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

Supplementary Material

In the Appendix, for clarity and ease of navigation, Tab. 7 lists all abbreviations used in the paper along with their corresponding full names and the locations where they are referenced. We also provide additional experimental details and extended results for the medical image segmentation task. These include **Section A**, which contains detailed descriptions of the additional modules discussed in the *Distribution-wise Consistency Learning (DSCL)* section 3.2 of the main paper; **Section B** presents further experimental settings for the medical segmentation task; and **Section C** includes supplementary experimental validations. These validations consist of: (***C.1***) the results of comparative methods under the TTA setting on the fundus tasks, (***C.2***) comparison experiments under continual mixed distribution shifts on the polyp tasks, (***C.3***) visualization of sample distributions of the test-time data generated by the *Learnable Instance-wise Decorator (LID)* module constrained by the *Source Knowledge Anchored Recalibration (SKAR)* loss functions, (***C.4***) ablation studies on the weights of different loss functions used in the SKAR method, (***C.5***) ablation Study of Random noise used in the DSCL method, and (***C.6***) the effect of Different Vision Foundation Model Backbones in DSCL.

A. Details of the Adapter and Feature Refinement Module (FRM) in DSCL

PEFT for Foundation Models. To enhance the adaptability of vision foundation models (VFMs) to the medical domain and improve feature alignment with the specialized model, we adopt a two-stage strategy. First, we integrate an Adapter module following the design in DAPSAM [56], which enables parameter-efficient fine-tuning [20] at every layer of the foundation model. In our framework, this adapter is incorporated into the FAFPI module within TanGo's Distribution-wise Consistency Learning branch (DSCL), allowing the model to absorb domain-relevant variations while keeping the backbone largely frozen. The formulation of the adapter is given as follows:

$$F_i^{\mathcal{V}\prime} = F_i^{\mathcal{V}} + \mathrm{MLP}_{\mathrm{up}}(\mathrm{GELU}(\mathrm{MLP}_{\mathrm{down}}(F_i^{\mathcal{V}}))) \quad (14)$$

In this adapter module, the input visual feature $F_i^{\mathcal{V}}$ is first projected into a lower-dimensional space by $\mathrm{MLP}_{\mathrm{down}}$, then processed through a GELU activation function, and subsequently projected back to the original dimension via $\mathrm{MLP}_{\mathrm{up}}$. This bottleneck structure allows the model to learn domain-specific adjustments with minimal additional parameters. The final output $F_i^{\mathcal{V}\prime}$ is obtained by adding the adapter's output to the original input feature $F_i^{\mathcal{V}}$ through a residual connection. This design ensures stable training by preserving the original feature representation while introducing a small, tunable transformation for effective domain adaptation.

Second, we propose the Feature Refinement Module (FRM) to further refine the extracted features, making them more suitable for downstream tasks and specialized models. At its core, FRM utilizes residual blocks to perform feature dimension transformation and alignment, ensuring effective adaptation of foundation model representations. The operation for each feature is formulated as:

$$\hat{F}_i^{\mathcal{V}} = \mathrm{ReLU}\left(\mathrm{BN}\left(\mathrm{Conv}_i\left(F_i^{\mathcal{V}\prime}\right)\right)\right), \quad i = 3, 6, 9, 12 \quad (15)$$

In this formulation, Conv_i denotes a convolutional layer (either standard or transposed) that performs two critical functions: it projects the channel dimension of the input feature to a target size C_i and simultaneously adjusts its spatial resolution via a specified stride based on a pre-defined scale factor. The Batch Normalization (BN) layer stabilizes the training process, and the ReLU activation function introduces non-linearity. The resulting refined features $\hat{F}_i^{\mathcal{V}}$ possess aligned channel dimensions and coordinated spatial scales, making them suitable for effective low-frequency fusion in Frequency Aware Foundation Prior Injection with the source model f_S.

B. Further Experimental Settings for the medical segmentation tasks

Training the source model with Distribution-wise Consistency Learning (DSCL). For all medical segmentation tasks, we train the model on a single source domain and directly evaluate it on the remaining target domains. Following VPTTA [6], the source domain model f_S is implemented as a ResUNet-34 for the OC/OD segmentation task and as a PraNet [12] with a Res2Net backbone for the polyp segmentation task. During training, only the foundation model f_V, which corresponds to the SAM [25] encoder, is utilized. We further integrate an adapter [56] for fine-tuning and a Feature Refinement Module to align the feature dimensions. We employ the AdamW optimizer for training the source model, with momentum parameters $\beta = [0.9, 0.999]$ and an initial learning rate $l_0 = 1 \times 10^{-4}$. The learning rate follows a polynomial decay schedule defined as $l_t = l_0 \times \left(1 - \frac{t}{T}\right)^{0.9}$, where t and T denote the current and total epochs, respectively. Both tasks are trained for 100 epochs with a batch size of 8. At test time, we follow protocol in [6] and optimize the

Table 7. Summary of abbreviations of metohds used in the paper, their corresponding locations, and relevant annotations.

