

A Large Scale Benchmark for Test Time Adaptation Methods in Medical Image Segmentation

Wenjing Yu^{a,*}, Shuo Jiang^{a,*}, Yifei Chen^{b,*}, Shuo Chang^a, Yuanhan Wang^a,
Beining Wu^a, Jie Dong^a, Mingxuan Liu^b, Shenghao Zhu^a, Feiwei Qin^{a,**},
Changmiao Wang^c, Qiyuan Tian^{b,**}

^aHangzhou Dianzi University, Hangzhou, 310018, China

^bTsinghua University, Beijing, 100084, China

^cShenzhen Research Institute of Big Data, Shenzhen, 518172, China

Abstract

Test time Adaptation is a promising approach for mitigating domain shift in medical image segmentation; however, current evaluations remain limited in terms of modality coverage, task diversity, and methodological consistency. We present MedSeg-TTA, a comprehensive benchmark that examines twenty representative adaptation methods across seven imaging modalities, including MRI, CT, ultrasound, pathology, dermoscopy, OCT, and chest X-ray, under fully unified data preprocessing, backbone configuration, and test time protocols. The benchmark encompasses four significant adaptation paradigms: Input-level Transformation, Feature-level Alignment, Output-level Regularization, and Prior Estimation, enabling the first systematic cross-modality comparison of their reliability and applicability. The results show that no single paradigm performs best in all conditions. Input-level methods are more stable under mild appearance shifts. Feature-level and Output-level methods offer greater advantages in boundary-related metrics, whereas prior-based methods exhibit strong modality dependence. Several methods degrade significantly under large inter-center and inter-device shifts, which highlights the importance of principled method selection for clinical deployment. MedSeg-TTA provides standardized datasets, validated implementations, and a public leaderboard, establishing a rigorous foundation for future research on robust, clinically reliable test-time adaptation. All source codes and open-source datasets are available at <https://github.com/wenjing-gg/MedSeg-TTA>.

*These authors contributed equally to this work.

**Corresponding authors: qinfeiwei@hdu.edu.cn, qiyuantian@tsinghua.edu.cn.

Keywords:

Benchmarking, Domain Shift, Input-level Transformation, Feature-level Alignment, Output-level Regularization, Prior Estimation

1. Introduction

Medical image segmentation plays a pivotal role in clinical diagnosis, therapeutic evaluation, and preoperative planning [1, 68]. Its precision directly determines the reliability of lesion quantification, organ morphometric analysis, and the development of individualized treatment strategies [2, 69, 67]. However, in practical applications, distribution shifts across institutions, scanners, and populations severely hinder the generalization capacity of deep learning models, leading to substantial performance degradation when tested on the target domain [10, 11, 12]. To address this challenge, Test-Time Adaptation (TTA) has recently attracted considerable attention [6, 14]. Unlike conventional domain adaptation methods, TTA neither requires access to source domain data nor an additional training stage in the target domain [13, 70]. Instead, it dynamically adjusts model parameters during deployment and instantaneously corrects predictions in a batch-wise or streaming manner, thereby offering greater clinical applicability [33]. Liang *et al.* [28] systematically categorized existing TTA approaches into four groups based on their adaptation mechanisms: Input-level Transformation, Feature-level Alignment, Output-level Regularization, and Prior Estimation. This taxonomy provides a clear framework for subsequent investigations.

Input-level Transformation methods adjust the appearance and statistical characteristics of test samples in the pixel or frequency domain, making their distributions more similar to those of the source domain, thereby enabling inference under familiar input conditions. A representative example is the reliable source approximation method proposed by Zeng *et al.* [29], which uses diffusion models to map target-domain images into the source-domain style before prediction. Feature-level Alignment focuses on reducing distributional discrepancies between the source and target domains in the intermediate representation space. A typical approach is adversarial training, in which Hu *et al.* [30] encouraged discriminative features to become domain-invariant through the interplay of a domain discriminator and a feature extractor. In addition, lightweight adaptation and statistical recalibration can be integrated to achieve latent space alignment without modifying the backbone architecture [5]. Output-level Regularization methods treat the model’s predictions as supervisory signals and impose self-supervised constraints

directly on the output distribution to enhance cross-domain robustness. Common strategies include entropy minimization, Maximum Squares, Batch Nuclear-norm Maximization, and mutual information maximization. For instance, Weihsbach *et al.* [31] improved the stability of out-of-distribution predictions by introducing a dual-branch consistency mechanism. Prior Estimation methods explicitly incorporate relatively stable categorical or anatomical priors during inference to constrain the reasoning process via energy functions, regularization terms, or conditional prompts. For example, Hu *et al.* [32] utilized prompt learning to encode both task semantics and domain-level priors, thereby guiding segmentation models toward convergent alignment when source data are unavailable.

Despite substantial advances in related research, the application of TTA to medical image segmentation still encounters several bottlenecks. The main limitations are as follows: most existing studies demonstrate effectiveness only on single-modality, single-task, or single-center datasets, with little systematic evaluation across diverse imaging modalities and tasks [4]. Evaluation protocols are also highly heterogeneous, and there is no agreement on critical design choices, such as whether cross-sample accumulation of statistics or external updates to batch normalization parameters are allowed. Some experimental setups further pose risks of source data or label leakage. In addition, evaluation metrics are often oversimplified, with many works reporting only the Dice coefficient while neglecting structurally sensitive measures such as Sen and PPV, thereby weakening comparability and clinical interpretability. Collectively, these issues hinder fair benchmarking and reproducibility, making it difficult to faithfully characterize the strengths, limitations, and failure modes of different TTA paradigms under heterogeneous-domain and multi-task conditions [3].

Inspired by the comprehensive review of Liang *et al.* [28], this study introduces a unified evaluation framework that, for the first time in the field of medical imaging, provides a systematic benchmark across various modalities, centers, and tasks. We assemble datasets from seven primary imaging modalities: MRI, CT, US, PATH, Dermoscopy (DER), OCT, and CXR. Under a unified backbone, data partitioning, and testing protocol, we systematically compare the performance of the four paradigms: Input-level Transformation, Feature-level Alignment, Output-level Regularization, and Prior Estimation.

The main contributions are summarized as follows:

- **Multi-modal and multi-center open-source dataset:** We construct a dataset covering tumor, organ, and lesion segmentation across seven imaging modalities: MRI, CT, US, PATH, DER, OCT, and CXR. The dataset uses stan-

dardized preprocessing and partitioning, reflecting distribution shifts across institutions, scanners, and populations, thereby providing a basis for unified TTA in medical image segmentation benchmarks.

- **Strong and reproducible TTA baselines:** We design a unified TTA protocol that fixes the backbone and strictly forbids source-domain access as well as any implicit information leakage. In this setting, we systematically reproduce and validate 20 representative state-of-the-art TTA methods for medical image segmentation across 4 paradigms.
- **Unified benchmark and public leaderboard:** Building on these implementations, we establish a public benchmark and leaderboard that support comparisons across imaging modalities, organs, and tasks using both region-overlap and structurally sensitive metrics such as Dice and HD95.
- **Paradigm taxonomy and applicability lineage:** We categorize TTA methods into four paradigms according to their locus of operation and, based on evaluations across modalities, organs, and tasks, construct *lineage maps* that highlight effective and ineffective regimes. This delineates applicability boundaries and provides practical guidance for future method selection.

2. Related Work

From a functional perspective, TTA methods for medical image segmentation can be broadly categorized into four paradigms [28]. Input-level Transformation achieves cross-domain consistency by adjusting appearance and statistical styles while preserving structural semantics as much as possible. Feature-level Alignment reconstructs robust latent representations and establishes distributional consistency, thereby enhancing feature resilience against domain shifts. Output-level Regularization imposes constraints on the prediction distribution and decision boundaries, enabling models to maintain stability and coherence under uncertain or out-of-distribution conditions. Prior Estimation introduces relatively stable forms of knowledge, such as anatomy, shape, and class frequency, to narrow the feasible solution space and improve interpretability and consistency of predictions. Around these four themes, TTA has gradually developed into a complete technical taxonomy. The following sections review representative methods and analyze their advantages, applicability boundaries, and common challenges in multi-modal and multi-center scenarios.

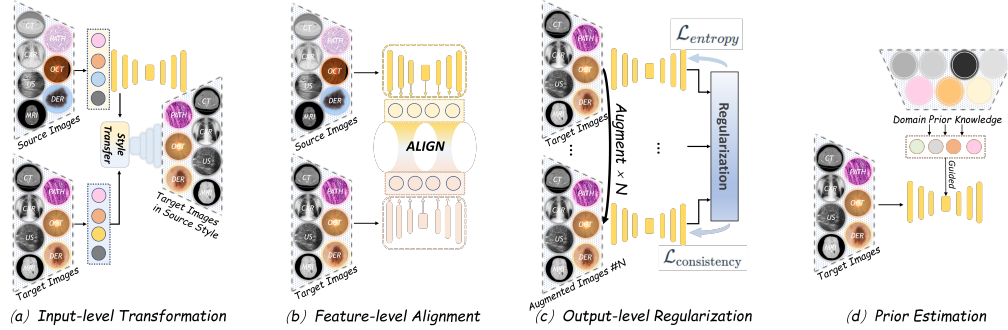


Figure 1: **Schematic diagram of the four paradigms in TTA.** (a) Input-level Transformation, adapting the model to target domain images by altering image appearances; (b) Feature-level Alignment, aligning features from the source and target domains; (c) Output-level Regularization, regularizing the model outputs for target domain adaptation; (d) Prior Estimation, forcing the model to update towards the target domain using prior information of the target domain.

2.1. Input-level Transformation TTA

Input-level methods generally follow the principle of aligning appearance first, then inferring with the pretrained model. That is, they attenuate cross-center imaging discrepancies by aligning styles or reconstructing images while retaining anatomical semantics, thereby achieving recategorized results without modifying model parameters. Depending on the implementation, they can be divided into style-replacement and generative model approaches.

2.1.1. Style Replacement

Style replacement aims to mitigate the discrepancies between the source and target domains by modifying the visual appearance of cross-domain images [9]. A representative work is that of Yang *et al.* [49], who proposed replacing low-frequency amplitudes in the frequency domain to achieve cross-domain style transfer while effectively preserving boundaries and delicate textures. This method assumes that low-frequency differences are the primary source of appearance shifts and demonstrates strong adaptability in 3D medical imaging. Nevertheless, frequency-domain style replacement still faces challenges, including ensuring anatomical consistency and avoiding structural distortions across varying slice thicknesses, window widths, and scanning protocols.

2.1.2. Generative Models

Generative model approaches reconstruct target-domain images at the explicit pixel level into a source-style representation, which a pre-trained segmentation

model then processes [8]. Their powerful modeling capacity enables stronger and more robust appearance remapping across centers and modalities. Existing works can be broadly categorized into two main types: generative adversarial networks and diffusion models [7, 29]. The former generates images consistent with the target domain through the antagonistic interplay between a generator and a discriminator and has been widely applied in medical imaging. The latter produces high-quality images via a progressive denoising process, thereby substantially enhancing cross-domain segmentation performance. However, generative models still face challenges with computational overhead and potential artifacts, particularly in 3D imaging tasks.

2.2. Feature-level Alignment TTA

Feature-level Alignment aims to alleviate distributional shifts in the latent space, rendering intermediate representations more robust and discriminative in the target domain. Omid *et al.* [60] analyzed skull stripping from adults to neonates, demonstrating that relying solely on Input-level Transformation is insufficient to handle substantial morphological differences, as style normalization may still result in inter-class entanglement and boundary ambiguity. To address this issue, Zhang *et al.* proposed TestFit [50], which applies one-time, plug-and-play calibration on intermediate representations within the network, achieving lightweight adaptation without access to source data. However, this method remains vulnerable to noise and gradient drift under sequential domain shifts and small-batch inference. Chen *et al.* [61] introduced a gradient alignment strategy, which constrains parameter update directions to suppress noise amplification and catastrophic forgetting, although stability remains limited. Subsequently, Hu *et al.* [52] proposed mutual information maximization alignment, which enhances feature clustering and inter-cluster separation under unlabeled conditions, thereby improving discriminability when significant appearance discrepancies exist. Nevertheless, this approach is sensitive to the quality of pretrained representations, and optimization may fail to converge when initial features are unsuitable. To further enhance robustness, Wen *et al.* [62] combined denoising pretraining with test-time alignment, achieving more stable improvements in multi-center segmentation tasks. Feature-level Alignment directly regularizes representations without modifying inputs and incurs low computational overhead. However, it remains sensitive to small-batch statistics and anatomical variability, making it challenging to consistently enforce clear boundaries when used alone.

2.3. Output-level Regularization TTA

Output-level Regularization methods treat the model’s prediction distribution in the target domain as a supervisory signal, establishing lightweight optimization objectives during testing that focus on entropy minimization, self-training with pseudo labels, and consistency regularization. This line of work offers low deployment costs and zero reliance on source data, making it a mainstream strategy for TTA in medical image segmentation. Currently, methods can be primarily divided into TENT-type methods, which focus on entropy minimization, and self-supervised methods, which center on self-supervision or pseudo-labeling.

