Test-Time Domain Generalization via Universe Learning: A Multi-Graph Matching Approach for Medical Image Segmentation

Supplementary Material

A. Theoretical Analysis

Lemma 1 (Cycle-consistency, Universe Matching)**.** *Given a set of pairwise (partial) matching matrices* $\{\mathbf{X}_{ij}\}_{i,j=1}^{m}$*, it is cycle-consistent iff there exists a collection of universe matching matrices* $\{U_i \in \mathbb{U}_{n_i,d}\}_{i=1}^{m}$ *such that for each graph pair* $(\mathcal{G}_i, \mathcal{G}_j)$*, we have*

$$\mathbf{X}_{ij} = U_i U_j^{\mathsf{T}}.$$

Proof. We need to prove that the matching matrices $\{\mathbf{X}_{ij}\}_{i,j=1}^{m}$ satisfy cycle-consistency, which means that:

$$\mathbf{X}_{ij}\mathbf{X}_{jk} = \mathbf{X}_{ik}. \forall i, j, k \in [m] \tag{1}$$

By the assumption of the *lemma 1*, there exists a collection of universe matching matrices $\{U_i\}_{i=1}^{m}$ such that for each pair $(\mathcal{G}_i, \mathcal{G}_j)$,

$$\mathbf{X}_{ij} = U_i U_j^{\mathsf{T}}. \tag{2}$$

Therefore, we have

$$\mathbf{X}_{ij} = U_i U_j^{\mathsf{T}}, \quad \mathbf{X}_{jk} = U_j U_k^{\mathsf{T}}, \quad \mathbf{X}_{ik} = U_i U_k^{\mathsf{T}}. \tag{3}$$

Now we compute $\mathbf{X}_{ij}\mathbf{X}_{jk}$:

$$\begin{aligned} \mathbf{X}_{ij}\mathbf{X}_{jk} &= (U_i U_j^{\mathsf{T}})(U_j U_k^{\mathsf{T}}) \\ &= U_i (U_j^{\mathsf{T}} U_j) U_k^{\mathsf{T}}. \end{aligned} \tag{4}$$

Assume that each matrix U_i satisfies $U_i^{\mathsf{T}} U_i = I$ (i.e., U_i is an orthogonal matrix or has unit inner product property). Thus, Eq. (4) simplifies further to:

$$\mathbf{X}_{ij}\mathbf{X}_{jk} = U_i I U_k^{\mathsf{T}} = U_i U_k^{\mathsf{T}} = \mathbf{X}_{ik}. \tag{5}$$

This shows that the matching matrices $\{\mathbf{X}_{ij}\}_{i,j=1}^{m}$ satisfy cycle-consistency. □

B. Algorithm Pipeline

Algorithms 1 and 3 outline the procedures for the source training phase and the test-time adaptation phase, respectively, while Algorithm 2 provides a detailed explanation of the HiPPI [1] method used in Algorithm 1.

C. Additional Experiments

C1. Single Source DG in Retinal Fundus

For the retinal fundus segmentation task, we conducted single-source domain generalization experiments. Unlike the experiments described in the main text, this setup simulates a more realistic scenario where test data may originate from arbitrarily complex real-world distributions, i.e., mixed distribution shifts. Specifically, data from one site was selected and split $8:2$ into training and validation sets ($S = 1$), while the remaining sites ($T = |\mathcal{D}_s \cup \mathcal{D}_t| - 1$) were shuffled and used entirely as the testing dataset. Notably, all models encountered these target domains for the first time during testing.

As shown in Table 1, our approach achieved SOTA performance across all five transfer experiments for the DSC metric. In the average results, we outperformed the second-best method (DeY-Net [17]) by 4.85%, 1.78%, and 1.81% in the DSC, E_ϕ^{max}, and S_α, respectively. These results validate the effectiveness of our method for medical image segmentation tasks.

The retinal fundus dataset is characterized by significant low-level visual differences and features segmentation targets that are not singular, often exhibiting overlapping and fixed structures. We attribute our superior performance to the comprehensive learning of the morphological knowledge of organs. This enables our method to robustly distinguish organ instances and their shape features—a domain-invariant property—even under severe domain shifts that degrade the performance of other methods.

C2. Multi Source DG in MRI Prostate

Datasets. We conducted experiments on the prostate segmentation task using T2-weighted MRI scans collected from six different clinical centers, denoted as Domain RUNMC, BMC, I2CVB, UCL, BIDMC, and HK. These centers are sourced from three publicly available datasets: NCI-ISBI13 [11], I2CVB [7], and PROMISE12 [9].

Implementation Details. We followed the data preprocessing pipeline of [19] to ensure consistency. Specifically, we used 30 labeled cases from RUNMC as the source dataset and evaluated the model on 30, 19, 13, 12, and 12 unlabeled cases from the five remaining clinical sites. Each MRI axial slice was resized to 384×384 pixels and normalized to have zero mean and unit variance. Before normalization, we clipped the 5%–95% intensity range of the histograms to reduce outlier influence.

For feature extraction, we employed a ResNet-50 backbone pre-trained on ImageNet. During both the source model training and test-time adaptation (TTA) stages, we maintained a batch size of 8. Given that edge precision is crucial in MRI prostate segmentation, and considering the complex shape variations of the prostate, we selected Dice Score

Table 1. Single source domain generalization in the retinal fundus segmentation. The performance (mean ± standard deviation) of three trials for our method and eight SOTA methods. "A $\rightarrow \{B, C, D, E\}$" represents models trained on Site A and tested on the mixed distribution of Sites B-E, and similar for others. Best results are colored as red.

