Tri-KA: Tri-level knowledge anchoring test-time adaptation for source-free cross-site MRI rectal cancer segmentation

Wang Bo[1,2,3], Ting Xue[4], Leyang Pan[4], Dingfu Huang[1,2], Zaiyi Liu[5,6], Shiyuan Liu[4] ✉, and Shaohua Kevin Zhou[1,2,7,8,9] ✉

[1] School of Biomedical Engineering, Division of Life Sciences and Medicine, University of Science and Technology of China (USTC), Hefei, China
[2] Medical Imaging, Robotics, Analytic Computing & Learning (MIRACLE) Lab, YRD-RIGHT, USTC Suzhou Institute for Advanced Research, Suzhou, China
[3] China Ocean Aviation Group Limited, Beijing, China
[4] Department of Radiology, Second Affiliated Hospital of Navy Medical University, Shanghai, China
[5] Department of Radiology, Guangdong Provincial People's Hospital, Guangdong Academy of Medical Sciences, Southern Medical University, Guangzhou, China
[6] Guangdong Provincial Key Laboratory of Artificial Intelligence in Medical Image Analysis and Application, Guangzhou, China
[7] Jiangsu Provincial Key Lab of Multimodal Digital Twin Technology, Suzhou, China
[8] Biomedical Basic Research Center (BBRC) of Jiangsu Province, Suzhou, China
[9] State Key Laboratory of Precision and Intelligent Chemistry, Hefei, China

Abstract. Cross-site generalization in rectal cancer MRI segmentation is challenged by severe domain shifts across different scanners and imaging protocols. While Test-Time Adaptation (TTA) mitigates this by dynamically updating models during inference, generic TTA methods often cause semantic drift and disrupt the optimized architectures of 3D medical networks. To address this under strict patient privacy constraints, we propose Tri-KA, a Tri-level Knowledge Anchoring TTA framework. Tri-KA safely distills and transmits privacy-preserving source priors across three dimensions without altering the base network structure or requiring raw source data. Specifically, (1) at the input level, an energy-calibrated Fourier frequency module aligns global image contrast to prevent initial gradient instability; (2) at the feature level, a deep-to-shallow prototype guidance mechanism freezes deep layers to anchor invariant semantic topologies while adapting shallow representations to target textures; and (3) at the output level, an anatomical KL-divergence regularization suppresses extreme macroscopic hallucinations. Extensive validation on a challenging cross-site cohort of highly anisotropic sagittal MRI demonstrates that Tri-KA effectively balances target-domain adaptation with the robust preservation of essential anatomical structures, achieving significant improvements over state-of-the-art TTA approaches. Code is available at `https://github.com/WayneBo98/Tri-KA`.

Keywords: Test-Time Adaptation · Source-Free Domain Adaptation · Rectal Cancer Segmentation · Magnetic Resonance Imaging.

1 Introduction

Accurate segmentation of rectal tumors and the surrounding rectum wall from Magnetic Resonance Imaging (MRI) is essential for clinical staging and treatment planning [2,20,10,3]. However, automated segmentation is challenged by extreme intra- and inter-patient anatomical heterogeneity, low tumor-to-background contrast, and the presence of confounding adjacent structures [4,19,18]. These challenges are exacerbated in cross-site clinical deployments, where variations in scanner vendors and imaging protocols introduce severe domain shifts [9,8,6], frequently leading to performance degradation in deep learning models [12,23,1].

Test-Time Adaptation (TTA) dynamically addresses inference-stage domain shifts. However, generic TTA introduces critical incompatibilities in 3D medical segmentation. First, standard continual frameworks (e.g., SHOT [13], SAR [16], CoTTA [24], EATA [21]) predominantly rely on entropy minimization or monolithic updates. In low-contrast medical scenarios, these noisy gradients frequently cause severe semantic drift and structural distortion. Second, conventional methods depend on Batch Normalization (BN) modulation [15,5,11], fundamentally conflicting with Instance Normalization (IN) pipelines in standard 3D architectures like nnU-Net. Finally, diffusion-based generative strategies [7,22,14] impose prohibitive computational overheads, hindering efficient online adaptation. This necessitates a strategy respecting both architectural and anatomical hierarchies.