2.3.1. TENT-Type Methods

Wang *et al.* [33] proposed TENT, which minimizes prediction entropy and updates only batch normalization statistics and affine parameters. This approach achieves cross-domain improvements without source data and establishes the baseline for Output-level Regularization. However, in small-batch and strong domain-shift scenarios, entropy minimization alone tends to cause overconfidence and cumulative drift. To address this issue, Niu *et al.* [34] introduced EATA, which alleviates noise amplification through uncertainty filtering and conservative regularization, although it is mainly applicable to one-shot domain shifts and struggles with continuous distributional changes. Wang *et al.* [35] subsequently proposed CoTTA, which incorporates teacher-student exponential moving averages and random replay mechanisms to balance short-term adaptation with long-term memory. However, this approach introduces additional storage and computational overhead in 3D inference. Later, Weihsbach *et al.* [31] presented DG-TTA, which combines domain-generalization pretraining with test-time entropy regularization to improve cold-start stability, although its performance remains limited in adapting to small lesions and scale variations. Additionally, Li *et al.* proposed SaTTCA [57], which incorporates scale-aware geometric priors during testing, thereby improving boundary delineation and recall of small targets; however, it relies heavily on interactions and prior knowledge.

2.3.2. Self-Supervised Methods

Another line of research introduces self-supervision or pseudo-labeling mechanisms to mitigate noisy supervision by leveraging consistency regularization and teacher-student frameworks. Sun *et al.* [41] proposed TTT, which augments representations during inference by paralleling self-supervised tasks, although its semantic coupling with the main task is weak. Liu *et al.* [56] proposed TTT++, which enhances task-relevant self-supervised signals via robust optimization to

improve adaptation stability, but it still lacks direct constraints in the output space. Weihsbach *et al.* [31] further proposed dual-branch consistency regularization, which establishes lightweight calibration directly on unlabeled predictions; however, it is susceptible to noise amplification in class-imbalanced settings. Wu *et al.* [58] introduced UPL-SFDA, which reduces noise accumulation by filtering pseudo-labels with uncertainty-aware filtering, thereby significantly improving the quality of pseudo-labels under source-free conditions. However, in continuous and non-stationary scenarios, single regularization remains prone to suboptimality and requires integration with Feature-level Alignment or anatomical priors.

2.4. Prior Estimation TTA

Prior-based methods emphasize explicitly incorporating anatomical and shape priors during testing to constrain the rationality and interpretability of predictions. Bateson *et al.* [51] proposed Shape-Moment TTA, which employs geometric moments, such as area and perimeter, as global constraints to ensure predictions adhere to plausible shapes. However, static global priors struggle to capture individual variability. Zhang *et al.* introduced PASS [74], which encodes style and semantic shapes into test-time prompts to achieve organ-level controllable alignment, thereby alleviating the under-constrained nature of global statistics. However, prompt acquisition and cross-center generalization remain limited. Chen *et al.* [63] further proposed case-specific customized prompts, dynamically injecting location and shape information during inference to enhance individual robustness, though balancing prompt-representation coupling remains challenging. Hu *et al.* [32] employed prompt learning to explicitly encode task semantics and domain-level priors, enabling alignment under source-free conditions while ensuring privacy and compliance. Omolegan *et al.* [59] explored combining structural priors with TTA, providing stable anchors in fetal brain 3D ultrasound, though real-time budgets and 3D accuracy remain challenging. As a lightweight approach, DenseCRF can rapidly refine thin-wall structures during testing by leveraging regional similarity and boundary smoothing [64]. This method improves local consistency but lacks global control and is therefore often integrated with shape or prompt priors to achieve overall stability.

3. Benchmarks and Implementation

This section outlines the benchmarking protocol, including the datasets and implementation, to facilitate a comprehensive and equitable comparison of TTA approaches for medical image segmentation across multi-modality, multi-task,

and multi-center conditions. Section 3.1 systematically examines seven representative dataset pairs from the same modality and task, namely MRI, CT, US, PATH, DER, OCT, and CXR, which exhibit domain shifts, and then introduces a unified evaluation framework. Section 3.2 presents standardized implementation, training, and testing protocols that specify the default configurations and constraints for the four TTA paradigms.

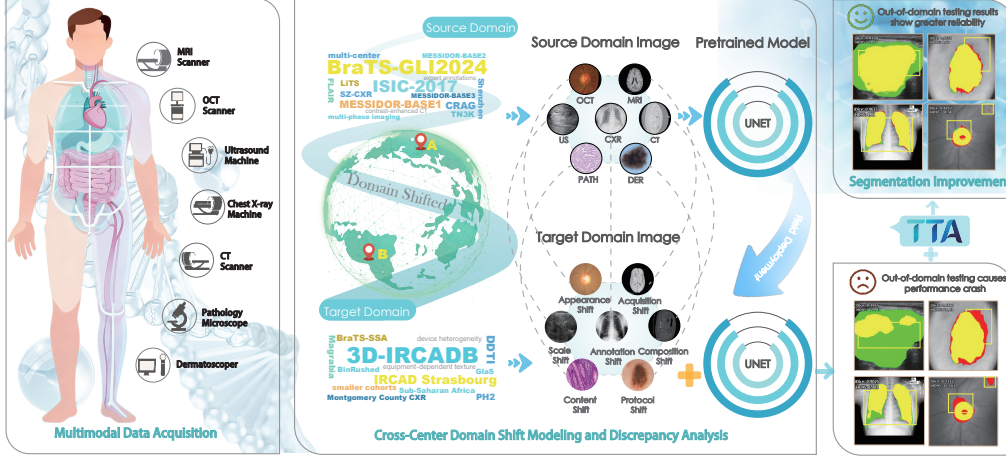


Figure 2: **The overall workflow of TTA methods for medical image segmentation.** It begins with the collection of multi-center data across seven modalities, which include MRI, CT, US, PATH, DER, OCT, and CXR. Next, a pretrained network is trained on source domain data, and benchmark testing is conducted on the target domain. Finally, out-of-domain segmentation performance improves with TTA methods.

3.1. Datasets and Evaluation Metrics

3.1.1. MRI-based Brain Tumor Segmentation Datasets

Source domain: BraTS-GLI2024 [65], Target domain: BraTS-SSA [16], as shown in Fig. 4. The BraTS series provides large-scale, multi-center, multi-sequence MRI data, including T1, T1c, T2, and FLAIR, with standardized annotations, serving as the de facto benchmark for adult glioma segmentation. We use 1,350 cases from BraTS-GLI2024 as the source-domain training set, with labels for enhancing tumor (ET), tumor core (CT), and whole tumor (WT). BraTS-SSA contains 60 multi-sequence MRI cases from the SSA region. Substantial differences in scanner conditions, acquisition protocols, and population composition naturally induce cross-region and cross-device domain shifts, making it an appropriate target domain for TTA evaluation under resource-limited scenarios.

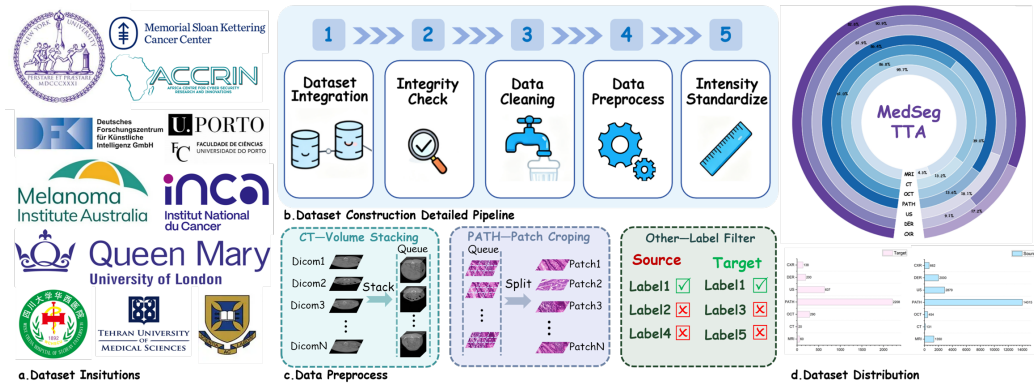


Figure 3: **Overview of the multi-institutional benchmark and dataset construction pipeline.** (a) participating institutions contributing data for the benchmark; (b) unified construction pipeline comprising dataset integration, integrity checking, data cleaning, data processing, and intensity standardization; and (c) data preprocessing procedures, including CT volume stacking from DICOM slices to 3D volumes, PATH patch cropping from whole-slide images, and source–target label filtering to harmonize label sets. (d) distribution of datasets across imaging modalities and their assignment to source or target domains;

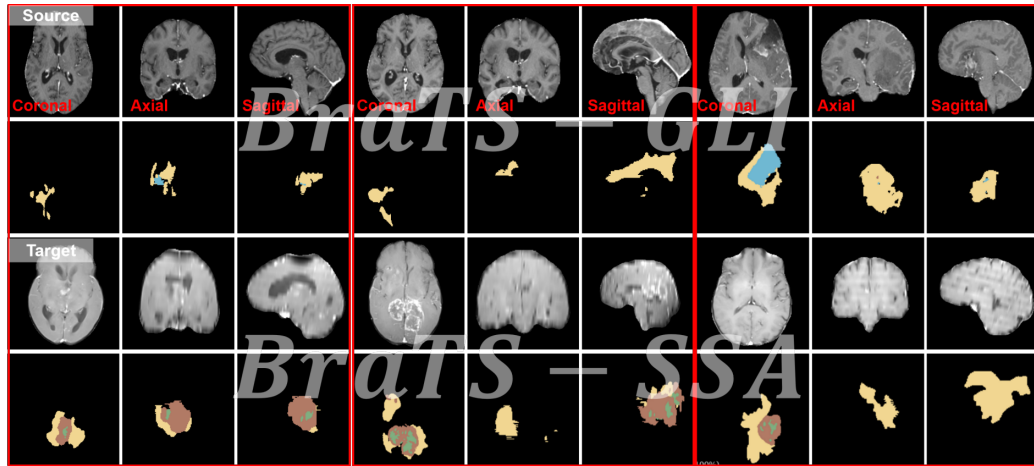


Figure 4: **Multi-sequence MRI for brain tumor segmentation.** Cases from the source dataset BraTS-GLI2024 and the target dataset BraTS-SSA. For the source dataset area, the first row displays T1, T1-weighted contrast-enhanced, T2 and FLAIR sequence images and enhancing tumor, tumor core and whole tumor annotations from anatomical views including axial, coronal and sagittal planes; the second row presents the corresponding annotation masks. The bottom section shows the equivalent multi-sequence MRI images and annotations for the target dataset.

3.1.2. CT-based Liver Segmentation Datasets

Source domain: LiTS [17], Target domain: 3D-IRCADB [18], as shown in Fig. 5. The LiTS dataset comprises multi-institutional, multi-protocol CT volumes of the liver and tumors, from which we select 131 publicly available training cases as the source domain. 3D-IRCADB, collected at IRCAD in France, comprises 20 enhanced CT cases with distinct acquisition devices, contrast phases, and a smaller patient cohort, resulting in a substantial domain shift. While systematic differences exist in contrast phases, convolution kernels, lesion burdens, and institutional sources, the anatomical structure and segmentation task remain consistent, rendering this dataset pair highly suitable for rigorous TTA evaluation.

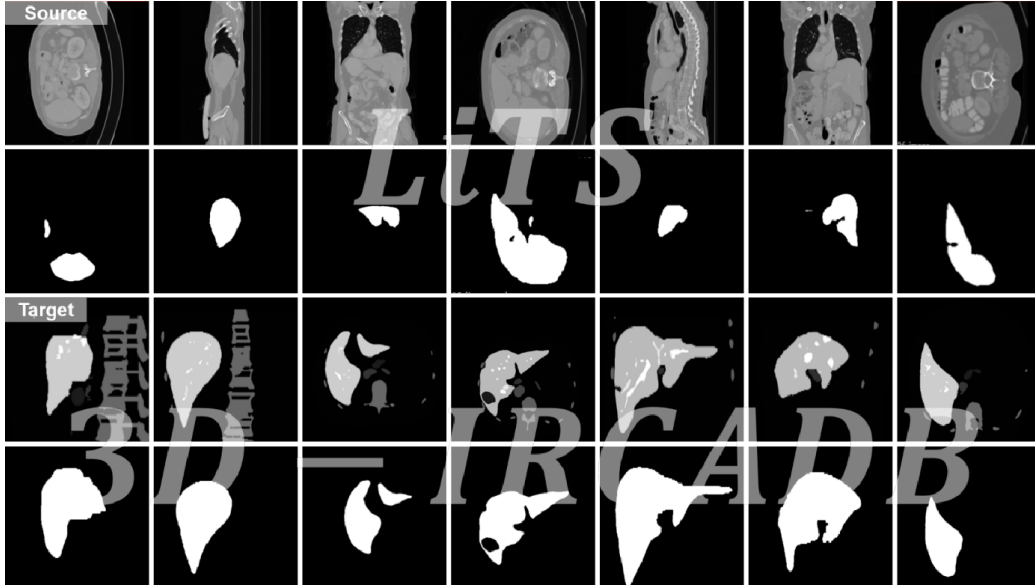


Figure 5: **CT volumes for liver segmentation.** Cases from the source dataset LiTS and the target dataset 3D-IRCADB. For the source dataset area, the first row displays CT images and liver-tumor annotations from anatomical views, including axial and coronal planes; the second row presents the corresponding annotation masks. The bottom section shows the equivalent CT images and annotations for the target dataset.