Methods	A $\rightarrow \{B, C, D, E\}$			B $\rightarrow \{A, C, D, E\}$			C $\rightarrow \{A, B, D, E\}$			D $\rightarrow \{A, B, C, E\}$			E $\rightarrow \{A, B, C, D\}$			Average		
	DSC	E_ϕ^{max}	S_α	*DSC*	E_ϕ^{max}	S_α	*DSC*	E_ϕ^{max}	S_α	*DSC*	E_ϕ^{max}	S_α	*DSC*	E_ϕ^{max}	S_α	*DSC* ↑	E_ϕ^{max} ↑	S_α ↑
No Adapt (U-Net [14])	70.60±10.01	86.92±1.17	80.36±0.88	77.08±6.90	91.58±0.84	85.44±0.99	66.24±8.45	86.49±0.77	80.01±0.91	71.21±9.47	82.90±1.48	80.11±0.88	72.26±7.60	86.51±1.02	86.05±0.68	71.47	86.88	82.39
TASD (AAAI'22) [10]	79.89±5.91	93.26±0.21	87.12±0.13	82.63±3.24	93.20±0.25	86.00±0.10	78.03±4.29	92.47±0.14	86.71±0.09	76.30±7.81	86.09±0.15	80.94±0.11	79.99±1.29	93.24±0.10	87.08±0.08	79.36	91.65	85.57
DLTTA (TMI'22) [18]	74.96±7.20	89.24±0.25	84.02±0.11	78.27±5.66	92.40±0.40	85.36±0.11	75.84±5.14	90.96±0.20	84.11±0.10	65.55±9.35	84.80±0.18	78.33±0.08	71.68±4.99	87.79±0.17	85.13±0.09	73.26	89.03	83.39
SAR (ICLR'23) [12]	74.20±6.09	89.07±0.33	83.64±0.10	80.34±2.86	92.70±0.41	85.95±0.12	72.58±4.46	91.20±0.20	83.47±0.09	70.30±8.98	85.10±1.09	79.42±0.72	70.31±5.77	88.56±0.81	86.79±0.30	73.54	89.32	83.85
DomainAdaptor (CVPR'23) [20]	77.23±3.97	90.22±0.40	84.20±0.11	76.41±4.28	91.80±0.31	85.73±0.12	70.17±8.01	91.32±0.24	83.50±0.09	67.39±9.82	84.16±0.21	78.02±0.08	76.97±4.59	89.30±0.13	85.46±0.08	73.63	89.36	83.38
DeY-Net (WACV'24) [17]	80.03±8.31	94.42±0.20	86.35±0.84	84.30±7.09	94.25±0.23	87.16±0.47	80.32±7.85	93.40±0.32	88.41±0.33	78.67±5.31	86.12±0.78	80.45±0.30	76.81±3.79	90.09±0.55	86.30±0.25	80.02	91.65	85.73
VPTTA (CVPR'24) [3]	73.57±6.60	92.68±0.03	84.14±0.01	78.21±2.40	94.07±0.09	86.16±0.01	69.26±4.29	92.78±0.08	82.66±0.02	60.11±8.05	85.18±0.10	76.24±0.03	72.58±5.21	91.16±0.13	84.74±0.04	70.74	91.17	82.78
NC-TTT (CVPR'24) [13]	78.21±2.74	93.87±0.25	85.49±0.11	82.13±3.30	93.19±0.29	86.78±0.08	77.50±5.29	91.99±0.14	84.08±0.03	74.14±3.50	87.53±0.25	80.56±0.11	80.53±1.08	92.73±0.10	85.81±0.07	78.50	91.86	84.54
Ours	85.25±2.33	94.68±0.09	88.52±0.13	85.34±3.08	93.18±0.20	86.83±0.11	86.19±1.99	94.57±0.20	89.24±0.13	81.52±4.25	91.20±0.30	84.53±0.28	86.08±3.08	94.60±0.23	88.58±0.11	84.87	93.64	87.54

(DSC) and Hausdorff Distance (HD95) as the primary evaluation metrics to provide a comprehensive performance assessment.

Experimental Results. The MRI prostate segmentation results are presented in Table 2, where we compare our method against several SOTA approaches, including the latest TTA segmentation method, PASS [19]. As shown in the results, the performance of existing TTA methods remains relatively close across both DSC and HD95 metrics. While PASS exhibits strong segmentation performance, our method surpasses it with a 1.69% improvement in DSC, demonstrating its effectiveness. Given the inherent challenges of MRI prostate segmentation, characterized by diverse imaging modalities and complex morphological variations, our results highlight the robust generalization capability of our approach. Nonetheless, further enhancing edge precision remains an important focus for our future work.

C3. Natural Image Classification

Datasets. For natural image classification tasks, we selected two benchmark datasets: PACS [8] and VLCS [4], which are widely used in domain generalization and test-time adaptation studies. The PACS [8] dataset consists of large images spanning 7 classes evenly distributed across 4 domains, i.e. A (Art), C (Cartoons), P (Photos), and S (Sketches), with a total of 9,991 images. The VLCS [4] dataset comprises 10,729 images across 5 classes (bird, car, chair, dog, and person), evenly distributed across 4 domains: C (Caltech101), L (LabelMe), S (SUN09), and V (VOC2007).