To overcome these limitations under stringent privacy constraints, we propose Tri-KA, a Tri-level Knowledge Anchoring framework for cross-site rectal cancer segmentation. While regulations prohibit sharing raw source data, transmitting highly compressed, non-identifiable statistical priors incurs negligible communication cost and minimal privacy risk. Without this knowledge, target models undergo unconstrained adaptation, exacerbating structural distortions. Therefore, rather than relying on unstable monolithic updates, our core insight is to safely distill macroscopic source priors—specifically, frequency amplitudes, deep semantic prototypes, and population-level distribution baselines—offline at the source center. During online deployment, Tri-KA leverages these privacy-preserving priors as hierarchical anchors to guide adaptation.

Specifically, our main contributions are summarized as follows:

- To our knowledge, this is the first work to address source-free TTA for rectal cancer segmentation on highly anisotropic sagittal MRI. By explicitly tackling domain shifts in this challenging modality, we provide a practical, privacy-preserving solution for cross-site clinical application.
- We propose Tri-KA, a system-level paradigm rather than a simple aggregation of isolated modules. This integration specifically aligns with the physical properties of highly anisotropic MRI to prevent semantic drift by anchoring knowledge across three dimensions: (1) an input-level Fourier module to align global contrast while maintaining structural integrity; (2) a feature-level deep-to-shallow prototype mechanism to anchor invariant semantic topologies while adapting shallow representations; and (3) an output-level KL-divergence regularization to suppress extreme macroscopic hallucinations.

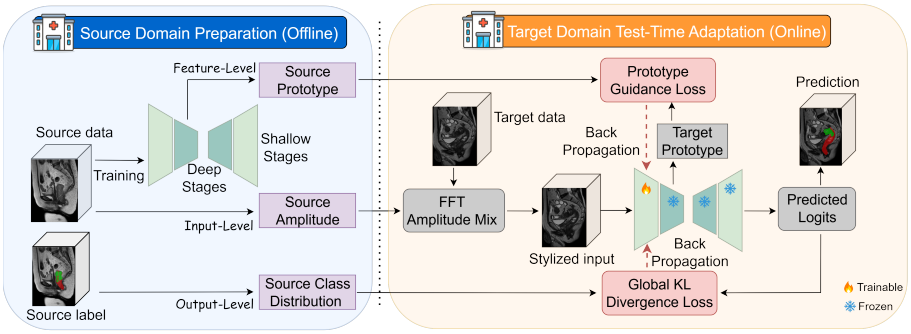

Fig. 1. Overview of the proposed Tri-KA framework. Privacy-preserving source priors are extracted offline, guiding the online adaptation via three different levels.

– Evaluations on a challenging cross-site cohort demonstrate that Tri-KA significantly improves segmentation accuracy over state-of-the-art TTA methods, effectively balancing target-specific adaptation with the robust preservation of essential anatomical structures.

2 Methodology

2.1 Overview of the Proposed Framework

Tri-KA safely adapts a pretrained model to an unseen target domain using a single target volume x_t, without accessing source training data. To maximize source information utilization while ensuring privacy, we shift away from monolithic fine-tuning, proposing a tri-level adaptation strategy that leverages precomputed heterogeneous source priors to prevent semantic drift. As illustrated in Fig. 1, our framework distills privacy-preserving source knowledge across three complementary dimensions: (1) Input-Level (Style Prior): A Fourier-based frequency module mitigates global contrast discrepancies before the image enters the network, aligning the target appearance with the source distribution. (2) Feature-Level (Semantic Prior): Deep prototypes extracted from the source domain act as stable semantic anchors to guide the texture adaptation of the shallow encoder. (3) Output-Level (Distribution Prior): A soft anatomical prior regularization penalizes extreme macroscopic deviations, effectively suppressing severe structural hallucinations. Through this multi-dimensional knowledge-anchoring mechanism, the model robustly adapts to target-specific appearances while preserving essential anatomical topology.