Abbrev.	Full Name	Location / Notes
TanGo	Training to Adapt with Foundation Guidance and Continual Style Calibration	Tab. 1, Tab. 2, Tab. 3, Tab. 8, Tab. 9, Tab. 10; Our framework
DSCL 3.2	Distribution-wise Consistency Learning	Tab. 4, Tab. 5, Tab. 12, Tab. 13, Fig. 3 ; DSCL = w/FAFPI + w/FCL
FAFPI 3.2.1	Frequency Aware Foundation Prior Injection	Tab. 4; None
FCL 3.2.2	Foundation-informed Consistency Learning	Tab. 4; None
ISAC 3.3	Instance-wise Style Adaptive Calibration	Fig. 7; ISAC = w/LID + w/SKAR
LID 3.3.1	Learnable Instance-aware Decorator	Fig. 4
SKAR 3.3.2	Source Knowledge Anchored Recalibration	Tab. 5; SKAR = $\mathcal{L}_{sad}$ + $\mathcal{L}_{sas}$ + $\mathcal{L}_{ent}$
$\mathcal{L}_{sad}$	Source-Anchored Distributional Loss	Tab. 6, Tab. 11
$\mathcal{L}_{sas}$	Source-Anchored Semantic Loss	Tab. 6, Tab. 11
$\mathcal{L}_{ent}$	Self-supervised Entropy Loss	Tab. 6, Tab. 11

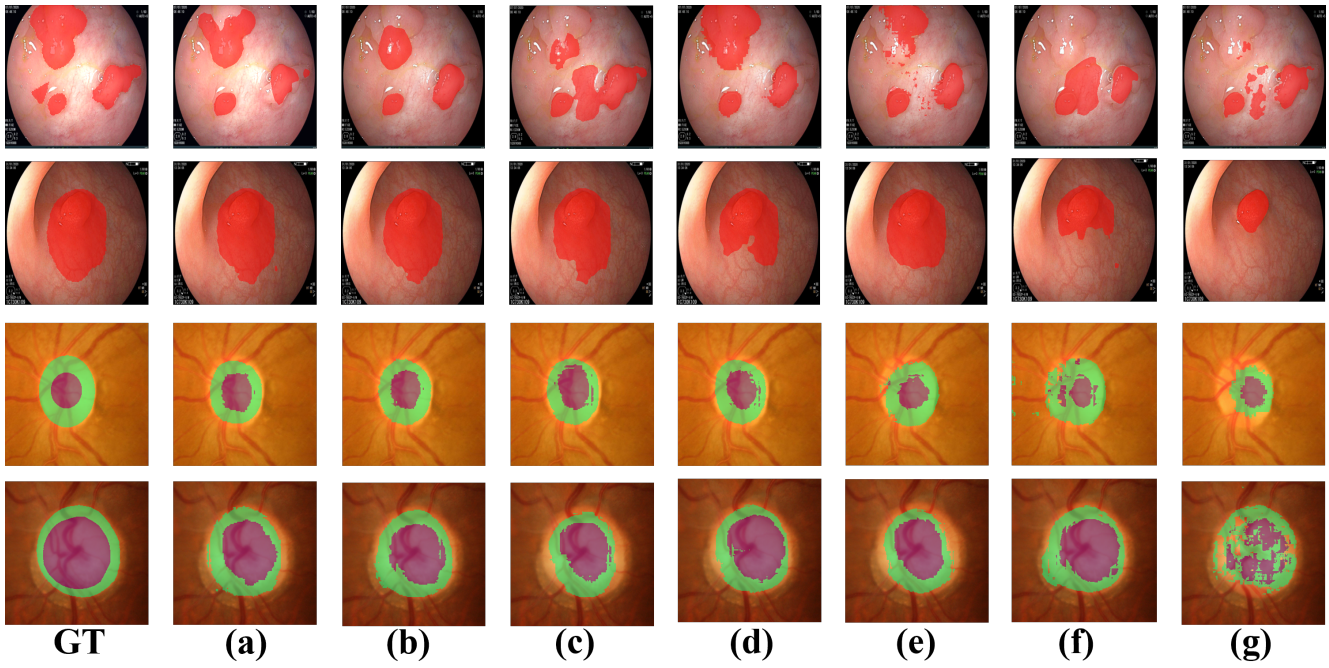

Figure 6. Comparisons across different TTA Methods on Fundus and Polyp tasks with ground truth (GT) and predictions (a-g). The subfigures (a) to (g) correspond to: (a) ***TanGo***, (b) VPTTA [6], (c) CertainTTA [10], (d) DomainAdaptor [59], (e) SAR [39], (f) DLTTA [57], (g) TENT [53].