Source model training. For all experiments, we employed an ImageNet-pretrained ResNet-50 [6] as the feature extractor, with an MLP layer provided by the DomainBed [5] benchmark serving as the classifier. We used the SGD optimizer with a learning rate of 1×10^{-5}. The batch size was set to 32, and training was conducted for 10,000 iterations. All images were resized to 224×224, and data augmentation techniques—including random cropping, flipping, color jittering, and intensity adjustments—were applied during source training.

Implementation details of test-time adaptation setup. We evaluated our framework against six methods (i.e. Empirical Risk Minimization (ERM) [15], DomainAdaptor [20], ITTA [2], VPTTA [3], NC-TTT [13]) under fair comparison conditions, following the leave-one-out training strategy using the publicly available DomainBed [5] framework. For deploying our framework at the test-time phase, we employed SGD with a learning rate of 0.005, a batch size of 16, and a universe size of $d = 60$. Notably, as all images in the natural image classification task contain only a single instance class, the class-wise similarity matrix described in Section 3.2 of the main text was not utilized.

Experimental Results. The classification results across different domains for natural images are presented in Tables 3 and 4. While the ERM method shows strong performance compared to existing approaches, our method achieves higher classification accuracy, surpassing ERM by 3.01% on the PACS dataset and 2.49% on the VLCS dataset. Additionally, our approach demonstrates competitive performance against state-of-the-art test-time adaptation methods designed for natural images. Natural images present greater challenges compared to medical images due to the lack of consistent morphological priors typically observed in the latter. However, unlike segmentation tasks, classification does not require determining the specific class of every pixel in an image. Our framework's graph construction effectively captures spatial correspondences for each instance, further enhancing its performance.

C4. Additional Visualization

We conducted additional visualization experiments, with segmentation results for retinal fundus and polyp images shown in Figures 1 and 2, respectively. Each row represents images from a distinct domain (Site), and we ensured that the model performing inference had not encountered images from that domain before.

Retinal fundus segmentation, in particular, presents a challenge due to the presence of two overlapping substructures. Lower clarity and contrast in images (e.g., rows 1 and 2 of Figure 1) further complicate the model's ability to accurately differentiate and segment these structures. By incorporating morphological priors of the organ within a multi-graph matching network, our method effectively learns robust substructure representations while minimizing domain-related noise. This approach overcomes issues like repeated, missing, or blurred edge pixels commonly seen in other methods, providing a more precise segmentation outcome.

Table 2. Test-time domain generalization results on the MRI prostate datasets. The performance (mean ± standard deviation) of three trials for our method and six SOTA methods. Best results are colored as red.

Methods	BMC		I2CVB		UCL		BIDMC		HK		Avg.	
	DSC	HD95	*DSC*	HD95	*DSC*	HD95	*DSC*	HD95	*DSC*	HD95	*DSC* ↑	HD95 ↓
No Adapt	74.30±5.31	16.08±12.41	66.47±13.50	37.16±18.24	75.28±6.20	16.77±12.10	52.08±7.71	50.09±20.85	80.51±9.35	8.79±9.06	69.72±8.29	25.77±15.41
TENT (ICLR'21) [16]	77.45±3.79	12.09±9.88	69.10±10.47	30.78±19.22	79.69±4.81	14.71±11.01	52.01±6.80	42.63±10.13	84.58±2.73	4.07±5.38	72.56±4.26	20.85±13.64
TASD (AAAI'22) [10]	76.28±2.35	15.11±15.17	68.30±7.88	31.43±24.10	80.25±3.54	10.59±16.39	56.08±3.82	51.90±24.82	81.09±1.79	4.26±4.16	72.40±5.72	22.65±17.53
SAR (ICLR'23) [12]	77.24±4.26	20.48±10.12	68.99±8.27	49.07±15.66	79.27±8.48	18.03±5.89	50.81±10.60	54.35±19.31	85.40±3.08	3.87±3.55	72.34±4.80	29.16±16.28
DomainAdaptor (CVPR'23) [20]	76.49±2.59	19.27±8.13	69.07±9.14	32.57±10.40	80.41±5.08	16.24±9.88	49.99±14.28	48.40±10.28	85.20±1.90	3.25±6.94	72.23±6.82	23.94±14.55
VPTTA (CVPR'24) [3]	77.42±4.38	12.93±7.09	70.25±5.18	30.01±13.68	82.07±6.27	13.28±18.09	57.49±8.46	40.11±12.05	83.27±2.96	3.40±5.45	74.10±4.79	19.94±12.99
PASS (TMI'24) [19]	80.07±7.14	10.50±9.57	71.41±6.28	28.26±9.97	84.39±8.81	10.68±12.27	57.27±11.48	36.94±16.43	84.88±3.71	3.03±5.05	75.60±5.13	17.88±12.14
Ours	79.63±4.71	9.99±11.10	74.09±9.80	25.70±13.07	86.30±7.25	11.08±17.47	60.33±13.59	39.52±14.20	86.12±2.08	2.84±8.08	77.29±3.98	17.82±11.06