2.2 Input-Level: Energy-Calibrated Frequency Stylization

Cross-site MRI scans frequently exhibit significant global appearance shifts. To alleviate this domain gap without corrupting anatomical semantics, we introduce

a privacy-preserving, Fourier-based frequency stylization prior to the network forward pass. For a given target slice x_t, we extract its spatial amplitude $\mathcal{A}_t$ and phase $\mathcal{P}_t$ via 2D Fast Fourier Transform (FFT, $\mathcal{F}$). A pre-computed averaged source amplitude $\mathcal{A}_s$ serves as our privacy-preserving style prior.

Directly swapping low-frequency spectra in z-score normalized medical volumes often induces artifacts due to global energy mismatch. To safely blend styles, we employ a central low-pass Gaussian mask M alongside an *energy-calibrated* mixing strategy. Let E_t and E_s denote the localized energy (mean amplitude) within the masked regions of the target and source, respectively. The blended amplitude is formulated as:

$$\mathcal{A}'_t = \mathcal{A}_t + \beta \cdot M \cdot \left(\mathcal{A}_s \frac{E_t}{E_s} - \mathcal{A}_t \right) \tag{1}$$

where $\beta \in [0, 1]$ controls the stylization intensity, and the dynamic scaling factor E_t/E_s effectively prevents intensity overflow. The adapted slice is then reconstructed via Inverse $\mathcal{F}$ using $\mathcal{A}'_t$ and $\mathcal{P}_t$. Crucially, to accommodate architectures highly sensitive to baseline intensities (e.g., nnU-Net), we preserve the target's Direct Current (DC) component and restrict the adaptation exclusively to the foreground anatomy via a median-thresholding mask.

2.3 Feature-Level: Deep-to-Shallow Prototype Guidance

To prevent catastrophic forgetting during TTA, we propose a deep-to-shallow adaptation strategy tailored to the physical properties of anisotropic medical volumes. For small, highly anisotropic datasets like sagittal rectal MRI, CNN architectures (e.g., nnU-Net) remain the clinical gold standard. In such models, the encoder exhibits a distinct two-stage hierarchy. Driven by the progressive downsampling and receptive field expansion in 3D networks, the shallow layers perform anisotropic operations to align spatial resolutions, making them highly sensitive to cross-site variations (e.g., slice thickness) and domain-variant textures. Conversely, the deep isotropic stages encode stable, domain-invariant semantic topologies. Capitalizing on this, we freeze these deep layers and the entire decoder, exclusively fine-tuning the shallow anisotropic encoder ($\theta_{shallow}$) to safely map target textures into the invariant decoding space.

To guide $\theta_{shallow}$ without violating privacy, we utilize multi-level source prototypes extracted explicitly from the deep isotropic stages. Since these layers encode abstract, domain-invariant anatomy, they serve as optimal supervision signals. Prior to deployment, we apply K-Means clustering ($K = 5$) directly to the source deep features, generating a robust set of L_2-normalized semantic anchors $p_{c,l}$ for each class c at deep layer l.

During the online TTA phase, we dynamically filter the target prediction $\hat{Y}$ to establish reliable pseudo-labels. First, we identify high-confidence pixels: $V = \{i | \max_c \hat{Y}_{i,c} > \tau\}$. To prevent cascading errors, we apply a dual-gated spatial prior. Since tumors originate near the rectal wall, confident tumor predictions outside the dilated rectum are explicitly removed from V, conditioned

on the rectum anchor itself also meeting the confidence threshold τ. This robust anatomical filtering is primarily stabilized by the FFT module (Section 2.2), which effectively mitigates initial localization collapse.