parameters θ in Eq. (13) using SGD with a learning rate of 0.001, and performed one iteration adaptation for each batch of test data with a batch size of 1 on all experiments of our TanGo and other competing methods. All experiments are conducted on the NVIDIA RTX 4090 GPUs.

C. Supplementary experimental Validations

C.1. Results of comparative methods under the TTA setting on the Fundus dataset

Additionally, quantitative segmentation results on the fundus tasks under the static TTA protocol are reported in Tab. 8. As expected, the static TTA setting is less demanding than Continual TTA (CTTA), and thus most methods, including TENT and its variants DLTTA and DomainAdaptor, achieve higher performance under static TTA. Crucially, our method still attains the highest mean performance across the 12 evaluated settings, outperforming the second-best approach by 5.6% in DSC. This gap demonstrates the robustness and effectiveness of our approach under test-time adaptation.

C.2. Comparison experiments under mixed distribution shifts on the both tasks

To simulate the arbitrary and unpredictable nature of test data encountered in complex real-world environments, we performed experiments under mixed distribution shifts in both

Table 8. Comparison with 'No Adapt' and nine competing methods under TTA, on the OD/OC segmentation task. The best and second-best results in each column are highlighted in **bold** and underline, respectively. Using Dice coefficient % as the evaluation metric.

Methods	DSC												Average ↑
	A→ B	A→ C	A→ D	A→ E	B→ A	B→ C	B→ D	B→ E	C→ A	C→ B	C→ D	C→ E	
No Adapt (UNet)	62.63	53.26	67.18	55.22	60.21	62.13	54.28	68.02	54.10	62.91	53.53	71.23	60.39
TENT (ICLR 2021) [53]	73.07	58.26	61.24	60.81	67.20	62.13	70.00	72.36	54.68	68.75	56.24	76.53	65.11
COTTA (CVPR 2022) [54]	77.39	65.98	69.14	65.99	70.40	67.98	72.30	75.18	55.03	74.85	60.28	79.05	69.46
DLTTA (TMI 2022) [57]	79.11	65.85	69.89	69.64	70.71	69.84	73.11	76.04	55.19	76.93	62.25	80.67	70.20
DUA (CVPR 2022) [35]	79.28	66.59	70.13	70.17	71.38	69.31	73.64	78.48	56.22	78.93	63.55	82.81	71.71
SAR (ICLR 2023) [39]	79.55	65.71	70.78	71.40	71.72	70.03	74.28	78.05	56.27	79.46	63.94	83.45	72.05
DIGA (CVPR 2023) [61]	80.78	64.28	71.30	71.50	73.03	70.38	75.28	79.10	57.26	80.04	64.82	83.19	72.58
DomainAdaptor (CVPR 2023) [59]	81.58	65.95	71.81	72.16	70.55	70.01	75.42	80.26	56.90	79.28	65.26	84.55	72.81
TIPI (CVPR 2023) [36]	81.66	65.18	70.42	72.61	71.37	70.84	76.25	81.65	58.05	81.40	65.62	84.90	73.33
VPTTA (CVPR 2024) [6]	82.17	65.22	70.45	73.04	73.29	71.94	76.88	81.56	58.48	81.93	65.25	85.02	73.77
CertainTTA [10]	84.96	68.18	74.24	73.76	74.68	75.49	77.95	83.13	64.59	84.03	68.54	89.39	76.58
TanGo (Ours)	**86.50**	**87.52**	**74.60**	**88.87**	**76.60**	**85.90**	**88.73**	**87.72**	**82.19**	**88.23**	**84.27**	**92.03**	**85.26**

Table 9. Performance of our TanGo, 'Source Only' baseline, and six competing methods on the OD/OC segmentation task. The best and second-best results in each column are highlighted in **bold** and underline, respectively. Using Dice coefficient % as the evaluation metric.