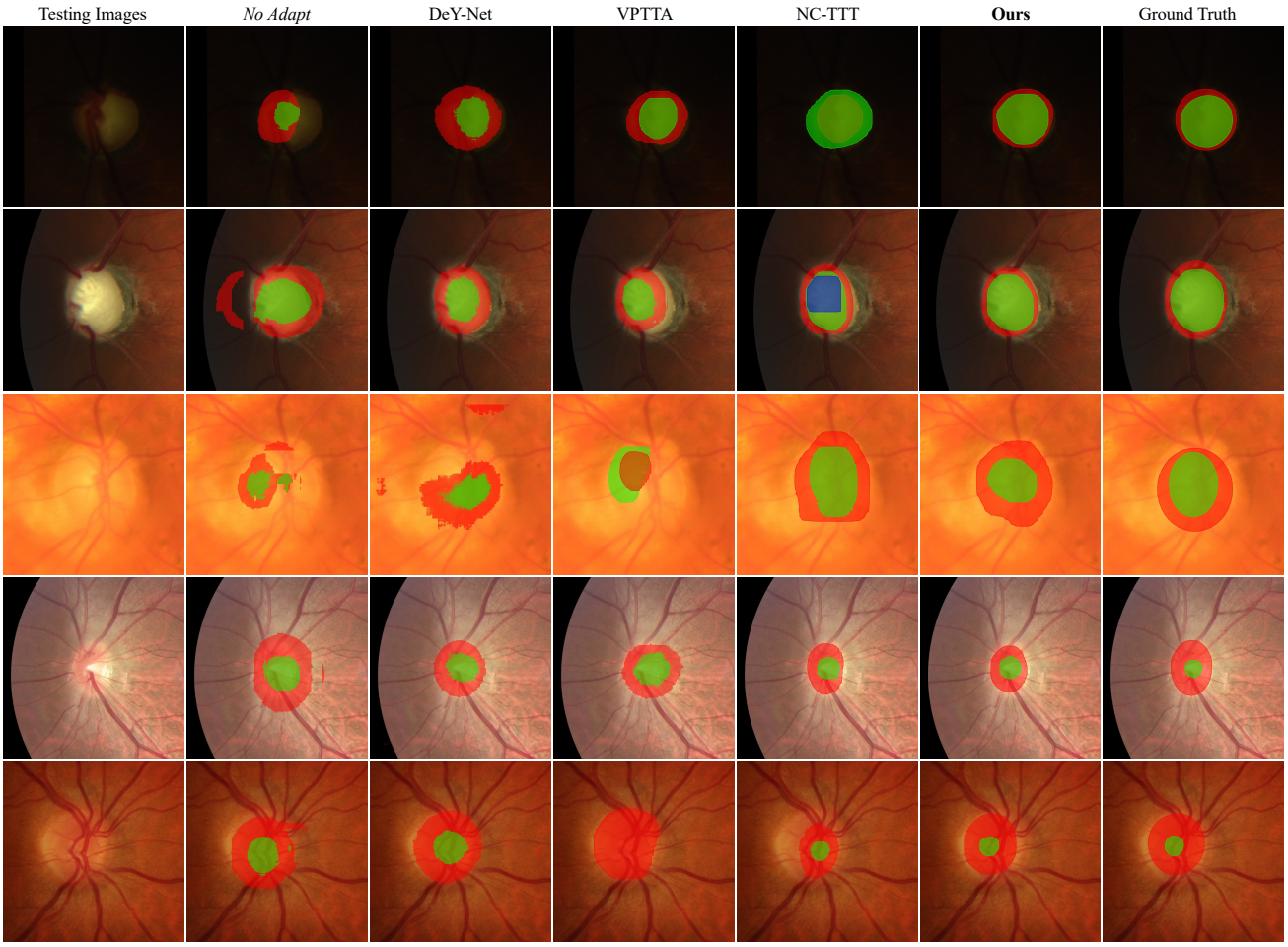


Figure 1. Visualization comparison of segmentation results for the *No Adapt* baseline, DeY-Net [17], VPTTA [3], NC-TTT [13], and our method in retinal fundus segmentation. The five rows from top to bottom display the final segmentation results for tests conducted on Sites A to E. Different colors represent the segmentation instances of different classes identified by the network.

The segmentation of polyps presents a greater challenge than that of retinal fundus imaging due to the highly variable appearance, with marked differences in shape, size, and color across domains. This variability demands precise, pixel-level classification from the network. Furthermore, we have not designated polyp segmentation as a single-object task; the model independently classifies and segments multiple classes during testing, using different colors to distinguish each segmented object in the visualization. As illustrated in Figure 2, the masks generated by our method are in close alignment with expert annotations and effectively avoid pixel misclassification into different categories, a common issue in other methods.

