biology*, pages 3065–3068. IEEE, 2010. [5](#)
- [61] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45:4396–4415, 2021. [3](#)
- [62] Ziqi Zhou, Lei Qi, and Yinghuan Shi. Generalizable medical image segmentation via random amplitude mixup and domain-specific image restoration. In *European Conference on Computer Vision*, pages 420–436. Springer, 2022. [1](#)
- [63] Chunzheng Zhu, Jianxin Lin, Guanghua Tan, Ningbo Zhu, Kenli Li, Chunlian Wang, and Shengli Li. Advancing ultrasound medical continuous learning with task-specific generalization and adaptability. In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 3019–3025. IEEE, 2024. [1](#)
- [64] Chunzheng Zhu, Yangfang Lin, Shen Chen, Yijun Wang, and Jianxin Lin. Medeyes: Learning dynamic visual focus for medical progressive diagnosis. *arXiv preprint arXiv:2511.22018*, 2025. [1](#)
- [65] Chunzheng Zhu, Yangfang Lin, Jialin Shao, Jianxin Lin, and Yijun Wang. Pathology-aware prototype evolution via llm-driven semantic disambiguation for multicenter diabetic retinopathy diagnosis. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 9196–9205, 2025. [2](#)