3.1.3. OCT-based fundus Segmentation Datasets

Source domain combined: MESSIDOR-BASE1, BASE2, and BASE3 [21], Target domain combined: Magrabia and BinRushed [21], as shown in Fig. 6. The

RIGA+ collection aggregates fundus images from diverse sources. The MESSIDOR series, acquired with various cameras and workflows, offers a large set of consistently annotated images. Magrabia and BinRushed, collected in Middle Eastern institutions, exhibit marked differences in equipment, illumination, and demographics. We merge glaucomatous fundus images and their pixel-level segmentation annotations from MESSIDOR-BASE1, BASE2, and BASE3 and refer to this dataset as RIGA+ (MES), comprising 454 cases, to form the source domain. We combine Magrabia and BinRushed as RIGA+ (MB), totaling 290 cases, as the target domain. This setup ensures a stable domain shift within the same modality and anatomical region, enabling robust evaluation of TTA methods for fundus segmentation.

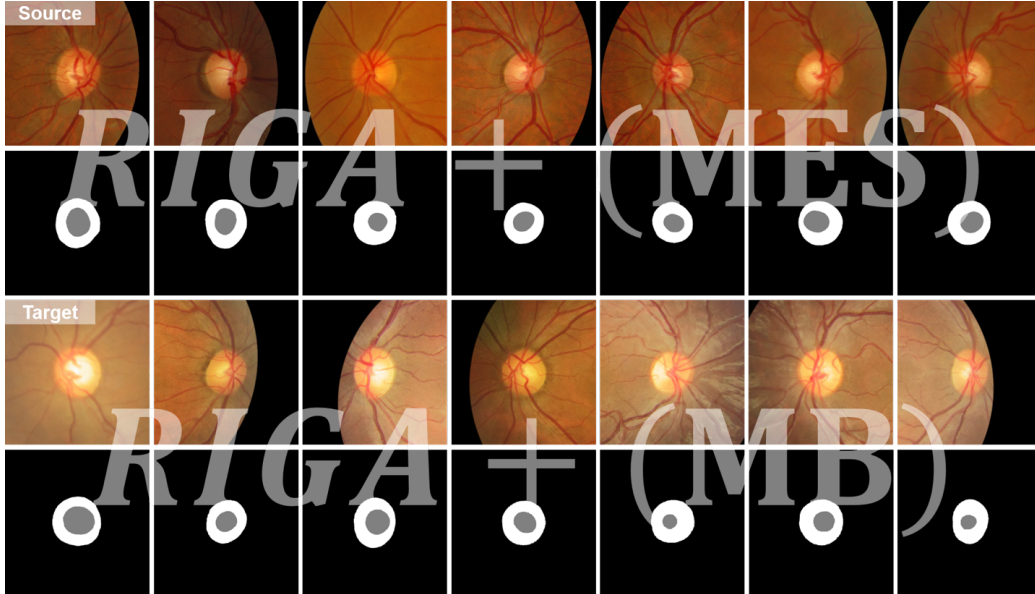


Figure 6: **Fundus images and annotations for fundus segmentation.** Cases from the source dataset RIGA+ (MES) and the target dataset RIGA+ (MB). In the top source dataset section, the first row displays original glaucomatous fundus images and pixel-level segmentation annotations combined from MESSIDOR-BASE1, BASE2, and BASE3; the second row presents the corresponding annotation masks. The bottom target dataset section shows the combined fundus images and annotations from Magrabia and BinRushed.

3.1.4. Ultrasound-based Thyroid Nodule Segmentation Datasets

Source domain: TN3K [26], Target domain: DDTI [27], as shown in Fig. 7. TN3K is a large-scale, multi-center ultrasound dataset of thyroid nodules, com-

prising 3,493 images from 2,421 patients. It was collected across diverse probe brands and acquisition workflows, making it a suitable starting point for multi-source generalization. We adopt its training split of 2879 images as the source domain. DDTI, in contrast, originates from a single device with 637 images showing more concentrated sources and distinct distributions of image resolution and appearance. The two datasets differ markedly in probe models, gain, and dynamic range configurations, as well as in image texture statistics, yet they share the identical task of nodule segmentation, thereby constituting a compelling testbed for TTA in ultrasound scenarios.

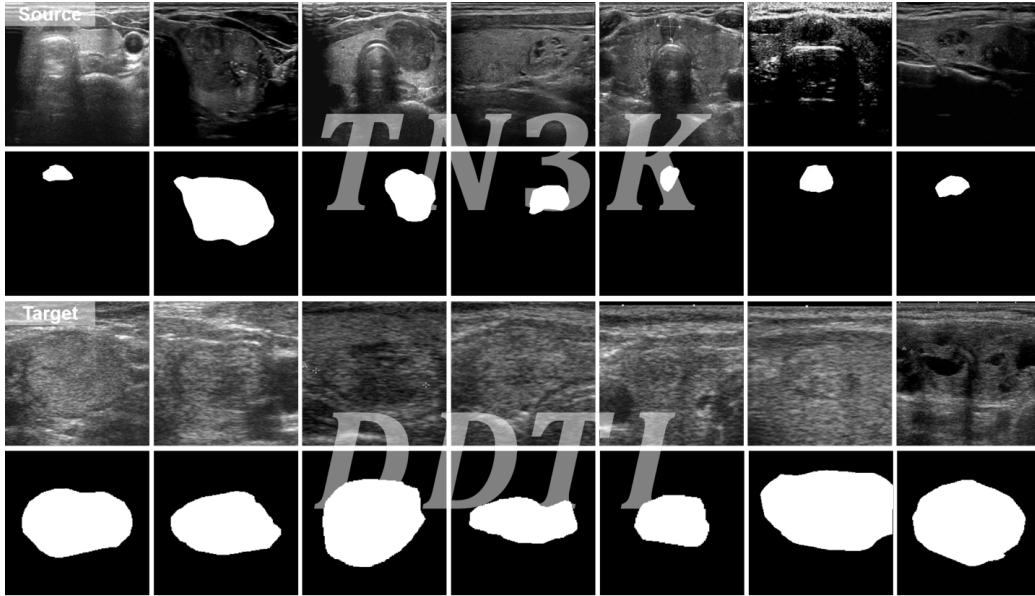


Figure 7: **Ultrasound images for thyroid nodule segmentation.** Cases from the source dataset TN3K and the target dataset DDTI. For the source dataset, the first row displays ultrasound images and thyroid nodule annotations; the second row presents the corresponding annotation masks. The bottom section shows the equivalent ultrasound images and annotations for the target dataset.

3.1.5. Pathology-based Colorectal Adenocarcinoma Tissue Segmentation Datasets

Source domain: CRAG [22], Target domain: GlaS [23], as shown in Fig. 8. CRAG contains 213 adenocarcinoma slides scanned at 20× magnification with detailed instance-level annotations, whereas GlaS comprises 165 images derived from 16 slides, reflecting pronounced variations in staining batches, scanners, magnifications, and tissue morphology. We harmonize the labels of the two datasets

into binary masks with foreground and background classes, and we employ sliding-window cropping with a 256×256 size and 50% overlap to accommodate substantial resolution differences. After discarding all background patches with less than 1% foreground ratio, we obtain 14,015 image-mask pairs from CRAG and 2,208 pairs from GlaS. This dataset pair, while maintaining task consistency, exhibits substantial appearance and texture domain shifts, making it an ideal testing ground for pathology segmentation under real-world TTA settings.

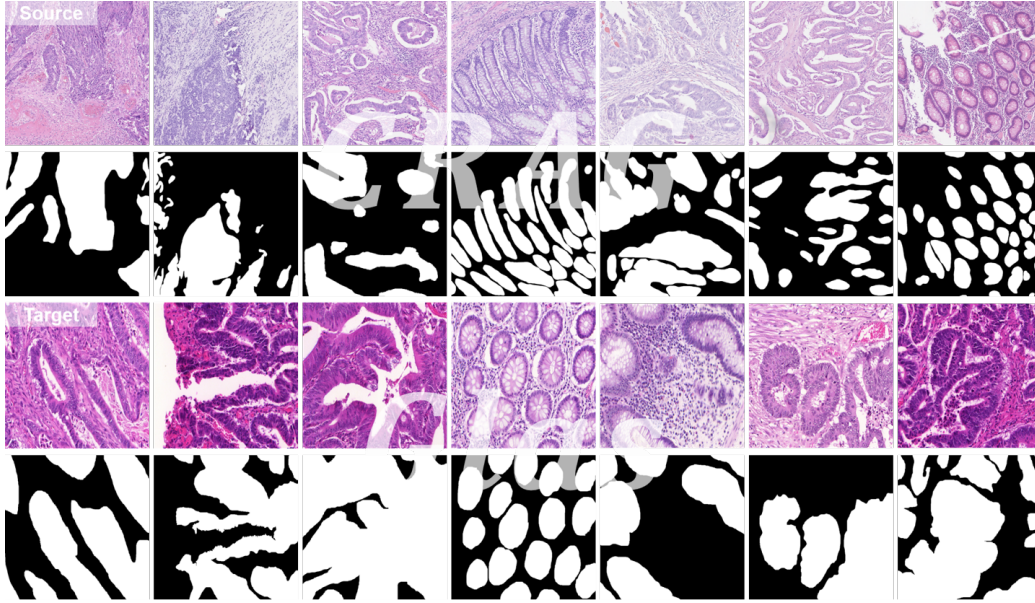


Figure 8: **Histopathological images for colorectal adenocarcinoma tissue segmentation.** Cases from the source dataset CRAG and the target dataset GlaS. For the source dataset area, the first row displays 256×256 cropped images after sliding-window processing and binary segmentation, with foreground as tumor tissue and background as normal tissue; the second row presents the corresponding annotation masks. The bottom section shows the equivalent cropped images and annotations for the target dataset.

3.1.6. Dermoscopy-based Melanoma Segmentation Datasets

Source domain: ISIC-2017 [66], Target domain: PH² [20], as shown in Fig. 9. ISIC-2017, released by the International Skin Imaging Collaboration, is a large-scale and diverse dermoscopy dataset comprising 2,000 training images, 150 validation images, and 600 test images. We use its 2,000 training images as the source domain. PH², collected at a hospital in Lisbon, Portugal, includes 200 dermoscopy images with meticulous annotations, obtained from a single device and a narrower

patient base. The two datasets show stable differences in camera settings, illumination, skin tones, and lesion spectra, leading to pronounced inter-dataset shifts despite sharing the same task: lesion segmentation. This pair provides a benchmark to assess the benefits and risks of TTA in small target domains.

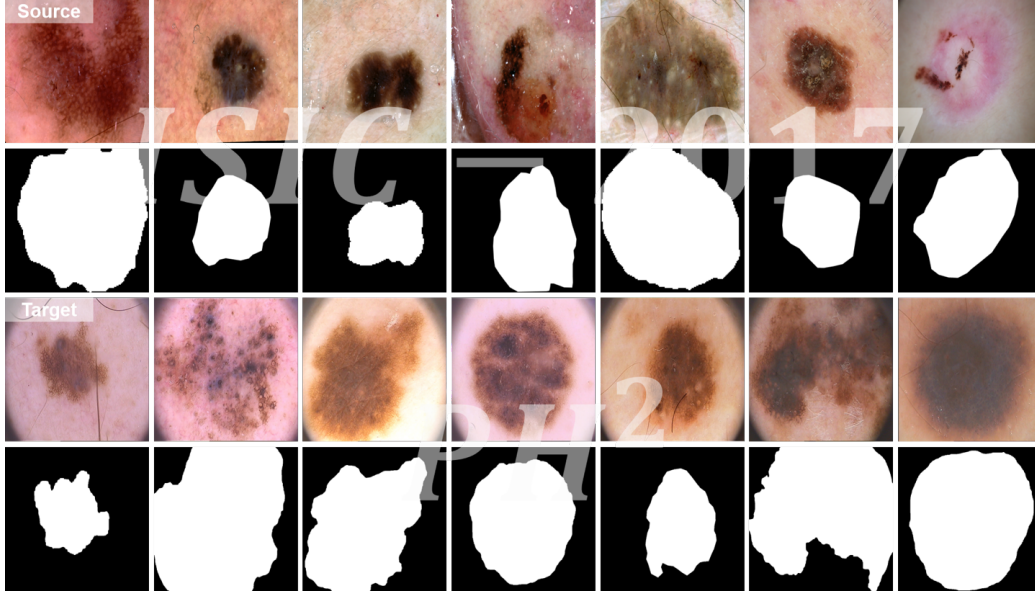


Figure 9: **Dermoscopy images for melanoma segmentation.** Cases from the source dataset ISIC-2017 and the target dataset PH². The first row displays dermoscopy images and melanoma lesion annotations; the second row presents the corresponding annotation masks. The bottom rows show the equivalent dermoscopy images and annotations for the target dataset.

3.1.7. Chest X-ray-based Lung Segmentation Datasets

Source domain: Shenzhen (SZ-CXR) [25], Target domain: Montgomery CXR Set [24], as shown in Fig. 10. Source domain: Shenzhen (SZ-CXR) [25], Target domain: Montgomery CXR Set [24], as shown in Fig. 10. The SZ-CXR dataset, collected at Shenzhen Third People’s Hospital, comprises 662 chest radiographs with lung-field masks, all of which are used as source-domain training data. The Montgomery dataset, curated by the Montgomery County Department of Health and NIH, contains 138 posteroanterior chest radiographs with lung masks, used as target test data. Pronounced differences exist in population demographics, imaging devices, and resolution distributions between Chinese and American cohorts. Under identical modality and anatomical focus, this dataset pair provides cross-

national and cross-device shifts, forming a canonical setup for TTA evaluation in chest X-ray segmentation.