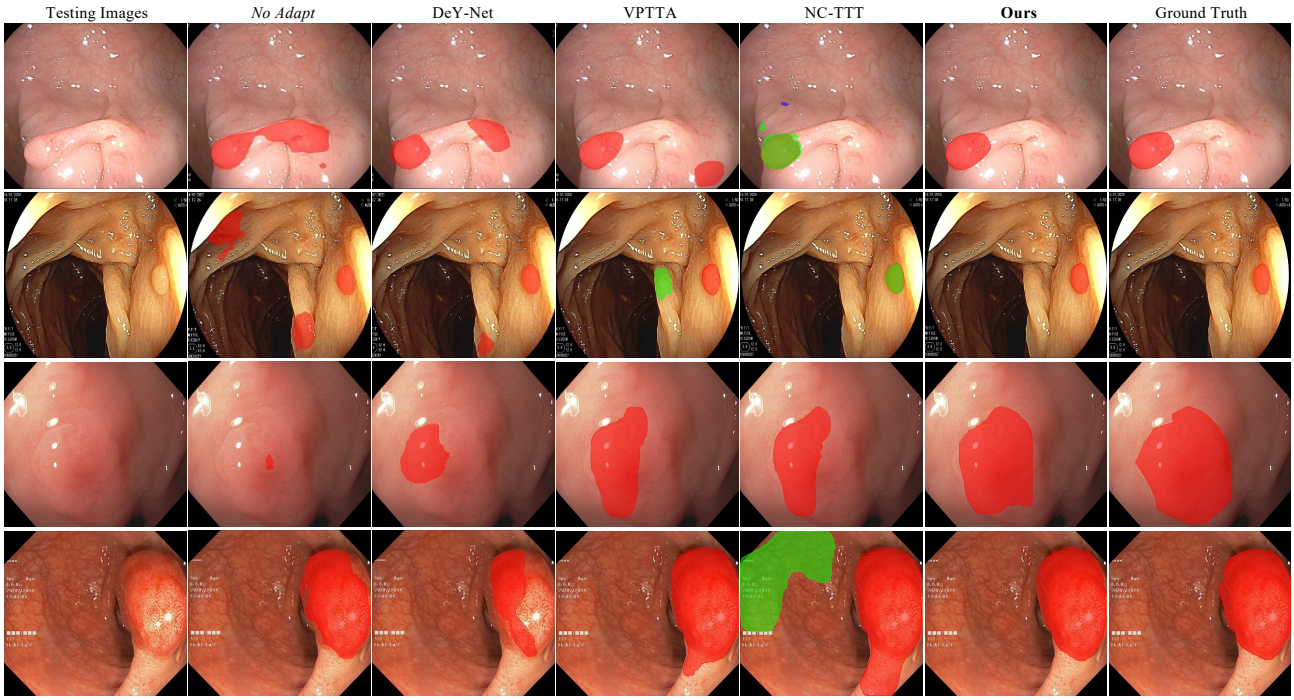


Figure 2. Visualization comparison of segmentation results for the *No Adapt* baseline, DeY-Net [17], VPTTA [3], NC-TTT [13], and our method in polyp segmentation. The four rows from top to bottom display the final segmentation results for tests conducted on Sites A to D. Different colors represent the segmentation instances of different classes identified by the network.

Table 3. Test-time domain generalization results on the PACS [8] dataset using a ResNet-50 backbone. Each column (A, C, P, S) indicates the domain used as the test set, while the remaining domains are used for training. The best results are highlighted in red.

Method	A	C	P	S	Avg.
ERM [15]	84.07	80.21	97.06	81.99	85.83
TENT (ICLR'21) [16]	82.34	78.63	97.93	82.72	85.40
DomainAdaptor (CVPR'23) [20]	86.15	82.02	98.40	84.38	88.45
ITTA (CVPR'23) [2]	85.63	84.30	97.27	84.09	87.82
VPTTA (CVPR'24) [3]	86.50	83.77	97.09	85.10	88.12
NC-TTT (CVPR'24) [13]	83.81	80.44	96.53	82.36	85.79
Ours	85.08	83.93	98.61	87.76	88.84

Table 4. Test-time domain generalization results on the VLCS [4] dataset using a ResNet-50 backbone. Each column (C, L, S, V) indicates the domain used as the test set, while the remaining domains are used for training. The best results are highlighted in red.

Method	C	L	S	V	Avg.
ERM [15]	97.63	64.20	70.39	74.41	76.66
TENT (ICLR'21) [16]	96.88	64.46	71.07	73.52	76.48
DomainAdaptor (CVPR'23) [20]	98.69	69.18	73.66	76.01	79.39
ITTA (CVPR'23) [2]	97.30	66.09	72.31	75.10	77.70
VPTTA (CVPR'24) [3]	97.25	67.69	71.78	75.22	77.98
NC-TTT (CVPR'24) [13]	96.72	65.58	73.04	76.83	78.04
Ours	98.49	68.72	74.12	75.30	79.15