The valid target features $f_{i,l}$ extracted from the deep isotropic layers are then optimized via cosine similarity against the matching source prototypes:

$$\mathcal{L}_{proto} = \frac{1}{|L_{deep}|} \sum_{l \in L_{deep}} (1 - \frac{1}{|\tilde{C}|} \sum_{c \in \tilde{C}} \frac{1}{|V_c| + \epsilon} \sum_{i \in V_c} \frac{f_{i,l} \cdot p_{c,l}}{||f_{i,l}||_2 ||p_{c,l}||_2}) \tag{2}$$

where $\tilde{C}$ denotes valid predicted classes ($|V_c| > 0$), and $\epsilon = 1 \times 10^{-8}$ ensures numerical stability. Crucially, while the supervision signal $\mathcal{L}_{proto}$ is computed in the deep semantic space, the deep layers themselves are frozen. Consequently, the gradients backpropagate through the invariant deep layers and flow exclusively into $\theta_{shallow}$. This deep-to-shallow mechanism forces the shallow anisotropic layers to adapt their texture representations such that the resulting deep features align closely with the domain-invariant semantic priors.

2.4 Output-Level: Anatomical Prior Regularization

While feature-level guidance stabilizes local semantics, extreme domain shifts frequently trigger catastrophic macroscopic hallucinations (e.g., massive false positives). To suppress these anatomically implausible outliers, we introduce a soft population-level prior based on the source domain's average class distribution. We define this prior as $P_{prior} \in \mathbb{R}^C$, empirically calculated from the source domain for each class.

During test-time, the target's macroscopic prediction distribution is computed as the spatial mean of the probability map for each class c:

$$P_{target}(c) = \frac{1}{DHW} \sum_{d,h,w} \hat{Y}_{d,h,w,c} \tag{3}$$

where $\hat{Y}_{d,h,w,c}$ denotes the predicted softmax probability at voxel (d, h, w) for class c. We penalize extreme deviations from this population baseline using the Kullback-Leibler (KL) divergence:

$$\mathcal{L}_{prior} = \sum_{c=1}^{C} P_{target}(c) \log \left(\frac{P_{target}(c)}{P_{prior}(c)} \right) \tag{4}$$

The overall objective for exclusively updating the shallow encoder $\theta_{shallow}$ is:

$$\mathcal{L}_{total} = \mathcal{L}_{proto} + \lambda \mathcal{L}_{prior} \tag{5}$$

Crucially, $\mathcal{L}_{prior}$ acts as a soft constraint to penalize gross volume anomalies (e.g., organ-scale false positives), thereby allowing $\mathcal{L}_{proto}$ to preserve natural patient-level tumor heterogeneity without catastrophic failure.

Table 1. Comparison with SOTA methods. Metrics are Mean ± SD. Best in **bold**, second underlined. Significance vs. ours: * $p < 0.05$, *** $p < 0.001$.

Method	Rectum (Anatomical Anchor)		Tumor (Clinical Target)	
	Dice (%, ↑)	ASD (mm, ↓)	Dice (%, ↑)	ASD (mm, ↓)
Baseline	58.64 ± 4.95***	2.98 ± 0.27***	45.65 ± 1.45***	8.59 ± 0.57***
TTAug (8×)	59.20 ± 4.56***	2.94 ± 0.38***	46.05 ± 2.09***	8.25 ± 0.88***
TENT [23]	58.45 ± 6.09***	**2.68 ± 0.21****	46.23 ± 1.17***	**7.81 ± 0.46**
SHOT [13]	61.40 ± 1.71	3.33 ± 0.41*	45.21 ± 1.75***	8.07 ± 0.43
CoTTA [24]	58.97 ± 4.72***	3.16 ± 0.20***	47.26 ± 0.96***	8.02 ± 0.91*
SAR [16]	59.55 ± 4.01***	3.01 ± 0.24***	46.20 ± 1.71***	7.92 ± 0.82***
EATA [21]	58.63 ± 5.26***	2.88 ± 0.27***	45.92 ± 1.77***	8.19 ± 0.77***
GALA [17]	58.63 ± 5.13***	2.80 ± 0.22***	46.06 ± 1.43***	8.05 ± 0.63***
Ours	**61.53 ± 1.92**	3.70 ± 0.39	**48.66 ± 1.91**	7.88 ± 0.72