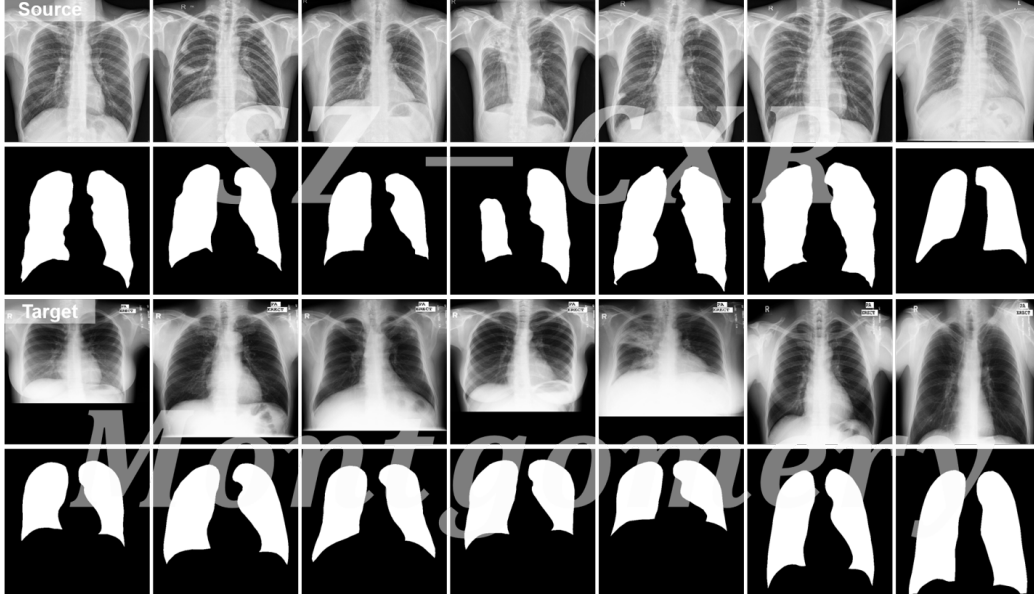


Figure 10: **Chest X-ray images for lung segmentation.** Cases from the source dataset SZ-CXR and the target dataset Montgomery County CXR Set. For the source dataset area, the first row displays posteroanterior chest radiographs and lung field annotations; the second row presents the corresponding annotation masks. The bottom section shows the equivalent chest X-ray images and annotations for the target dataset.

3.1.8. Summary of Datasets

Fig. 3 illustrates representative examples of the seven domain-shift pairs, while Table 1 summarizes their modalities, tasks, dimensions, and repartitioning details. Collectively, these seven source-to-target combinations adhere strictly to the principles of modality and task consistency with matched categories and annotation granularity, while simultaneously exhibiting systematic discrepancies across institutions, devices, and protocols, populations, or preparation pipelines. Moreover, they reflect the practical deployment setting characterized by large-sample source domains and small-sample target domains.

The evaluation is organized into two categories. For MRI brain tumor segmentation, the metrics are computed separately for ET, CT, and WT regions in accordance with the BraTS evaluation protocol, allowing performance to be captured across distinct tumor components. For other modalities such as CT, OCT, and US,

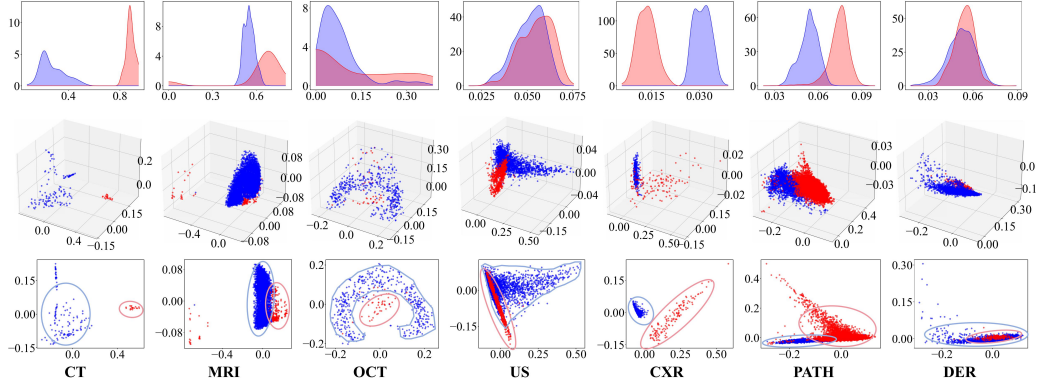


Figure 11: **Multi-perspective visualization of feature distributions across heterogeneous domains.** Row 1 illustrates the marginal frequency distribution of feature values, with the abscissa and ordinate denoting feature magnitude and frequency, respectively. Rows 2 and 3 present the spatial distribution of features in the low-dimensional embedding space obtained via dimensionality reduction. Source-domain and target-domain samples are encoded in red and blue, respectively.

Table 1: **Source and target dataset pairs for the TTA benchmark in medical image segmentation.** The Category indicates the number of segmentation classes after aligning the annotation categories between the source and target domain datasets. "Binary" refers to binary segmentation tasks, while "4-Class" refers to four-class segmentation tasks, which include the background class. The Reprocess column uses ✓ to indicate reprocessed datasets and ✗ to indicate datasets used in their original form.

Modal	Dataset	Domain	Category	Quantity	Year	Reprocess	Source
MRI	BraTS-GLI2024 [65]	Source	4-Class	1k~2k	2024	✗	Link
	BraTS-SSA [16]	Target	4-Class	60	2023	✗	Link
CT	LiTS [17]	Source	Binary	<0.5k	2017	✓	Link
	3D-IRCADB [18]	Target	Binary	20	2010	✓	Link
DER	ISIC-2017 [66]	Source	Binary	2k~3k	2017	✗	Link
	PH ² [20]	Target	Binary	<0.5k	2015	✗	Link
Ultrasound	TN3K [26]	Source	Binary	2k~3k	2021	✗	Link
	DDTI [27]	Target	Binary	0.5k~1k	2015	✗	Link
X-Ray	SZ-CXR [25]	Source	Binary	0.5k~1k	2018	✗	Link
	Montgomery [24]	Target	Binary	<0.5k	2021	✗	Link
Fundus	RIGA+(MES) [21]	Source	Binary	<0.5k	2021	✓	Link
	RIGA+(MB) [21]	Target	Binary	<0.5k	2021	✓	Link
Histopathology	CRAE [22]	Source	Binary	<0.5k	2019	✓	Link
	Glas [23]	Target	Binary	<0.5k	2017	✓	Link

which predominantly involve binary segmentation of foreground and background, the metrics are computed on the foreground regions to assess both target recognition and boundary stability. This unified and fine-grained metric design enables a more precise characterization of the adaptability and robustness of different TTA methods across diverse modalities and tasks.

3.2. Implementation Details

3.2.1. Unified Pretraining for Medical Image Segmentation

To eliminate confounding factors introduced by backbone differences, all source domain models are pretrained using a standard UNet architecture. The 2D and 3D configurations, including input sizes, encoder depths, channel widths, output channels, and BraTS-specific settings, are listed in Table 2. These configurations are consistently applied across modalities and remain reusable, with only minimal adjustments required for dimensionality in 2D or 3D, as well as the number of classes. This ensures that TTA comparisons have equivalent starting points.

Table 2: **Unified architectural configurations of UNet backbones for pretraining.** UNet2d and UNet3d share the same encoder depth and base width C_0 , while the BraTS-specific UNet3d variant uses four input and output channels for multi-class tumor segmentation.

Components	UNet2d	UNet3d	UNet3d (BraTS)
Input size	256×256	$128 \times 128 \times 128$	$128 \times 128 \times 128$
Input channel	1	1	4
Output channel	2	2	4
Encoder blocks	5	5	5
Block depth D	256	256	256
Basic width C_0	32	32	32
Channels of encoders	$C_0-2C_0-4C_0-8C_0-16C_0$	$C_0-2C_0-4C_0-8C_0-16C_0$	$C_0-2C_0-4C_0-8C_0-16C_0$

3.2.2. Unified Implementation of the TTA for Medical Image Segmentation

To enable a fair comparison among various TTA approaches, we adopt UNet as the unified, pre-trained backbone, without altering its network topology. Adjustments are made solely during testing, following each method’s prescribed strategy to update permissible parameters, statistics, and adaptation modules. All methods share the following standard settings: (1) identical source domain initialization and data preprocessing. (2) identical target domain input sequences with streaming or mini-batch configurations. (3) a unified evaluation interface and standardized metric system. (4) a strict prohibition on accessing source domain data or labels, thereby preventing both explicit and implicit leakage.

4. Experiments

To systematically and comprehensively evaluate TTA strategies for medical image segmentation, we benchmarked twenty representative methods across seven imaging modalities. These methods encompass four major paradigms: Input-level Transformation, Feature-level Alignment, Output-level Regularization, and Prior Estimation, each with five representative methods. Such broad coverage ensures that the experimental findings possess strong generalizability and reference value across diverse imaging modalities and segmentation tasks. It is worth noting that many of these methods were initially designed and optimized for a specific data dimensionality. For instance, FSM [49] primarily targets 2D images, whereas DANN [30] is tailored for 3D data. Given that our benchmark encompasses both 2D and 3D medical images, we extended these single-dimensional methods to support both settings, enabling fair and comprehensive cross-dimensional evaluation. Tables 3 and 4 summarize all investigated methods, along with the key principles behind their dimensional extensions.

4.1. Overall Performance Analysis

To assess domain quality gaps and distribution shifts, we split the source domain into training and validation at an 8:2 ratio for in-domain evaluation and used the entire target domain for out-of-domain testing. Under a unified UNet backbone, all four TTA paradigms achieved higher average segmentation accuracy than the no-adaptation baseline across modalities, confirming the practical value of TTA for medical image segmentation. As summarized in Table 5, Input-level Transformation methods achieved the highest Dice in most modalities. In ultrasound, Prior Estimation was slightly better; yet, Input-level methods still outperformed the remaining paradigms by a clear margin. From the perspective of HD95, a boundary-sensitive complement to Dice, the top performers did not concentrate within a single paradigm. In OCT, pathology, and MRI, the leading methods often came from different categories. These patterns suggest that appearance alignment effectively reduces intensity and texture discrepancies but cannot, by itself, resolve complex deformations and boundary noise. We also observe task-dependent failures: certain methods degrade on specific segmentation tasks and can even underperform the no-adaptation baseline, revealing limited generalization and underscoring the need for careful paradigm selection.

More specifically, Tables 6 through 12 report the performance of twenty methods across seven imaging modalities: MRI, CT, OCT, US, PATH, DER, and CXR. While the mean performance of each paradigm exceeds the non-adaptive

Table 3: **Implementation details of compared TTA methods.** Covering their paradigms, original modalities, and dimensions, with paradigms marked by symbols: ♠ for Input-level Translation, ♣ for Feature-level Alignment, ♥ for Output-level Regularization, and ♦ for Prior Estimation.

Method	Paradigm	Original Modal	Original Imension
♠ SFDA-FSM [49]	Input-level Translation	Endoscope	2D
♠ RSA [29]	Input-level Translation	MRI	2D
♠ DL-TTA [71]	Input-level Translation	PATH	2D
♠ STDR [72]	Input-level Translation	MRI	2D
♠ AIF-SFDA [73]	Input-level Translation	OCT	2D
♣ DANN [30]	Feature-level Alignment	MRI	3D
♣ Testfit [50]	Feature-level Alignment	CT/PATH	General
♣ GraTa [61]	Feature-level Alignment	OCT	2D
♣ UDA-MIMA [52]	Feature-level Alignment	MRI/CT	3D
♣ DeTTA [62]	Feature-level Alignment	CT	2D
♥ TENT [33]	Output-level Regularization	General Image	General
♥ DG-TTA [31]	Output-level Regularization	MRI/CT	3D
♥ SaTTCA [57]	Output-level Regularization	CT	3D
♥ UPL-SFDA [58]	Output-level Regularization	CMR/MRI	General
♥ SmART [70]	Output-level Regularization	MRI	3D
♦ PASS [74]	Prior Estimation	OCT	2D
♦ ProSFDA [32]	Prior Estimation	OCT	2D
♦ AdaMI [51]	Prior Estimation	MRI/CT	3D
♦ ExploringTTA [59]	Prior Estimation	US	3D
♦ VPTTA [63]	Prior Estimation	OCT	2D

baseline, within each paradigm, some methods drop below the baseline on specific modality-task pairs, revealing the limits of these TTA methods and their inability to generalize uniformly across medical segmentation problems. Overall, Input-level Transformation shows the most outstanding stability and reproducibility across modalities. For example, SFDA-FSM achieves the top Dice score in multiple modalities, including MRI, CXR, DER, OCT, PATH, and US. Although it does not lead in CT, it still surpasses the non-adaptive baseline without TTA. On average, SFDA-FSM delivers an absolute Dice increase of 0.118, with the most significant gain observed in the US modality, where the Dice rises by approximately 0.35, and a steady improvement in CT, where the increase is close to 0.12. In a similar vein, RSA and TestFit provide consistent gains across several modalities. While less stable than FSM, they achieve top or near-top HD95 on boundary-sensitive tasks such as CXR and DER, highlighting the effectiveness of

Table 4: **Implementation details of compared TTA methods.** Covering method and cross-dimensional extension, marked by symbols: ♠ for Input-level Translation, ♣ for Feature-level Alignment, ♥ for Output-level Regularization, ♦ for Prior Estimation.