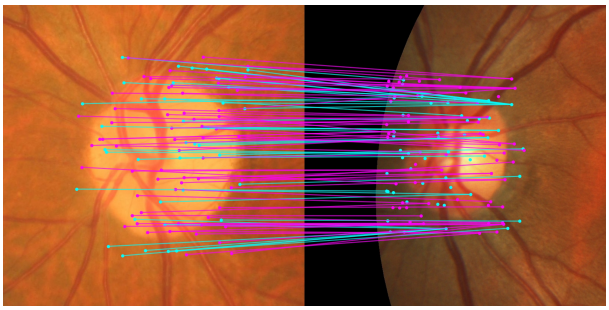
D. Additional Analysis

D1. Effectiveness of the class-wise similarity matrix

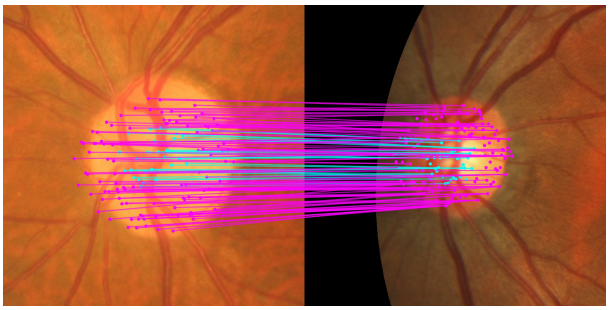
The class-wise similarity matrix $\mathbf{W}$ is introduced to mitigate category confusion in graphs caused by nodes belonging to different classes. Such confusion often results in mismatches, semantic deviations, and redundant computations. By reordering the adjacency matrix based on the labels Y_i of each node $\mathcal{V}_i$, our method strengthens the capacity to identify and learn class-specific information during the source training phase. To validate the above perspective, we conducted experiments comparing the final TTA segmentation results with and without $\mathbf{W}$ (denoted as with $\mathbf{W}$ and *w/o* $\mathbf{W}$). As illustrated in Figure 4, *w/o* $\mathbf{W}$ results in a measurable decline in DSC performance. Furthermore, we visualized the effect of *w/o* $\mathbf{W}$ in multi-object segmentation scenarios, as shown in Figure 5. While the masks generated by the model closely align with the ground truth, the model misclassified the categories of two segmented instances.

D2. Effectiveness of Morphological Priors

We visualized cross-site pairing without morphological priors, as shown in Fig. 3(a), and compared it with the results obtained after incorporating priors, as shown in Fig. 3(b). Without priors, the graph nodes were not correctly sampled within the corresponding organs, leading to mismatches. By



(a) Visualization of graph matching **without** morphological priors



(b) Visualization of graph matching **with** morphological priors

Figure 3. Visualization of graph pair matching.

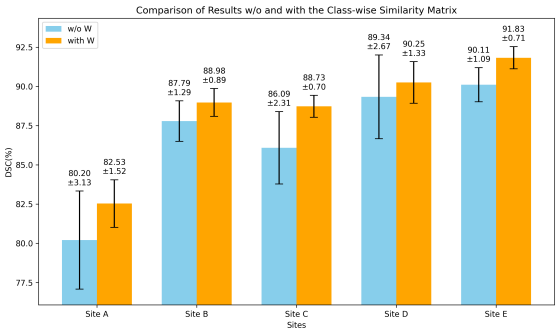


Figure 4. Ablation study on the impact of the Class-wise Similarity Matrix $\mathbf{W}$ in retinal fundus segmentation: comparison of results with and without (*w/o*) $\mathbf{W}$.

introducing priors, this issue was effectively resolved, and multigraph matching ensured more stable pairing across multiple domains. For a quantitative evaluation of the impact of without priors, please refer to Table 4 in the main text.

D3. Impact of Batch Size on Segmentation

As shown in Table 5, increasing the number of simultaneously matched graphs leads to a significant increase in both FLOPs and inference time, while the improvement in segmentation quality remains marginal. To achieve a balance between segmentation performance and computational efficiency, we set the mini-batch size to 4 during the TTA phase in retinal fundus datasets. However, this is a tunable hyper-

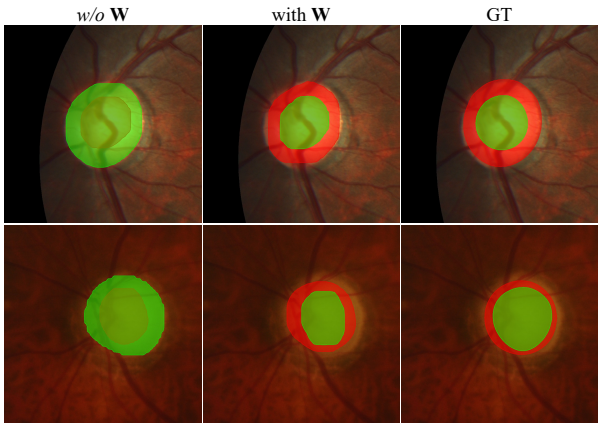


Figure 5. Visualization of segmentation results with and without (*w/o*) the Class-wise Similarity Matrix $\mathbf{W}$.

Table 5. Ablation study on batch size during TTA for retinal fundus segmentation. "Avg. DSC" represents the average DSC across the five sites, while "time" indicates the inference time per image.

Batch Size	Avg. DSC	FLOPs (G)	time (s/img)
2	85.20	3.012	0.277
4	88.46	4.255	0.392
8	88.93	20.43	0.780
16	89.15	80.96	1.831
32	88.31	223.1	3.715

parameter rather than a fixed value, as it depends on factors such as the size of the segmented objects and the input image resolution. Empirically, we find that a mini-batch size between 4 and 8 provides an optimal trade-off.

E. Limitations

Unlike mainstream TTA methods that update only the Batch Normalization layers, our approach optimizes all network parameters during test time, achieving superior segmentation performance. However, the increased computational overhead limits deployment on portable devices, making efficiency optimization a key focus for future work.

In our experiments, we also observed that when both large and small organs are present, the model tends to perform better on larger organs while often overlooking smaller ones. This is due to the uniform sampling of foreground nodes, which can lead to diminished segmentation accuracy for small targets. To address this, we plan to incorporate stronger regularization in future work to better guide the sampling and learning of small structures.

Our method is well-suited for medical imaging compared to natural image tasks. In natural images, objects often exhibit significant variation due to intrinsic properties, motion, and state changes. In contrast, organs in medical images remain relatively stable, which aligns well with the prior knowledge.