3 Experiments and Results

3.1 Experimental Setup

Datasets. We evaluate Tri-KA on two in-house sagittal T2 MRI datasets for rectal cancer. Source Dataset A (80 volumes from Guangdong Provincial People's Hospital) is randomly split into 5 folds to train independent base models. Target Dataset B (77 anisotropic volumes from Second Affiliated Hospital of Navy Medical University) exhibits severe domain shifts. Target slices are resized to 320×320 to approximate the source domain's voxel spacing ratio.

Implementation & Metrics. Implemented in PyTorch (NVIDIA 4090), Tri-KA adapts five base models independently, reporting average results. $\theta_{shallow}$ is optimized using Adam (LR=3×10^{-5}) for 20 steps per subject. Hyperparameters are fixed as: $\tau = 0.95, \lambda = 2.0, \beta = 0.5$. Dice and ASD evaluate accuracy; Wilcoxon signed-rank tests assess significance ($^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$). Whole-volume inference with compatible padding preserves the global anatomical context essential for stable feature-level adaptation.

3.2 Quantitative Results

We benchmark against the unadapted baseline (with/without 8× test-time augmentation, i.e., TTAug) and state-of-the-art TTA methods using largest connected component (LCC) post-processing. As shown in Table 1, our approach achieves the highest Dice scores (rectum: 61.53%, tumor: 48.66%), significantly outperforming most competitors ($p < 0.001$). Furthermore, reducing tumor ASD to 7.88 mm demonstrates effective malignant boundary delineation.

Existing methods often struggle to balance pathological and healthy structures. For instance, TENT yields the lowest tumor ASD (7.81 mm) due to a conservative predictive tendency. Unconstrained entropy-driven adaptation forces

Table 2. Ablation of Tri-KA components. Best in **bold**, second underlined. Significance vs. full method: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Components			Rectum (Anatomical Anchor)		Tumor (Clinical Target)	
FFT	Proto	KL	Dice (%, ↑)	ASD (mm, ↓)	Dice (%, ↑)	ASD (mm, ↓)
×	×	×	58.64±4.95***	2.98±0.27***	45.65±1.45***	8.59±0.57***
✓	×	×	59.13±4.84***	**2.92±0.30****	47.49±1.42**	8.09±0.84
×	✓	×	60.06±3.02***	3.52±0.32***	47.07±2.11***	8.32±1.16***
✓	✓	×	60.64±3.66*	3.43±0.39***	48.03±2.30	7.95±1.38***
✓	✓	✓	**61.53±1.92**	3.70±0.39	**48.66±1.91**	**7.88±0.72**

Table 3. Ablation of feature-level guidance strategy. P/T denotes Proto./Tuned layers (D: Deep, S: Shallow). Best in **bold**, second underlined.

Strat.	P/T	Rectum		Tumor	
		Dice (%, ↑)	ASD (mm, ↓)	Dice (%, ↑)	ASD (mm, ↓)
Base	-	58.64±4.95***	**2.98±0.27*****	45.65±1.45**	8.59±0.57
(a)	D/D	58.63±4.38***	3.96±0.71	46.25±2.89	9.35±1.52
(b)	S/S	58.40±5.11***	3.06±0.21***	46.36±1.91*	**8.24±0.76**
(c) **Ours**	D/S	**60.06±3.02**	3.52±0.32	**47.07±2.11**	8.32±1.16

overconfidence in initial biased predictions, capturing only the central core while under-segmenting ambiguous boundaries (TENT tumor Dice: 46.23%). Similarly, SHOT achieves competitive rectum Dice (61.40%) but noticeably degrades tumor Dice below baseline (45.21%).