Method	Cross-Dimensional Method Extension
♠ SFDA-FSM [49]	2D to 3D: The <code>fft2/i fft2</code> operations are replaced with <code>fftn/i fftn</code> . The low-frequency square replacement is upgraded to a cubic structure. Case-level mask sharing and sliding-window fusion are introduced to ensure inter-slice consistency.
♠ RSA [29]	2D to 3D: The diffusion U-Net is extended to a 3D architecture, with noise and timestep embeddings augmented along the depth dimension. Training is on volumetric patches, and inference incorporates consistency constraints to ensure coherence.
♠ DL-TTA [71]	2D to 3D: The reconstructor is replaced by a 3D autoencoder or generator. Patch-level reconstructions are normalized and aggregated into volumetric outputs, which are then fed into the pretrained segmentation network.
♠ STDR [72]	2D to 3D: Style and statistical alignment are extended from image level to volume level, with statistics shared to ensure inter-slice consistency and correspondence.
♠ AIF-SFDA [73]	2D to 3D: Adaptive frequencies are computed and aligned in the 3D spectrum, frequency-band cubes are mixed and stabilized through case-level normalization to maintain volumetric appearance consistency.
♣ DANN [30]	3D to 2D: The domain discriminator and feature extractor are downgraded from 3D to 2D convolutions. Domain labels are shared across slices, with case-level feature aggregation and gradient reversal applied.
♣ Testfit [50]	None: Native support for both 2D and 3D.
♣ GraTa [61]	2D to 3D: Gradient alignment across multiple enhanced branches on the 3D volume, update normalized affine parameters, and accumulate case-level stable direction estimates.
♣ UDA-MIMA [52]	3D to 2D: Mutual information estimation is reduced from volumetric to slice level. A 2.5D neighborhood mechanism is introduced to preserve depth context.
♣ DeTTA [62]	2D to 3D: Single-slice masks are expanded into patch-level random masks. Multiple generated inputs are fused volumetrically, with case-level parameter resets preventing cross-domain interference.
♥ TENT [33]	None: Natively compatible with both 2D and 3D settings.
♥ DG-TTA [31]	3D to 2D: Dual-branch consistency is reduced from volumetric to slice level. EMA teachers and consistency masks are shared to enforce smoothness across slices.
♥ SaTTCA [57]	3D to 2D: 3D click prompts are projected into 2D Gaussian heatmaps. Radii are rescaled based on the pixel spacing and jointly optimized with entropy regularization.
♥ UPL-SFDA [58]	None: Natively compatible with both 2D and 3D settings.
♥ SmarT [70]	3D to 2D: The loss function is adapted from a $3 \times 3 \times 3$ region to a 3×3 region, and the connectivity criterion is redefined for 2D scenarios.
♦ PASS [74]	2D to 3D: Shape and style prompts are elevated to 3D tokens, shared across three orthogonal planes, and injected into the volumetric encoder through cross-attention.
♦ ProSFDA [32]	2D to 3D: 3D class prototypes are constructed from high-confidence voxels. Volumetric prototype recalibration and logits reweighting are performed, followed by case-level aggregation to enhance stability.
♦ AdaMI [51]	3D to 2D: The original network processes 2D slices from 3D images and can be directly adapted for inference on standalone 2D inputs.
♦ ExploringTTA [59]	3D to 2D: Structural and topological priors are reduced from volumetric to slice-level boundary and connectivity priors, with inter-slice smoothing ensuring consistency.
♦ VPTTA [63]	2D to 3D: The 2D FFT is replaced with 3D FFT. Batch normalization statistics are extended to 3D, and prompts are redesigned as 3D tensors.

Table 5: **Paradigm taxonomy and applicability lineage across modalities.** We categorize TTA methods by their locus of operation and, using evaluations across modalities, organs, and tasks, derive a *lineage map* that highlights effective, partially effective, and ineffective regimes, with effectiveness defined as Dice increasing and HD95 decreasing. The table reports Dice and HD95 with significance at $p < 0.05$.

Method	MRI		CT		OCT		US		PATH		DER		CXR	
	Dice↑	HD95↓	Dice↑	HD95↓	Dice↑	HD95↓	Dice↑	HD95↓	Dice↑	HD95↓	Dice↑	HD95↓	Dice↑	HD95↓
Intra-Domain	0.7049	3.33	0.8949	1.63	0.9550	0.41	0.7215	31.01	0.8360	32.66	0.9280	2.33	0.9557	1.02
Target-Domain (w/o TTA)	0.5951	23.04	0.6595	13.91	0.8951	13.32	0.4811	62.78	0.8026	33.66	0.8532	13.78	0.9093	5.15
▲ Input-level Translation	0.6411	13.72	0.7156	10.32	0.9075	9.77	0.5757	49.26	0.8107	31.63	0.8672	12.68	0.9308	3.36
▲ Feature-level Alignment	0.6269	9.67	0.6874	13.56	0.9000	10.28	0.5539	50.83	0.7930	35.39	0.8404	15.59	0.9194	4.69
♥ Output-level Regularization	0.6399	7.98	0.7240	14.24	0.9051	8.02	0.5235	58.03	0.8008	32.14	0.8606	13.92	0.9284	3.87
◆ Prior Estimation	0.6395	6.96	0.6969	14.01	0.8981	11.17	0.5776	46.95	0.7981	33.98	0.8526	15.78	0.9285	3.66

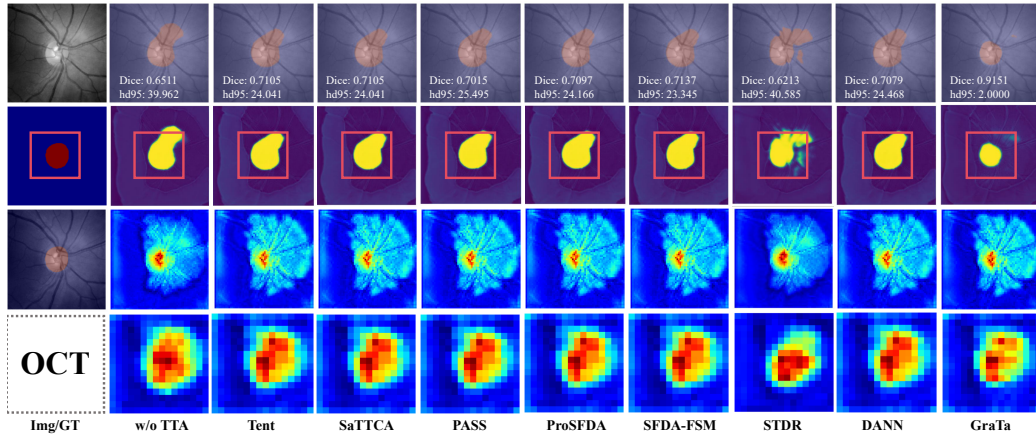


Figure 12: **Visual comparison of TTA efficacy across different methods on OCT datasets.** Rows 1 and 2 illustrate the segmentation outcomes. Rows 3 and 4 depict changes in features extracted by the Encoder after TTA. GT denotes the ground-truth.

the strategy that first aligns appearance and then performs inference.

4.2. TTA Under Mild Domain Shift

In tasks characterized by relatively mild domain shifts, such as OCT, PATH, DER, and CXR, Input-level methods generally achieved superior results, often excelling across both Dice and HD95 metrics. In addition, Prior Estimation methods exhibited advantages in scenarios with stable structural priors, indicating that anatomical regularities can enhance model adaptability.

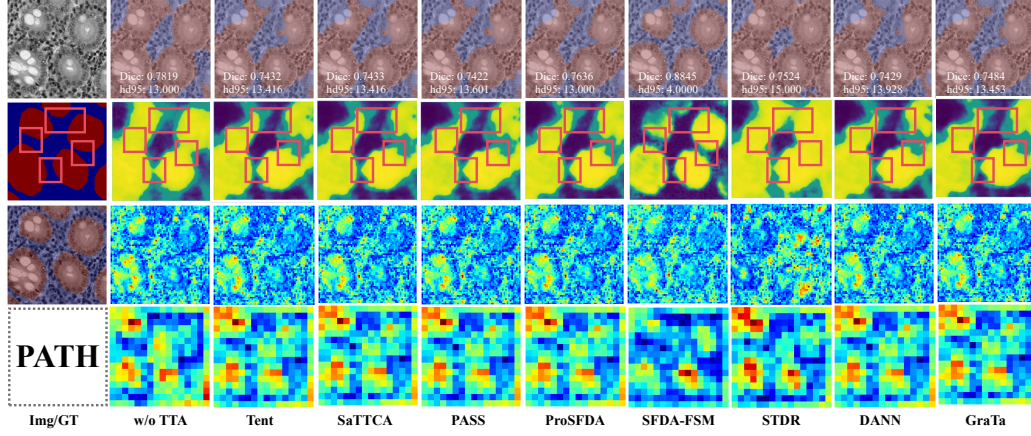


Figure 13: **Visual comparison of TTA efficacy across different methods on DER datasets.** Rows 1 and 2 illustrate the segmentation outcomes. Rows 3 and 4 depict changes in features extracted by the Encoder after TTA. GT denotes the ground-truth.

4.2.1. Optical Coherence Tomography Modality

As shown in Table 6, in the OCT task, 15 out of 20 methods outperformed the baseline Dice score of 0.8951, and 17 improved the HD95. Among them, SFDA-FSM achieved the best performance, with Dice increasing to 0.9319, a 4.1% gain, and HD95 decreasing to 9.87, a 3.45% reduction, indicating that appearance and frequency-domain alignment effectively alleviate intensity and texture discrepancies. In contrast, UPL-SFDA excelled in boundary accuracy, with HD95 reduced to 6.40, a 52 percent decrease, demonstrating the efficacy of Output-level Regularization in suppressing coarse deviations in predictive distributions. Feature alignment methods, such as UDA-MIMA and TestFit, and partially mismatched Prior Estimation approaches, such as VPTTA, suffer from performance degradation in this modality due to noisy pseudo-labels, suboptimal learning rates, and prior misalignment, which lead to oversmoothing and boundary degradation. Fig. 12 illustrates the effects of various TTA methods on the OCT task, as well as their influence on the encoder. Most models demonstrated positive outcomes.

4.2.2. Histopathology Modality

As reported in the Table 7, results on the Histopathology task showed greater divergence: only 6 of 20 methods improved Dice scores, while 13 methods experienced degradation. SFDA-FSM again achieved the best Dice score of 0.8713, representing an 8.6% increase, whereas UPL-SFDA yielded the most remarkable HD95 improvement, a 47.1% reduction. However, most Prior Estimation meth-

Table 6: **Comprehensive performance comparison of TTA methods on OCT Fundus Image Segmentation.** Results colored in **red**, **orange**, and **yellow** denote the best, second-best, and third-best performances, respectively. Rows with a **gray** background indicate Dice values lower than that of the Target-Domain(w/o TTA). Symbols denote paradigms: ♠ for Input-level Translation, ♣ for Feature-level Alignment, ♥ for Output-level Regularization, ♦ for Prior Estimation. All improvements are statistically significant at $p < 0.05$.