Algorithm 1 Source Training Phase per Mini-Batch

Output: $\mathcal{L}_{overall}$: The overall loss for training the segmentation network;
$\mathcal{U}$: the pre-trained universe embeddings integrate morphological priors;
Input: $\{x_i \in \mathbb{R}^{H\times W\times C}\}_{i=1}^{m}$: A batch of m images from one or multiple domains;
$\{y_i \in [0, 255] \cap \mathbb{Z}\}_{i=1}^{m}$: The ground truth masks corresponding to the input images;
$E(\cdot)$: Feature extractor (ResNet-50);
$S(\cdot)$: Segmentation head;
N: The total classes number of segmentation organ;
$\mathcal{U}$: Learnable universe embeddings;
(1) Segmentation Network Training.
Get the visual feature maps: $f_i \leftarrow E(x_i)$.
Get the predict segmentation masks: $\hat{y}_i \leftarrow S(f_i)$.
Get the supervised loss: $\mathcal{L}_{sup} \leftarrow \text{CE}(\hat{y}_i, y_i)$, where CE is Cross Entropy Loss.
(2) Graph Construction.
for each $i \in [1, m]$ **do**
  for each object $n \in [1, N]$ **do**
    $\{f_{i,k}^n\}_{k=1}^{K} \leftarrow$ Extract feature maps for object n from layers 1 to K based on f_i and y_i.
  end for
  Obtain object-specific features: $\{F_i^n\}_{n=1}^{N} \leftarrow Concat(f_{i,1}^n, \ldots, f_{i,K}^n)$ for each n in N.
  Build features of nodes and corresponding labels: $\{\mathcal{V}_i \in \mathbb{R}^{n_i\times h}, Y_i \in \mathbb{Z}^{n_i}\}_{i=1}^{m} \leftarrow \phi(\{F_i^n\}_{n=1}^{N})$, where ϕ is the spatially-uniform sampling, and n_i is the total number of nodes for x_i.
  $\mathcal{G}_i = (\mathcal{V}_i, \mathcal{A}_i)$, the weighted adjacency matrix $\mathcal{A}_i$ is obtained from Eq. (5).
end for
(3) Formulation of universe embeddings.
if $\mathcal{U}$ is not initialized **then**
  $\mathcal{U} = 1/d + 10^{-3}z$, where $z \sim N(0, 1)$.
end if
Universe matching matrices: $\mathbf{U} = [U_1^\mathsf{T}, \cdots, U_m^\mathsf{T}]^\mathsf{T}$, where $U_i = Sinkhorn(\mathcal{V}_i\,\mathcal{U}^\mathsf{T}, \tau) \in \mathbb{U}_{n_i d}$, d is the *universe size*.
Block-diagonal multi-adjacency matrix: $\mathbf{A} = diag(\mathcal{A}_1, \cdots, \mathcal{A}_m)$.
Compute the class-aware similarity matrix: $\tilde{\mathbf{A}} = \mathbf{W}^\mathsf{T}\mathbf{A}\mathbf{W}$, where $\mathbf{W} = [W_{ij}]_{ij}$, and $W_{ij} = Y_i Y_j^\mathsf{T}$.
HiPPI solving for stable convergence of $\mathbf{U}$ as in Eqs. (6-8).
Update $\mathcal{U}$ with $L(\mathcal{U})$ in Eq. (9).
Overall Loss of Source Training.
$\mathcal{L}_{overall} = \mathcal{L}_{sup} + L(\mathcal{U})$.

Algorithm 2 Higher-order Projected Power Iteration (HiPPI) [1]

Output: Cycle-consistent universe-matching $\mathbf{U}_t$.
Input: W: multi-graph similarity matrix;
$\mathbf{U}_0$: initial universe-matching $\mathbf{U}_0 \in \mathbb{U}_{n_i d}$;
Initialise: $t \leftarrow 0$, proj $\leftarrow Sinkhorn$.
repeat
  $V_t \leftarrow W\mathbf{U}_t\mathbf{U}_t^\mathsf{T}W\mathbf{U}_t$.
  $\mathbf{U}_{t+1} \leftarrow \text{proj}(V_t)$.
  $t \leftarrow t + 1$.
until $||U_i^t - U_i^{t-1}|| < 10^{-5}$

References

[1] Florian Bernard, Johan Thunberg, Paul Swoboda, and Christian Theobalt. Hippi: Higher-order projected power iterations for scalable multi-matching. In *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, pages 10284–10293, 2019. 1, 6

[2] Liang Chen, Yong Zhang, Yibing Song, Ying Shan, and Lingqiao Liu. Improved test-time adaptation for domain generalization. In *Proc. IEEE/CVF Comput. Vis. Pattern Recog.*, pages 24172–24182, 2023. 2, 4

[3] Ziyang Chen, Yongsheng Pan, Yiwen Ye, Mengkang Lu, and

Algorithm 3 Test-time Adaptation Phase per Mini-Batch

Output: $\{y_i^* \in \mathbb{R}^{H\times W\times C}\}_{i=1}^m$: The predicted masks of test dataset.
Input: $\{x_i^t \in \mathbb{R}^{H\times W\times C}\}_{i=1}^m$: A batch of m images from test dataset;
$E(\cdot)$: Pre-trained feature extractor (ResNet-50);
$S(\cdot)$: Pre-trained segmentation head;
N: The total classes number of segmentation organ;
$\mathcal{U}$: Pre-trained learnable universe embeddings;
MIter: Max iteration of multi-graph matching.