In contrast, our multi-dimensional guidance successfully coordinates the adaptation of both classes to recover more comprehensive anatomical structures. While this adaptation introduces an acceptable relaxation in the healthy rectum ASD (3.70 mm vs. 2.98 mm), this 0.72 mm sub-voxel variance remains well within the highly anisotropic MRI resolution limits. More importantly, this pragmatic clinical trade-off effectively improves highly challenging malignant tumor delineation (ASD reduced to 7.88 mm). Specifically, by prioritizing optimization on pathological regions over the stable rectal wall, Tri-KA ensures that clinical focus remains on accurate tumor localization, which is of paramount importance for cancer staging.

We evaluate the contributions of individual modules in Table 2. The input-level FFT stylization provides a stable initialization by mitigating frequency-domain shifts, which improves the tumor Dice from 45.65% to 47.49%. Independently, feature-level prototype guidance aids in recovering the global topology, raising the rectum Dice to 60.06%. Combining both mechanisms further improves segmentation performance. Finally, incorporating the output-level KL regularization effectively filters macroscopic false positives, achieving the highest Dice scores for both the rectum (61.53%) and the tumor (48.66%). Consistent with the

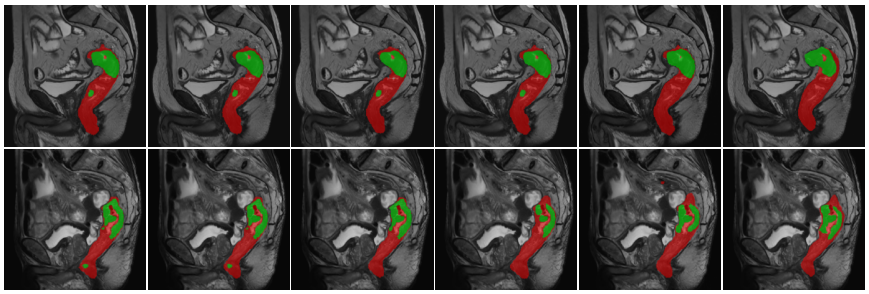

Fig. 2. Qualitative comparison of cross-site adaptation results across different methods. The red and green regions indicate segmented rectum and rectal tumor, respectively.

observations in the SOTA comparison, the comprehensive recovery of anatomical structures introduces an acceptable relaxation in the healthy rectum ASD (3.70 mm). Reiterating our clinical objective, this variance is a strategic trade-off that successfully secures robust malignant tumor localization, as reflected by the optimal tumor ASD of 7.88 mm.

To isolate the internal mechanics of the feature-level guidance, we ablate the layer-selection strategy on the raw baseline (without FFT or KL modules) in Table 3. Strategy (a) fine-tunes the deep layers, which inadvertently disrupts the network's domain-invariant topological knowledge. This modification leads to a decrease in rectum Dice and introduces instability in tumor localization (tumor ASD deteriorating to 9.35±1.52 mm). Alternatively, Strategy (b) utilizes shallow features for guidance. Because shallow representations are highly susceptible to domain-specific texture shifts, they provide less reliable optimization trajectories, resulting in an inferior tumor Dice (46.36%). By freezing the deep semantic anchors and the decoder, and exclusively updating the shallow encoder, our deep-to-shallow strategy (c) effectively resolves this structural-textural conflict. This configuration significantly improves the tumor Dice to 47.07% ($p < 0.01$) while successfully maintaining the global anatomical topology (rectum Dice: 60.06%). Furthermore, based on our existing parameter tuning logs, the framework demonstrates minimal performance fluctuations within the ranges of $\beta \in [0.4, 0.6]$ and $\lambda \in [1.5, 2.5]$. This stability confirms that our method relies on robust anatomical priors rather than heavy dataset-specific tuning.