Methods	**Domain A**	**Domain B**	**Domain C**	**Domain D**	**Domain E**	Average ↑
	DSC	*DSC*	*DSC*	*DSC*	*DSC*	
Source Only (ResUNet-34)	64.53	76.06	71.18	52.67	64.87	65.86
TENT-continual (ICLR 2021) [53]	71.50	77.96	72.79	42.97	69.56	66.96
CoTTA (CVPR 2022) [54]	73.71	76.31	72.43	53.04	71.14	69.33
DLTTA (TMI 2022) [57]	74.90	78.73	74.48	50.99	69.25	69.67
DUA (CVPR 2022) [35]	73.06	75.74	70.82	57.04	70.31	69.39
SAR (ICLR 2023) [39]	74.48	77.49	70.78	57.93	73.05	70.75
DomainAdaptor (CVPR 2023) [59]	74.50	76.39	71.81	56.78	70.55	70.01
VPTTA (CVPR2024) [6]	74.24	79.12	74.05	55.84	76.47	71.94
GraTa (AAAI2025) [7]	76.42	78.56	76.12	66.85	73.79	74.35
TanGo (Ours)	**86.52**	**84.87**	**83.26**	**84.01**	**82.76**	**84.28**

the polyp and fundus benchmarks. Specifically, models were trained on a single source domain with our Distribution-wise Consistency Learning (DSCL) and evaluated on a mixed domain constructed from the remaining target domains. For all methods, we shuffled the data of the target domains using a random seed of 2024, with a batch size of 1. The outcomes for two medical segmentation benchmark tasks are summarized in Tab. 9 and Tab. 10. A consistent pattern emerges: only DUA, VPTTA, and our method (TanGo) surpass the baseline, while the remaining approaches degrade because their adaptation strategies amplify misleading gradients from overconfident yet inaccurate predictions. Moreover, the results across both tables indicate that our approach consistently achieves the highest overall performance across all domains in both tasks, emphasizing its enhanced applicability and robustness in dynamic and unstructured settings.

C.3. Distribution visualizations of sample generated by LID

In this section, we visualize the feature distributions of test samples generated by the learnable Instance-aware Decorator (LID) in the Instance-wise Style Adaptive Calibration module (ISAC) of TanGo using t-SNE. As shown in Fig. 7, we compare our method with the baseline (TENT) and a recent representative approach (VPTTA). Different colors denote distinct target domains. The visualizations are derived from the last-layer features of the segmentation encoder for the adapted test samples. It can be observed that the features obtained by our method are distributed more uniformly, with significantly reduced inter-domain discrepancies. This indicates that TanGo effectively mitigates domain shift during test-time adaptation and exhibits strong domain generalization capability.

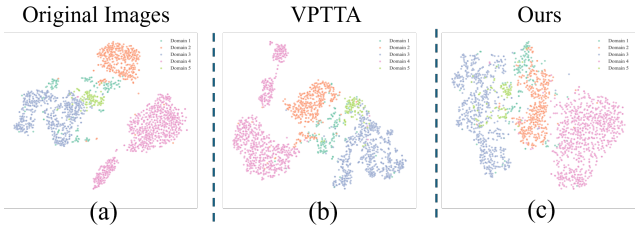

Figure 7. The t-SNE visualization uses colors to distinguish domains in the first row, our method (third column) demonstrates superior discriminative ability and more uniform distribution coherence compared to the sota VPTTA method.

Table 10. Performance of our TanGo, 'Source Only' baseline, and six competing methods on the polyp segmentation task. The best and second-best results in each column are highlighted in **bold** and underline, respectively. Using Dice coefficient % as the evaluation metric.

Methods	Domain A			Domain B			Domain C			Domain D			Average ↑		
	DSC	$E_\phi^{\max}$	S_α	*DSC*	$E_\phi^{\max}$	S_α	*DSC*	$E_\phi^{\max}$	S_α	*DSC*	$E_\phi^{\max}$	S_α	*DSC*	$E_\phi^{\max}$	S_α
Source Only (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-continual (ICLR 2021) [53]	72.72	82.99	79.19	69.41	80.09	79.10	13.38	36.09	51.23	73.70	83.33	82.72	57.30	70.62	73.06
CoTTA (CVPR 2022) [54]	76.29	85.31	80.62	66.58	76.73	79.71	71.29	83.50	80.12	70.62	79.72	80.56	71.10	81.34	81.07
DLTTA (TMI 2022) [57]	75.52	84.69	81.66	66.77	77.21	79.34	63.75	78.79	75.55	70.79	81.14	83.32	69.18	80.46	80.02
DUA (CVPR 2022) [35]	78.79	87.14	83.93	69.13	80.62	79.34	74.66	84.96	82.07	86.63	93.62	90.06	77.30	86.58	83.77
SAR (ICLR 2023) [39]	76.48	85.89	81.41	66.45	77.38	75.08	71.46	83.23	79.40	70.41	80.11	81.07	71.20	81.65	81.08
DomainAdaptor (CVPR 2023) [59]	77.48	86.31	82.40	70.82	81.76	80.58	71.96	83.03	79.97	76.89	85.89	84.45	74.29	84.24	81.93
VPTTA [6]	80.65	88.62	84.78	76.94	87.64	84.10	76.48	86.56	83.04	86.37	93.54	89.87	80.11	89.09	85.45
TanGo (Ours)	**83.19**	**90.62**	**86.10**	**81.12**	**90.44**	**85.93**	**80.65**	**89.01**	**85.31**	**86.92**	**93.80**	**90.16**	**82.97**	**90.97**	**86.80**