Initialise: *Iter*$\leftarrow 0$. $V_i \leftarrow 0, \forall i \in [m]$. V_i is the gradient of Eq. (4) with respect to U_i. $\tau \leftarrow 0.05$.
Obtain the visual feature maps: $f_i^t \leftarrow E(x_i^t)$.
Obtain the pseudo segmentation masks: $\hat{y}_i^t \leftarrow S(f_i^t)$.
(1) Graph Construction.
$\mathcal{G}_i^t = (\mathcal{V}_i^t, \mathcal{A}_i^t)$, where $\mathcal{V}_i^t$ and $\mathcal{A}_i$ are obtained in the same manner as in Algorithm 1. (2).
(2) Unsupervised Multi-graph Matching.
Universe matching of $\mathcal{G}_i$: $U_i = Sinkhorn(\mathcal{V}_i^t\,\mathcal{U}^\mathsf{T}, \tau)$.
for $(\mathcal{G}_i, \mathcal{G}_j)$ in $\{\mathcal{G}_1, \cdots, \mathcal{G}_m\}$ **do**
  Affinity matrix $M_{ij} \leftarrow f_{mlp}\{\mathcal{V}_i^t\mathcal{W}_x^t \cdot (\mathcal{V}_j^t\mathcal{W}_y^t)^\mathsf{T}\}$, where $\mathcal{W}_x^t$ and $\mathcal{W}_y^t$ are two learnable linear projection, f_{mlp} is a multi-layer perception (MLP).
  repeat
    $V_i \leftarrow V_i + (\lambda\mathcal{A}_i^t U_i U_j^\mathsf{T}\mathcal{A}_j^t U_j + M_{ij}U_j)$.
    Relax V_i to the space of universe: $U_i \leftarrow sinkhorn(V_i, \tau)$.
  until $\{U_i\}$ *converged* OR *Iter* $>$ *MIter*
end for
Fine-tune the segmentation network (update E and S) using L_{mat} in Eq. (12).
(3) Final Inference Process.
$y_i^* = S(E(x_i))$

Yong Xia. Each test image deserves a specific prompt: Continual test-time adaptation for 2d medical image segmentation. In *Proc. IEEE/CVF Comput. Vis. Pattern Recog.*, pages 11184–11193, 2024. 2, 3, 4

[4] Chen Fang, Ye Xu, and Daniel N Rockmore. Unbiased metric learning: On the utilization of multiple datasets and web images for softening bias. In *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, pages 1657–1664, 2013. 2, 4

[5] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations*, 2021. 2

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. IEEE/CVF Comput. Vis. Pattern Recog.*, pages 770–778, 2016. 2

[7] Guillaume Lemaître, Robert Martí, Jordi Freixenet, Joan C Vilanova, Paul M Walker, and Fabrice Meriaudeau. Computer-aided detection and diagnosis for prostate cancer based on mono and multi-parametric mri: a review. *Computers in Biology and Medicine*, 2015. 1

[8] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, pages 5542–5550, 2017. 2, 4

[9] Geert Litjens, Robert Toth, Wendy Van De Ven, Caroline Hoeks, Sjoerd Kerkstra, Bram Van Ginneken, Graham Vincent, Gwenael Guillard, Neil Birbeck, Jindang Zhang, et al. Evaluation of prostate segmentation algorithms for mri: the promise12 challenge. *Medical Image Analysis*, 2014. 1

[10] 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 *AAAI*, pages 1756–1764, 2022. 2, 3

[11] Bloch Nicholas, Madabhushi Anant, Huisman Henkjan, Freymann John, Kirby Justin, et al. 2013 challenge: Automated segmentation of prostate structures. *The Cancer Imaging Archive*, 2015. 1

[12] Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Zhiquan Wen, Yaofo Chen, Peilin Zhao, and Mingkui Tan. Towards stable test-time adaptation in dynamic wild world. *ICLR*, 2023. 2, 3

[13] David Osowiechi, Gustavo A Vargas Hakim, Mehrdad Noori, Milad Cheraghalikhani, Ali Bahri, Moslem Yazdanpanah, Ismail Ben Ayed, and Christian Desrosiers. Nc-ttt: A noise constrastive approach for test-time training. In *Proc. IEEE/CVF Comput. Vis. Pattern Recog.*, pages 6078–6086, 2024. 2, 3, 4

[14] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241. Springer, 2015. 2

[15] Vladimir Vapnik. Statistical learning theory. *John Wiley & Sons google schola*, 2:831–842, 1998. 2, 4

[16] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Ol-

shausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *ICLR*, 2021. 3, 4

[17] Ruxue Wen, Hangjie Yuan, Dong Ni, Wenbo Xiao, and Yaoyao Wu. From denoising training to test-time adaptation: Enhancing domain generalization for medical image segmentation. In *Proc. IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 464–474, 2024. 1, 2, 3, 4

[18] 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. 2

[19] 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. 1, 2, 3

[20] Jian Zhang, Lei Qi, Yinghuan Shi, and Yang Gao. Domainadaptor: A novel approach to test-time adaptation. In *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, pages 18971–18981, 2023. 2, 3, 4