3.3 Qualitative Results

Fig. 2 visualizes cross-site results. In the first row, the unadapted baseline incorrectly predicts a macroscopic false-positive (FP) tumor region due to severe domain shifts. Conventional methods (TENT, SHOT, GALA) fail to eliminate this artifact, demonstrating their limited capacity to correct structural errors when relying solely on unconstrained entropy minimization. Conversely, guided by our multi-dimensional priors, Tri-KA effectively suppresses this FP, yielding a prediction highly consistent with Ground Truth.

The second row presents a complex case where competitors retain FPs and severely distort true tumor morphology. Tri-KA outperforms other methods in recovering fine-grained boundaries while substantially reducing FPs. This visually confirms our motivation: freezing deep semantic anchors and the entire decoder prevents implausible hallucinations, while exclusively adapting the shallow encoder safely refines target textures.

4 Conclusion

In this paper, we propose Tri-KA, a privacy-preserving Tri-level Knowledge Anchoring framework for cross-site rectal MRI test-time adaptation. By integrating macroscopic source priors across three dimensions—notably, freezing deep semantic anchors while adapting shallow layers to target textures—Tri-KA effectively prevents semantic drift. Cross-site evaluations demonstrate its superiority, providing a robust, clinically practical solution. While currently exploiting the anisotropic nature of sagittal MRI, extending this mechanism to isotropic modalities and conducting broader multi-center validations remain future work.

Acknowledgments. Supported by Natural Science Foundation of China (62271465), National Key R&D Program of China (2025YFC3408300), Natural Science Foundation of Jiangsu Province (BK20255001), and Suzhou Basic Research Program (SYG202338).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bateson, M., Kervadec, H., Dolz, J., Lombaert, H., Ben Ayed, I.: Source-relaxed domain adaptation for image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 490–499. Springer (2020)
2. Beets-Tan, R.G., Lambregts, D.M., Maas, M., Bipat, S., Barbaro, B., Curvo-Semedo, L., Fenlon, H.M., Gollub, M.J., Gourtsoyianni, S., Halligan, S., et al.: Magnetic resonance imaging for clinical management of rectal cancer: updated recommendations from the 2016 european society of gastrointestinal and abdominal radiology (esgar) consensus meeting. European radiology **28**(4), 1465–1475 (2018)
3. Benson, A.B., Venook, A.P., Al-Hawary, M.M., Azad, N., Chen, Y.J., Ciombor, K.K., Cohen, S., Cooper, H.S., Deming, D., Garrido-Laguna, I., et al.: Rectal cancer, version 2.2022, nccn clinical practice guidelines in oncology. Journal of the National Comprehensive Cancer Network **20**(10), 1139–1167 (2022)
4. Bo, W., Xue, T., Pan, L., Huang, D., Xiao, Y., Fan, L., Liu, Z., Liu, S., Zhou, S.K.: Rcsegnext: Efficient multi-scale convnext for rectal cancer segmentation from sagittal mri scans. In: Tasdizen, T., Elhabian, S., Summers, R., Chen, C., Koch, L., Zhuang, Y. (eds.) Proceedings of The 8th International Conference on Medical Imaging with Deep Learning. Proceedings of Machine Learning Research, vol. 301, pp. 167–182. PMLR (09–11 Jul 2026), https://proceedings.mlr.press/v301/bo26a.html