Method	Dice $\uparrow$	HD95 $\downarrow$	JI $\uparrow$	Sen $\uparrow$	PPV $\uparrow$
Intra-Domain	0.9550$\pm$0.0007	0.41$\pm$0.01	0.9145$\pm$0.0012	0.9607$\pm$0.0033	0.9512$\pm$0.0023
Target-Domain(w/o TTA)	0.8951$\pm$0.0951	13.32$\pm$28.89	0.8217$\pm$0.1392	0.9611$\pm$0.0387	0.8529$\pm$0.1516
♠ RSA (2021, MICCAI) [29]	0.8981 $\pm$ 0.0879	13.01 $\pm$ 23.45	0.8362 $\pm$ 0.1303	0.9447 $\pm$ 0.0531	0.8621 $\pm$ 0.1329
♠ DL-TTA (2022, TMI) [71]	0.8927 $\pm$ 0.1033	10.01 $\pm$ 14.25	0.8247 $\pm$ 0.1441	0.9523 $\pm$ 0.0415	0.8583 $\pm$ 0.1421
♠ SFDA-FSM (2022, MIA) [49]	0.9319 $\pm$ 0.0483	9.87 $\pm$ 26.50	0.8761 $\pm$ 0.0797	0.9652 $\pm$ 0.0364	0.9051 $\pm$ 0.0797
♠ STDR (2024, TMI) [72]	0.9171 $\pm$ 0.0862	7.85 $\pm$ 14.68	0.8589 $\pm$ 0.1312	0.9010 $\pm$ 0.0495	0.8617 $\pm$ 0.1441
♠ AIF-SFDA (2025, AAAI) [73]	0.8978 $\pm$ 0.1059	8.12 $\pm$ 24.09	0.8245 $\pm$ 0.1466	0.9605 $\pm$ 0.0774	0.8569 $\pm$ 0.1495
♣ DANN (2016, JMLR) [30]	0.9015 $\pm$ 0.0915	9.30 $\pm$ 22.15	0.8315 $\pm$ 0.1313	0.9550 $\pm$ 0.0608	0.8664 $\pm$ 0.1373
♣ DeTTA (2024, WACV) [62]	0.9069 $\pm$ 0.0855	9.52 $\pm$ 18.35	0.8392 $\pm$ 0.1233	0.9467 $\pm$ 0.0676	0.8836 $\pm$ 0.1305
♣ TestFit (2024, MIA) [50]	0.8930 $\pm$ 0.0951	13.83 $\pm$ 28.88	0.8186 $\pm$ 0.1391	0.9601 $\pm$ 0.0854	0.8505 $\pm$ 0.1516
♣ UDA-MIMA (2024, CMPB) [52]	0.8911 $\pm$ 0.1074	11.99 $\pm$ 25.17	0.8170 $\pm$ 0.1414	0.9498 $\pm$ 0.0734	0.8521 $\pm$ 0.1502
♣ GraTa (2025, AAAI) [61]	0.9073 $\pm$ 0.0965	6.75 $\pm$ 18.78	0.8410 $\pm$ 0.1232	0.9278 $\pm$ 0.0822	0.8981 $\pm$ 0.1307
♥ TENT (2021, ICLR) [33]	0.9022 $\pm$ 0.0921	7.22 $\pm$ 19.12	0.8330 $\pm$ 0.1325	0.9554 $\pm$ 0.0598	0.8682 $\pm$ 0.1392
♥ UPL-SFDA (2023, TMI) [58]	0.9099 $\pm$ 0.0849	6.40 $\pm$ 18.33	0.8441 $\pm$ 0.1226	0.9435 $\pm$ 0.0690	0.8911 $\pm$ 0.1285
♥ DG-TTA (2023, arXiv) [31]	0.9053 $\pm$ 0.0748	9.96 $\pm$ 23.44	0.8347 $\pm$ 0.1140	0.9295 $\pm$ 0.0665	0.8967 $\pm$ 0.1285
♥ SaTTCA (2023, MICCAI) [57]	0.9030 $\pm$ 0.0823	8.52 $\pm$ 8.03	0.8330 $\pm$ 0.0389	0.9452 $\pm$ 0.0364	0.8782 $\pm$ 0.0541
♥ SmaRT (2025, arXiv) [70]	0.8993 $\pm$ 0.0924	7.49 $\pm$ 16.54	0.8280 $\pm$ 0.1327	0.9738 $\pm$ 0.0348	0.8485 $\pm$ 0.1401
♦ AdaMI (2022, MICCAI) [51]	0.9013 $\pm$ 0.0943	10.25 $\pm$ 22.16	0.8285 $\pm$ 0.1337	0.9631 $\pm$ 0.0425	0.8624 $\pm$ 0.1461
♦ PASS (2024, TMI) [74]	0.9009 $\pm$ 0.0920	9.33 $\pm$ 22.15	0.8307 $\pm$ 0.1320	0.9552 $\pm$ 0.0609	0.8645 $\pm$ 0.1380
♦ VPTTA (2024, CVPR) [63]	0.8975 $\pm$ 0.0993	13.83 $\pm$ 28.89	0.8229 $\pm$ 0.1432	0.9643 $\pm$ 0.0426	0.8548 $\pm$ 0.1554
♦ ExploringTTA (2025, ISBI) [59]	0.8934 $\pm$ 0.1017	11.11 $\pm$ 24.09	0.8203 $\pm$ 0.1427	0.9567 $\pm$ 0.0737	0.8527 $\pm$ 0.1458
♦ ProSFDA (2026, PR) [32]	0.8973 $\pm$ 0.0943	11.33 $\pm$ 24.18	0.8254 $\pm$ 0.1360	0.9616 $\pm$ 0.0458	0.8557 $\pm$ 0.1460

ods, such as PASS, AdaMI, and VPTTA, performed poorly in this modality, highlighting the substantial impact of staining variations and scanning heterogeneity on prior-dependent strategies. Notably, our benchmark employs a patch-level evaluation protocol, which effectively circumvents the limitations and instability associated with WSI-level adaptation and GPU memory constraints. This also underscores that WSI-level TTA remains a critical open challenge. Fig. 13 presents a visualization of the effects of the TTA method on the Histopathology task, illustrating the changes in the feature extraction capabilities of the encoder. The segmentation results indicate a general refinement of the boundaries.

4.2.3. Dermoscopy Modality

In DER, as summarized in the Table 8, SFDA-FSM continued to excel with a Dice score of 0.9258, an improvement of 8.5 percent, and HD95 of 3.58, a

Table 7: **Comprehensive performance comparison of TTA methods on Colorectal Gland Segmentation in Histopathology Images.** Results colored in red, orange, and yellow denote the best, second-best, and third-best performances, respectively. Rows with a gray background indicate Dice values lower than that of the Target-Domain(w/o TTA). Symbols denote paradigms: ♠ for Input-level Translation, ♣ for Feature-level Alignment, ♥ for Output-level Regularization, ♦ for Prior Estimation. All improvements are statistically significant at $p < 0.05$.

Method	Dice ↑	HD95 ↓	JI ↑	Sen ↑	PPV ↑
Intra-Domain	0.8360±0.1966	32.66±56.88	0.7562±0.2288	0.9583±0.0826	0.7834±0.2342
Target-Domain (w/o TTA)	0.8026±0.1782	33.66±49.01	0.7001±0.2034	0.9422±0.1119	0.7307±0.2063
♠ RSA (2021, MICCAI) [29]	0.7810±0.1745	35.72±46.39	0.6691±0.2016	0.9314±0.1196	0.7092±0.2060
♠ DL-TTA (2022, TMI) [71]	0.7935±0.1748	34.62±47.45	0.6867±0.2035	0.9336±0.1095	0.7262±0.2087
♠ SFDA-FSM (2022, MIA) [49]	0.8713±0.1460	19.56±41.32	0.7944±0.1763	0.9142±0.1173	0.8526±0.1709
♠ STDR (2024, TMI) [72]	0.7993±0.1795	34.81±46.23	0.6952±0.2002	0.9579±0.1152	0.7118±0.2247
♠ AIF-SFDA (2025, AAAI) [73]	0.8082±0.1801	33.46±48.66	0.7053±0.2057	0.9397±0.1098	0.7342±0.2086
♣ DANN (2016, JMLR) [30]	0.7794±0.1900	38.37±51.65	0.6758±0.2132	0.9079±0.1262	0.7301±0.2302
♣ DeTTA (2024, WACV) [62]	0.8035±0.1617	35.49±49.71	0.7005±0.1842	0.9238±0.1650	0.7342±0.2209
♣ TestFit (2024, MIA) [50]	0.8037±0.1758	33.45±48.65	0.7010±0.2015	0.9454±0.1056	0.7300±0.2045
♣ UDA-MIMA (2024, CMPB) [52]	0.7890±0.1732	33.16±46.01	0.6795±0.1984	0.8462±0.1685	0.7920±0.2055
♣ GraTa (2025, AAAI) [61]	0.7892±0.1832	36.49±49.71	0.6832±0.2107	0.9139±0.1232	0.7372±0.2193
♥ TENT (2021, ICLR) [33]	0.7890±0.1833	36.50±49.68	0.6830±0.2108	0.9136±0.1236	0.7371±0.2193
♥ UPL-SFDA (2023, TMI) [58]	0.8438±0.1820	17.81±43.01	0.7629±0.2132	0.9435±0.1120	0.8038±0.2331
♥ DG-TTA (2023, arXiv) [31]	0.8018±0.1772	33.76±48.88	0.6987±0.2026	0.9455±0.1074	0.7276±0.2054
♥ SaTTCA (2023, MICCAI) [57]	0.7686±0.1186	40.49±32.23	0.6566±0.1358	0.8958±0.0770	0.7207±0.1568
♥ SmaRT (2025, arXiv) [70]	0.7385±0.1978	43.84±52.81	0.6183±0.2129	0.9160±0.1082	0.6613±0.2298
♦ AdaMI (2022, MICCAI) [51]	0.7930±0.1743	34.21±48.18	0.6854±0.2001	0.8708±0.1518	0.7773±0.2109
♦ PASS (2024, TMI) [74]	0.7938±0.1743	34.12±48.34	0.6866±0.2002	0.8712±0.1511	0.7781±0.2108
♦ VPTTA (2024, CVPR) [63]	0.7938±0.1875	38.45±42.91	0.6876±0.2149	0.9182±0.1274	0.7415±0.2234
♦ ExploringTTA (2025, ISBI) [59]	0.7835±0.1818	36.16±48.91	0.6745±0.2069	0.8743±0.1387	0.7573±0.2231
♦ ProSFDA (2026, PR) [32]	0.8263±0.1647	26.95±43.57	0.7308±0.1944	0.8772±0.1457	0.8173±0.1893

reduction of 74 percent, followed closely by SmaRT. In contrast, UDA-MIMA exhibited substantial degradation, with the Dice score decreasing by 7.3% and the HD95 score increasing by 59%. Overall, Input-level Transformation and Output-level Regularization showed clear advantages in both Dice and HD95, while feature alignment and Prior Estimation tended to fail. This stems from the domain-specific characteristics of dermoscopic imaging, including variations in color calibration and illumination, device-dependent imaging artifacts, and confounders such as hair and glare, which intensify boundary ambiguity. Additionally, the high proportion of small, low-contrast lesions amplifies pseudo-label noise, causing representational drift and undermining fine-grained alignment strategies. Fig. 14 demonstrates the effects of various TTA methods on the DER task, as well as their influence on the encoder. The performance of the different methods varies, exhibiting both positive and negative outcomes.

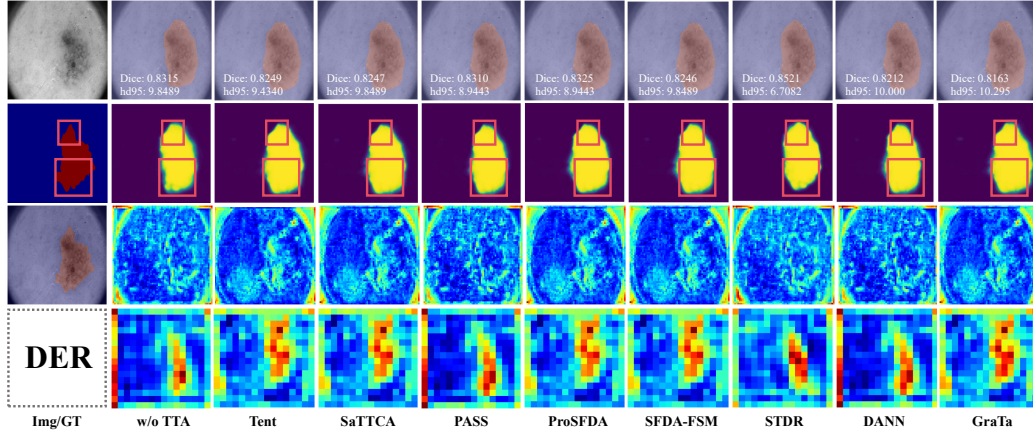


Figure 14: **Visual comparison of TTA efficacy across different methods on DER datasets.** Rows 1 and 2 illustrate the segmentation outcomes. Rows 3 and 4 depict changes in features extracted by the Encoder after TTA. GT denotes the ground-truth.

4.2.4. Chest X-ray Modality

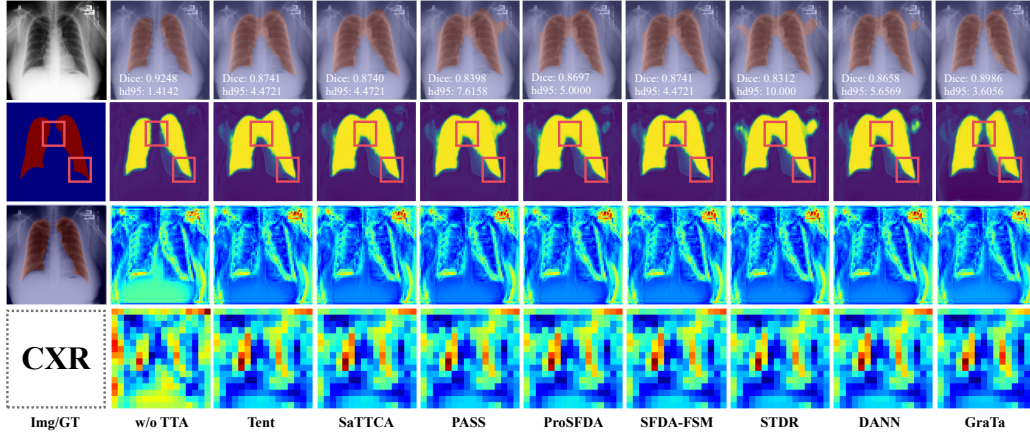


Figure 15: **Visual comparison of TTA efficacy across different methods on CXR datasets.** Rows 1 and 2 illustrate the segmentation outcomes. Rows 3 and 4 depict changes in features extracted by the Encoder after TTA. GT denotes the ground-truth.

In the Chest X-ray task shown in Table 9, most methods achieved substantial improvements, with 17 out of 20 surpassing the baseline Dice score on this challenging benchmark. SFDA-FSM again led the pack with a Dice score of 0.9663, representing an improvement of 6.3 percent, and an HD95 of 0.5576, a reduction of 89 percent. This was followed by ProSFDA, DG-TTA, AdaMI, and SaTTCA.

Table 8: **Comprehensive performance comparison of TTA methods on Melanoma Segmentation in DER Images.** Results colored in red, orange, and yellow denote the best, second-best, and third-best performances, respectively. Rows with a gray background indicate Dice values lower than that of the Target-Domain(w/o TTA). Symbols denote paradigms: ♠ for Input-level Translation, ♣ for Feature-level Alignment, ♥ for Output-level Regularization, ♦ for Prior Estimation. All improvements are statistically significant at $p < 0.05$.