Table 11. Parameter sensitivity analysis by varying parameter λ_1, λ_2 and λ_3 on **fundus** and **polyp** benchmarks. Using Dice coefficient % as the evaluation metric to assess performance.

λ_1	0.10	0.20	0.50	0.80	1.00
fundus	80.79	81.28	**81.92**	81.63	81.29
polyp	79.82	80.32	**80.91**	80.72	80.63
λ_2	0.20	0.50	0.60	0.80	1.00
fundus	84.26	84.72	**85.23**	83.97	83.12
polyp	82.63	82.71	**83.70**	83.12	83.65
λ_3	0.20	0.50	0.60	0.80	1.00
fundus	84.91	**85.23**	83.17	82.83	82.14
polyp	83.45	**83.70**	83.01	82.81	82.16

Table 12. Performance comparisons of Random noise and the distribution on polyp tasks. Using Dice coefficient % as the evaluation metric to assess performance.

	w/ $\mathcal{R}$ (ours)	*w/o* $\mathcal{R}$	Gaussian	Uniform (ours)
Dice ↑	**80.91**	**79.52**	80.42	**80.91**

C.4. Ablation studies on the loss functions used in IASC

We also present the results for detailed weight settings for loss terms in Tab. 11. Based on these results, our TanGo method is not particularly sensitive to the hyperparameters for the loss terms. In this work, we select a set of hyperparameters $[\lambda_1, \lambda_2, \lambda_3]$ = [0.5, 0.6, 0.5] that exhibit the best performance as the default settings.

C.5. Ablation Study of Random noise

In this subsection, we analyze the role of the random noise term $\mathcal{R}$ in Eq. (4) through an ablation study. In Tab. 12, the second and third columns report the results of the variants without and with the noise term. The performance gap between the two settings is modest, yet the variant that includes $\mathcal{R}$ consistently delivers slightly higher accuracy. This observation suggests that injecting noise into the source model features, after they have been augmented with the low frequency components from the vision foundation model, encourages the model to rely more on domain invariant phase information. As a result, the source model forms more stable and transferable representations, which further enhances its generalization ability. A further comparison between the fourth and fifth columns shows that noise drawn from a uniform distribution leads to better performance than Gaussian noise. This indicates that uniform perturbations introduce a more balanced and stable adjustment to the feature space, which appears to better support the learning of domain invariant representations.

Table 13. Ablation study of different VFMs in DSCL with the polyp tasks. The best results in each column are highlighted in **bold**. Using Dice coefficient % as the evaluation metric to assess performance.

Methods	Domain A	Domain B	Domain C	Domain D	Average ↑
	Dice Score Metric (DSC)				
VPTTA	81.00	76.87	77.58	86.39	80.46
TanGo + DINOv2	83.59	82.31	81.42	87.39	83.68
TanGo + DINOv3	**84.83**	**82.96**	**82.72**	**88.36**	**84.72**
Ours	83.95	82.44	81.11	87.29	83.70

C.6. Effect of Different Vision Foundation Model Backbones in DSCL

In this subsection, we present an ablation study on the choice of vision foundation model backbone within DSCL (e.g. DINOv2 [40] and DINOv3 [48]). To more thoroughly validate our method, we add comparisons that use the DINOv2 and DINOv3 backbones, as reported in Tab. 13. The results show that combining DSCL with either DINOv2 or DINOv3 yields a clear improvement compared with sota method VPPTA [6] in segmentation performance. In particular, dinov3 combined with DSCL attains the best results, which we attribute to DINOv3 being trained on a larger and more diverse dataset and to its improved pretraining strategy. These findings demonstrate that our approach can effectively transfer the generalization capabilities of different types of vision foundation models into the source model, thereby enriching the source domain representation and further enhancing the model's capacity for continual test adaptation.