5. Dong, H., Konz, N., Gu, H., Mazurowski, M.A.: Medical image segmentation with intent: Integrated entropy weighting for single image test-time adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5046–5055 (2024)
6. Dou, Q., Coelho de Castro, D., Kamnitsas, K., Glocker, B.: Domain generalization via model-agnostic learning of semantic features. Advances in neural information processing systems **32** (2019)
7. Gao, J., Zhang, J., Liu, X., Darrell, T., Shelhamer, E., Wang, D.: Back to the source: Diffusion-driven test-time adaptation. arXiv preprint arXiv:2207.03442 (2022)
8. Glocker, B., Robinson, R., Castro, D.C., Dou, Q., Konukoglu, E.: Machine learning with multi-site imaging data: An empirical study on the impact of scanner effects. arXiv preprint arXiv:1910.04597 (2019)
9. Guan, H., Liu, M.: Domain adaptation for medical image analysis: a survey. IEEE Transactions on Biomedical Engineering **69**(3), 1173–1185 (2021)
10. Horvat, N., Carlos Tavares Rocha, C., Clemente Oliveira, B., Petkovska, I., Gollub, M.J.: Mri of rectal cancer: tumor staging, imaging techniques, and management. Radiographics **39**(2), 367–387 (2019)
11. Joshi, S., Osuala, R., Garrucho, L., Kushibar, K., Kessler, D., Diaz, O., Lekadir, K.: Single image test-time adaptation via multi-view co-training. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 617–627. Springer (2025)
12. Karani, N., Erdil, E., Chaitanya, K., Konukoglu, E.: Test-time adaptable neural networks for robust medical image segmentation. Medical Image Analysis **68**, 101907 (2021)
13. Liang, J., Hu, D., Feng, J.: Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In: International conference on machine learning. pp. 6028–6039. PMLR (2020)
14. Liu, M., Li, X., Guo, M., Duan, L., Fang, H., Xu, Y.: Leveraging diffusion models for continual test-time adaptation in fundus image classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 337–347. Springer (2025)
15. Nado, Z., Padhy, S., Sculley, D., D'Amour, A., Lakshminarayanan, B., Snoek, J.: Evaluating prediction-time batch normalization for robustness under covariate shift. arXiv preprint arXiv:2006.10963 (2020)
16. Niu, S., Wu, J., Zhang, Y., Wen, Z., Chen, Y., Zhao, P., Tan, M.: Towards stable test-time adaptation in dynamic wild world. arXiv preprint arXiv:2302.12400 (2023)
17. Sahoo, S., ElAraby, M., Ngnawe, J., Pequignot, Y.B., Precioso, F., Gagné, C.: A layer selection approach to test time adaptation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 20237–20245 (2025)
18. Sasuga, S., Kudo, A., Kitamura, Y., Iizuka, S., Simo-Serra, E., Hamabe, A., Ishii, M., Takemasa, I.: Image synthesis-based late stage cancer augmentation and semi-supervised segmentation for mri rectal cancer staging. In: MICCAI Workshop on Data Augmentation, Labelling, and Imperfections. pp. 1–10. Springer (2022)
19. Shen, J., Lu, S., Qu, R., Zhao, H., Zhang, L., Chang, A., Zhang, Y., Fu, W., Zhang, Z.: A boundary-guided transformer for measuring distance from rectal tumor to anal verge on magnetic resonance images. Patterns **4**(4) (2023)
20. Sung, H., Ferlay, J., Siegel, R.L., Laversanne, M., Soerjomataram, I., Jemal, A., Bray, F.: Global cancer statistics 2020: Globocan estimates of incidence and mor-

tality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians **71**(3), 209–249 (2021)
21. Tan, M., Chen, G., Wu, J., Zhang, Y., Chen, Y., Zhao, P., Niu, S.: Uncertainty-calibrated test-time model adaptation without forgetting. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025)
22. Tsai, Y.Y., Chen, F.C., Chen, A.Y., Yang, J., Su, C.C., Sun, M., Kuo, C.H.: Gda: Generalized diffusion for robust test-time adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23242–23251 (2024)
23. Wang, D., Shelhamer, E., Liu, S., Olshausen, B., Darrell, T.: Tent: Fully test-time adaptation by entropy minimization. arXiv preprint arXiv:2006.10726 (2020)
24. Wang, Q., Fink, O., Van Gool, L., Dai, D.: Continual test-time domain adaptation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7201–7211 (2022)