Method	Dice $\uparrow$	HD95 $\downarrow$	JI $\uparrow$	Sen $\uparrow$	PPV $\uparrow$
Intra-Domain	0.9280$\pm$0.0767	2.33$\pm$3.47	0.8723$\pm$0.0979	0.9397$\pm$0.1006	0.9270$\pm$0.0797
Target Domain (w/o TTA)	0.8532$\pm$0.0864	13.78$\pm$13.69	0.7716$\pm$0.1052	0.8527$\pm$0.1323	0.9120$\pm$0.0556
♠ RSA (2021, MICCAI) [29]	0.8517 $\pm$ 0.1747	16.85 $\pm$ 42.47	0.7724 $\pm$ 0.2053	0.8179 $\pm$ 0.2256	0.9387 $\pm$ 0.1134
♠ DL-TTA (2022, TMI) [71]	0.8478 $\pm$ 0.1648	13.64 $\pm$ 25.37	0.7672 $\pm$ 0.1840	0.8576 $\pm$ 0.2318	0.9341 $\pm$ 0.1326
♠ SFDA-FSM (2022, MIA) [49]	0.9258 $\pm$ 0.0604	3.58 $\pm$ 9.86	0.8670 $\pm$ 0.0904	0.8979 $\pm$ 0.0937	0.9667 $\pm$ 0.0619
♠ STDR (2024, TMI) [72]	0.8526 $\pm$ 0.1682	15.65 $\pm$ 36.71	0.7692 $\pm$ 0.1883	0.8540 $\pm$ 0.2106	0.8623 $\pm$ 0.1139
♠ AIF-SFDA (2025, AAAI) [73]	0.8538 $\pm$ 0.1735	13.66 $\pm$ 36.71	0.7749 $\pm$ 0.1936	0.8597 $\pm$ 0.2159	0.8698 $\pm$ 0.1192
♣ DANN (2016, JMLR) [30]	0.8425 $\pm$ 0.1715	14.71 $\pm$ 27.68	0.7568 $\pm$ 0.1994	0.8489 $\pm$ 0.2276	0.9098 $\pm$ 0.1042
♣ TestFit (2024, MIA) [50]	0.8686 $\pm$ 0.1428	11.52 $\pm$ 23.84	0.7856 $\pm$ 0.1686	0.8925 $\pm$ 0.1916	0.8890 $\pm$ 0.0943
♣ DeTTA (2024, WACV) [62]	0.8479 $\pm$ 0.1785	15.84 $\pm$ 42.27	0.7645 $\pm$ 0.2182	0.8489 $\pm$ 0.2347	0.9018 $\pm$ 0.1385
♣ UDA-MIMA (2024, CMPB) [52]	0.7913 $\pm$ 0.1458	21.91 $\pm$ 26.37	0.6767 $\pm$ 0.1842	0.8260 $\pm$ 0.2347	0.8426 $\pm$ 0.1579
♣ GraTa (2025, AAAI) [61]	0.8516 $\pm$ 0.1665	13.96 $\pm$ 27.20	0.7694 $\pm$ 0.1956	0.8533 $\pm$ 0.2215	0.9090 $\pm$ 0.0945
♥ TENT (2021, ICLR) [33]	0.8558 $\pm$ 0.1627	14.33 $\pm$ 35.04	0.7794 $\pm$ 0.1911	0.8565 $\pm$ 0.2141	0.9107 $\pm$ 0.1195
♥ UPL-SFDA (2023, TMI) [58]	0.8683 $\pm$ 0.1534	13.26 $\pm$ 30.34	0.7891 $\pm$ 0.1739	0.8672 $\pm$ 0.1941	0.8990 $\pm$ 0.1045
♥ DG-TTA (2023, arXiv) [31]	0.8725 $\pm$ 0.1481	13.37 $\pm$ 40.52	0.7962 $\pm$ 0.1744	0.8586 $\pm$ 0.1938	0.8947 $\pm$ 0.1311
♥ SaTTCA (2023, MICCAI) [57]	0.8426 $\pm$ 0.0968	14.71 $\pm$ 15.67	0.7570 $\pm$ 0.1175	0.8491 $\pm$ 0.1547	0.9009 $\pm$ 0.0767
♥ SmaRT (2025, arXiv) [70]	0.9108 $\pm$ 0.0688	6.19 $\pm$ 13.86	0.8427 $\pm$ 0.1048	0.9679 $\pm$ 0.0628	0.8721 $\pm$ 0.1112
♦ AdaMI (2022, MICCAI) [51]	0.8425 $\pm$ 0.1738	16.84 $\pm$ 42.26	0.7597 $\pm$ 0.2136	0.8438 $\pm$ 0.2300	0.8970 $\pm$ 0.1342
♦ PASS (2024, TMI) [74]	0.8426 $\pm$ 0.1736	15.90 $\pm$ 35.99	0.7577 $\pm$ 0.2014	0.8445 $\pm$ 0.2296	0.9013 $\pm$ 0.1188
♦ VPTTA (2024, CVPR) [63]	0.8562 $\pm$ 0.1708	16.96 $\pm$ 27.20	0.7738 $\pm$ 0.1997	0.8576 $\pm$ 0.2253	0.9132 $\pm$ 0.0982
♦ ExploringTTA (2025, ISBI) [59]	0.8519 $\pm$ 0.1687	14.81 $\pm$ 35.29	0.7704 $\pm$ 0.1971	0.8533 $\pm$ 0.2221	0.9112 $\pm$ 0.1147
♦ ProSFDA (2026, PR) [32]	0.8696 $\pm$ 0.1579	14.41 $\pm$ 41.43	0.7947 $\pm$ 0.1852	0.8593 $\pm$ 0.2055	0.9191 $\pm$ 0.1199

Nevertheless, a few methods exhibited instability. For example, VPTTA slightly underperformed the baseline in Dice, while TestFit and UPL-SFDA showed worse HD95 values. Overall, all four paradigms demonstrated adaptive capability on CXR, though Prior Estimation and feature alignment approaches were more sensitive to parameter settings and data distribution. Fig 15 illustrates the effects of various TTA methods on the CXR task and their influence on the encoder. The majority of models exhibit adverse effects.

4.3. TTA Under Strong Domain Shift

In scenarios with substantial domain shifts or severe noise interference, such as those involving US, MRI, and CT, Input-level Transformation retained its overall advantage in terms of Dice scores, whereas the best HD95 performance typically

Table 9: **Comprehensive performance comparison of TTA methods on Lung Segmentation in Chest X-ray Images.** Results colored in red, orange, and yellow denote the best, second-best, and third-best performances, respectively. Rows with a gray background indicate Dice values lower than that of the Target-Domain(w/o TTA). Symbols denote paradigms: ♠ for Input-level Translation, ♣ for Feature-level Alignment, ♥ for Output-level Regularization, ♦ for Prior Estimation. All improvements are statistically significant at $p < 0.05$.

Method	Dice $\uparrow$	HD95 $\downarrow$	JI $\uparrow$	Sen $\uparrow$	PPV $\uparrow$
Intra-Domain	0.9557$\pm$0.0057	1.02$\pm$0.30	0.9170$\pm$0.0095	0.9573$\pm$0.0077	0.9569$\pm$0.0023
Target-Domain (w/o TTA)	0.9093$\pm$0.0695	5.15$\pm$9.11	0.8414$\pm$0.1038	0.8534$\pm$0.1059	0.9846$\pm$0.0272
♠ RSA (2021, MICCAI) [29]	0.9183 $\pm$ 0.0892	4.64 $\pm$ 9.46	0.8781 $\pm$ 0.0848	0.9531 $\pm$ 0.0423	0.9245 $\pm$ 0.1048
♠ DL-TTA (2022, TMI) [71]	0.9162 $\pm$ 0.0782	4.36 $\pm$ 10.38	0.8863 $\pm$ 0.0818	0.9408 $\pm$ 0.0513	0.9131 $\pm$ 0.0865
♠ SFDA-FSM (2022, MIA) [49]	0.9663 $\pm$ 0.0223	0.56 $\pm$ 3.01	0.9357 $\pm$ 0.0390	0.9648 $\pm$ 0.0398	0.9690 $\pm$ 0.0151
♠ STDR (2024, TMI) [72]	0.9260 $\pm$ 0.0674	3.87 $\pm$ 8.13	0.8662 $\pm$ 0.0913	0.9478 $\pm$ 0.0429	0.9100 $\pm$ 0.0876
♠ AIF-SFDA (2025, AAAI) [73]	0.9274 $\pm$ 0.0205	3.39 $\pm$ 4.20	0.8755 $\pm$ 0.0324	0.9531 $\pm$ 0.0135	0.9101 $\pm$ 0.0377
♣ DANN (2016, JMLR) [30]	0.9344 $\pm$ 0.0632	3.51 $\pm$ 10.14	0.8824 $\pm$ 0.0932	0.9608 $\pm$ 0.0418	0.9167 $\pm$ 0.0940
♣ DeTTA (2024, WACV) [62]	0.9095 $\pm$ 0.0716	4.35 $\pm$ 10.37	0.8796 $\pm$ 0.0751	0.9341 $\pm$ 0.0384	0.9064 $\pm$ 0.0798
♣ TestFit (2024, MIA) [50]	0.9120 $\pm$ 0.0849	5.59 $\pm$ 11.62	0.8477 $\pm$ 0.1230	0.9208 $\pm$ 0.1117	0.9409 $\pm$ 0.1022
♣ UDA-MIMA (2024, CMPB) [52]	0.9199 $\pm$ 0.0802	5.26 $\pm$ 12.70	0.8602 $\pm$ 0.1159	0.9567 $\pm$ 0.0415	0.8964 $\pm$ 0.1214
♣ GraTa (2025, AAAI) [61]	0.9212 $\pm$ 0.0733	4.73 $\pm$ 11.46	0.8611 $\pm$ 0.1074	0.9478 $\pm$ 0.0440	0.9061 $\pm$ 0.1145
♥ TENT (2021, ICLR) [33]	0.9325 $\pm$ 0.0630	3.69 $\pm$ 10.20	0.8791 $\pm$ 0.0942	0.9607 $\pm$ 0.0400	0.9133 $\pm$ 0.0961
♥ UPL-SFDA (2023, TMI) [58]	0.9042 $\pm$ 0.0831	5.47 $\pm$ 10.45	0.8557 $\pm$ 0.0953	0.9243 $\pm$ 0.0536	0.9147 $\pm$ 0.0857
♥ DG-TTA (2023, arXiv) [31]	0.9420 $\pm$ 0.0458	2.92 $\pm$ 7.72	0.8936 $\pm$ 0.0757	0.9175 $\pm$ 0.0685	0.9720 $\pm$ 0.0434
♥ SaTTCA (2023, MICCAI) [57]	0.9348 $\pm$ 0.0272	3.40 $\pm$ 4.21	0.8829 $\pm$ 0.0397	0.9605 $\pm$ 0.0208	0.9175 $\pm$ 0.0450
♥ SmaRT (2025, arXiv) [70]	0.9501 $\pm$ 0.0288	1.77 $\pm$ 6.10	0.9064 $\pm$ 0.0496	0.9572 $\pm$ 0.0413	0.9449 $\pm$ 0.0364
♦ AdaMI (2022, MICCAI) [51]	0.9363 $\pm$ 0.0628	3.49 $\pm$ 10.09	0.8857 $\pm$ 0.0926	0.9576 $\pm$ 0.0425	0.9232 $\pm$ 0.0933
♦ PASS (2024, TMI) [74]	0.9346 $\pm$ 0.0632	3.50 $\pm$ 10.13	0.8827 $\pm$ 0.0931	0.9606 $\pm$ 0.0418	0.9173 $\pm$ 0.0939
♦ VPTTA (2024, CVPR) [63]	0.8978 $\pm$ 0.0765	4.87 $\pm$ 10.44	0.8492 $\pm$ 0.0886	0.9177 $\pm$ 0.0469	0.9081 $\pm$ 0.0791
♦ ExploringTTA (2025, ISBI) [59]	0.9286 $\pm$ 0.0711	3.89 $\pm$ 10.67	0.8734 $\pm$ 0.1014	0.9625 $\pm$ 0.0402	0.9058 $\pm$ 0.1040
♦ ProSFDA (2026, PR) [32]	0.9451 $\pm$ 0.0445	2.56 $\pm$ 7.64	0.8990 $\pm$ 0.0716	0.9402 $\pm$ 0.0563	0.9549 $\pm$ 0.0603

came from feature alignment or output regularization. In particular, in 3D MRI and CT, appearance alignment alone was insufficient to ensure cross-patch consistency, and generative model-based approaches incurred high computational costs and low inference efficiency. By contrast, Output-level Regularization strategies, such as entropy minimization, showed greater robustness to heavy noise.

4.3.1. Magnetic Resonance Imaging Modality

In the Magnetic Resonance Imaging task shown in the Table 10 and Table 11, Input-level Transformation methods achieved the most notable improvement in average Dice, with DL-TTA performing the best. The average Dice reached 0.6675, representing a 12.2% increase. Within subregions, ET and TC improved by 0.1209 and 0.1007, respectively, while WT decreased slightly by 0.0045. SmaRT

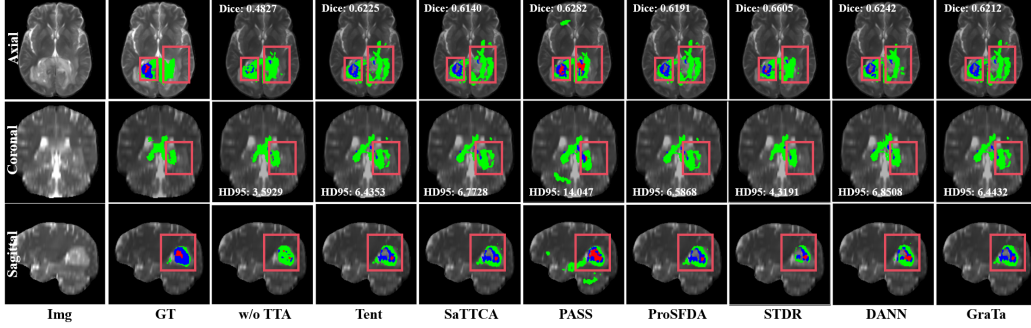


Figure 16: **Visual comparison of TTA efficacy across different methods on MRI datasets.** The three rows represent slices taken from three different orientations. GT denotes the ground-truth. The colors **red**, **green**, and **blue** represent categories 1, 2, and 3, respectively.

achieved the largest reduction in HD95, bringing it to 4.4402, an 80.7% decrease. Notably, only half of the methods outperformed the baseline in the WT subregion. This is primarily because WT relies heavily on the high FLAIR signal, which exhibits diffuse boundaries and significant inter-center variability. This reliance leads entropy-minimization and uncertainty-filtering approaches to focus on core regions while contracting peripheral edema areas, resulting in mild under-segmentation. In contrast, ET and TC regions exhibit stronger contrast and more compact morphology, yielding more stable pseudo-labels and more consistent convergence. However, it is essential to note that the baseline model used for the TTA methods is the standard UNet3D, which shows limited representation ability in more complex task scenarios, such as BraTS. As a result, its performance is subpar in the source domain. Nevertheless, when directly tested on the target domain, a clear domain shift is observed, with significant declines in metrics such as Dice and HD95. The performance improvements observed with other methods listed on the leaderboard further validate the effectiveness of the TTA approach. Fig. 16 presents the effects of various TTA methods on the MRI task. Most models demonstrate a positive impact on the Dice metric while negatively affecting the HD95 metric.

4.3.2. Computed Tomography Modality

In the Computed Tomography task reported in Table 12, all twenty methods increased the Dice score, and fifteen of them also reduced HD95, whereas four led to higher HD95. Taken together, the CT results show consistent gains across paradigms. DG-TTA achieves the highest Dice score, reaching 0.8500 and surpassing the baseline value of 0.6595. SFDA-FSM delivers the most considerable

Table 10: **Comprehensive performance comparison of TTA methods on MRI brain tumor segmentation: Dice and HD95 across ET/TC/WT.** Results colored in **red**, **orange**, and **yellow** denote the best, second-best, and third-best performances, respectively. Rows with a **gray** background indicate Dice values lower than that of the Target-Domain(w/o TTA). Symbols denote paradigms: ♠ for Input-level Translation, ♣ for Feature-level Alignment, ♥ for Output-level Regularization, ♦ for Prior Estimation. All improvements are statistically significant at $p < 0.05$.

Method	Dice↑			HD95↓		
	ET	TC	WT	ET	TC	WT
Intra-Domain	0.6320±0.1664	0.6379±0.1706	0.8480±0.0542	4.40±9.63	4.40±9.58	1.20±0.74
Target-Domain (w/o TTA)	0.5232±0.1279	0.5105±0.1217	0.7468±0.0834	26.85±2.31	37.04±3.27	15.22±8.34
♠ RSA (2021, MICCAI) [29]	0.5527±0.1681	0.5061±0.1267	0.7593±0.1580	17.65±4.31	17.67±4.14	15.80±13.03
♠ DL-TTA (2022, TMI) [71]	0.6441±0.1921	0.6160±0.1834	0.7423±0.0731	14.31±8.39	14.28±8.94	9.35±7.31
♠ SFDA-FSM (2022, MIA) [49]	0.6118±0.2369	0.5936±0.2754	0.6764±0.1287	18.39±48.30	16.82±48.07	12.27±12.87
♠ STDR (2024, TMI) [72]	0.5438±0.1749	0.5187±0.1743	0.7052±0.1289	15.17±4.93	14.88±4.46	12.29±4.86
♠ AIF-SFDA (2025, AAAI) [73]	0.6375±0.1856	0.6092±0.1768	0.7342±0.0734	10.30±8.38	9.27±4.93	7.34±7.30
♣ DANN (2016, JMLR) [30]	0.5394±0.1522	0.5006±0.0838	0.7056±0.1420	12.65±10.38	12.49±6.00	14.20±7.36
♣ DeTTA (2024, WACV) [62]	0.5975±0.1371	0.5742±0.0953	0.7489±0.0843	12.54±5.59	11.35±5.52	8.43±3.93
♣ TestFit (2024, MIA) [50]	0.5669±0.1083	0.5756±0.0837	0.7508±0.0251	16.11±3.40	6.87±3.60	2.51±1.82
♣ UDA-MIMA (2024, CMPB) [52]	0.5609±0.2135	0.5505±0.1488	0.7373±0.0763	16.65±9.41	15.56±9.99	7.63±9.93
♣ GraTa (2025, AAAI) [61]	0.5996±0.1413	0.5842±0.1242	0.7379±0.0396	6.84±5.00	7.29±4.14	3.96±3.95
♥ TENT (2021, ICLR) [33]	0.6145±0.1164	0.6010±0.0971	0.7291±0.0279	5.49±2.44	5.77±2.43	2.79±1.73
♥ UPL-SFDA (2023, TMI) [58]	0.5416±0.1842	0.5065±0.1337	0.7398±0.0957	17.30±14.83	17.15±8.30	9.58±6.79
♥ DG-TTA (2023, arXiv) [31]	0.5995±0.1409	0.5839±0.1140	0.7384±0.0394	6.95±2.00	7.38±4.14	3.94±4.00
♥ SaTTCA (2023, MICCAI) [57]	0.5929±0.1399	0.5717±0.1423	0.7395±0.0448	7.35±5.10	7.64±5.22	4.48±4.09
♥ SmaRT (2025, arXiv) [70]	0.6403±0.2392	0.6002±0.2794	0.7520±0.0279	4.80±4.52	5.36±4.98	3.16±2.79
♦ AdaMI (2022, MICCAI) [51]	0.5906±0.1481	0.5631±0.1057	0.7481±0.0367	6.66±5.14	7.64±5.22	3.62±4.53
♦ PASS (2024, TMI) [74]	0.5983±0.1083	0.5736±0.0977	0.7365±0.0347	6.64±5.60	7.46±5.41	3.47±4.04
♦ VPTTA (2024, CVPR) [63]	0.5962±0.1365	0.5724±0.0946	0.7475±0.0834	13.54±5.59	10.35±5.19	8.43±3.93
♦ ExploringTTA (2025, ISBI) [59]	0.6038±0.1433	0.5805±0.1473	0.7443±0.0396	6.55±5.60	7.36±5.15	3.44±3.94
♦ ProSFDA (2026, PR) [32]	0.5883±0.1801	0.5796±0.2058	0.7412±0.0583	6.78±5.13	7.05±5.53	5.35±5.83

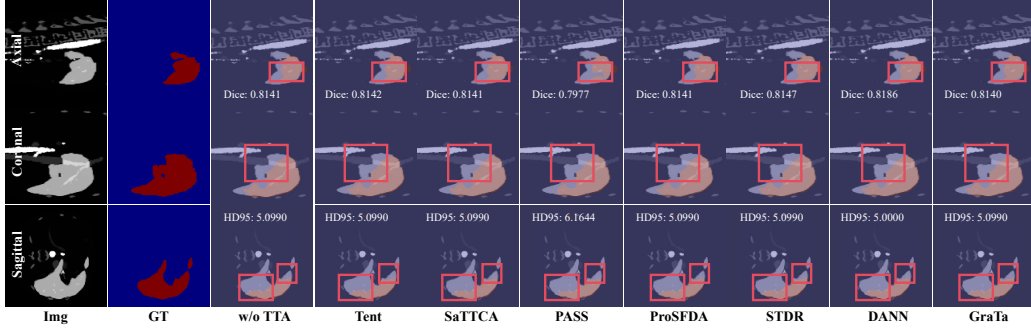


Figure 17: **Visual comparison of TTA efficacy across different methods on CT datasets.** The three rows represent slices taken from three different orientations. GT denotes the ground-truth.

reduction in HD95, lowering the value by 6.8811, which corresponds to an im-

Table 11: **Comprehensive performance comparison of TTA methods on MRI brain tumor segmentation: Sen and PPV across ET/TC/WT.** Results colored in red, orange, and yellow denote the best, second-best, and third-best performances, respectively. Rows with a gray background indicate Dice values lower than that of the Target-Domain(w/o TTA). Symbols denote paradigms: ♠ for Input-level Translation, ♣ for Feature-level Alignment, ♥ for Output-level Regularization, ♦ for Prior Estimation. All improvements are statistically significant at $p < 0.05$.

Method	Sen↑			PPV↑		
	ET	TC	WT	ET	TC	WT
Intra-Domain	0.6862±0.0181	0.6978±0.0209	0.7094±0.0083	0.7912±0.1511	0.7758±0.1677	0.8518±0.0559
Target-Domain (w/o TTA)	0.6328±0.0398	0.6302±0.0435	0.6771±0.0387	0.7472±0.0826	0.7802±0.0722	0.7881±0.0892
♠ RSA (2021, MICCAI) [29]	0.6455±0.0231	0.6559±0.1071	0.6828±0.0195	0.6162±0.1242	0.7021±0.0143	0.0993±0.0993
♠ DL-TTA (2022, TMI) [71]	0.6593±0.1631	0.6475±0.0518	0.6868±0.0488	0.7123±0.1351	0.7561±0.1362	0.7282±0.1551
♠ SFDA-FSM (2022, MIA) [49]	0.6671±0.0523	0.6553±0.0672	0.6971±0.0552	0.6312±0.2734	0.6764±0.0287	0.6518±0.2349
♠ STDR (2024, TMI) [72]	0.6312±0.0371	0.6355±0.0427	0.6606±0.0860	0.5784±0.1749	0.8053±0.0427	0.8401±0.1293
♠ AIF-SFDA (2025, AAAI) [73]	0.6527±0.0566	0.6409±0.1297	0.6802±0.1425	0.7052±0.1268	0.7358±0.1737	0.7216±0.1486
♣ DANN (2016, JMLR) [30]	0.6421±0.0531	0.6592±0.0429	0.6828±0.0485	0.5332±0.0838	0.6585±0.0706	0.0891±0.0891
♣ DeTTA (2024, WACV) [62]	0.6236±0.0298	0.6032±0.0349	0.6712±0.0324	0.6026±0.0953	0.7957±0.0928	0.7941±0.1444
♣ TestFit (2024, MIA) [50]	0.6312±0.0759	0.5989±0.0215	0.6783±0.0204	0.7590±0.0835	0.8323±0.0829	0.8609±0.0923
♣ UDA-MIMA (2024, CMPB) [52]	0.6421±0.0514	0.6544±0.0649	0.6871±0.0511	0.7005±0.0514	0.7402±0.1244	0.7129±0.0740
♣ GraTa (2025, AAAI) [61]	0.6378±0.0366	0.6618±0.0433	0.7466±0.0397	0.6842±0.1242	0.7593±0.1128	0.7466±0.1692
♥ TENT (2021, ICLR) [33]	0.6386±0.0731	0.6418±0.0724	0.6838±0.0718	0.7055±0.0973	0.7889±0.0633	0.7725±0.1221
♥ UPL-SFDA (2023, TMI) [58]	0.6612±0.0487	0.6691±0.0499	0.6828±0.1172	0.6495±0.0831	0.6691±0.1557	0.7417±0.0489
♥ DG-TTA (2023, arXiv) [31]	0.6375±0.0365	0.6613±0.0431	0.7463±0.0393	0.6698±0.1140	0.7613±0.1123	0.7463±0.1642
♥ SaTTCA (2023, MICCAI) [57]	0.6397±0.0366	0.6212±0.0367	0.6828±0.0391	0.6656±0.0579	0.7395±0.1369	0.7154±0.1742
♥ SmaRT (2025, arXiv) [70]	0.6374±0.2147	0.6256±0.0704	0.6719±0.0450	0.6494±0.2794	0.6929±0.2245	0.7019±0.0941
♦ AdaMI (2022, MICCAI) [51]	0.6249±0.0375	0.6033±0.0415	0.8301±0.0842	0.7474±0.1068	0.8108±0.1124	0.8301±0.1553
♦ PASS (2024, TMI) [74]	0.6320±0.0835	0.6107±0.0813	0.6766±0.0204	0.7187±0.0970	0.7058±0.0955	0.7991±0.0922
♦ VPTTA (2024, CVPR) [63]	0.6258±0.0291	0.6019±0.0342	0.6697±0.0318	0.6246±0.0946	0.7694±0.0921	0.7928±0.1437
♦ ExploringTTA (2025, ISBI) [59]	0.7322±0.1016	0.6082±0.0412	0.8008±0.0837	0.7322±0.1472	0.7554±0.1627	0.8008±0.1507
♦ ProSFDA (2026, PR) [32]	0.6318±0.0423	0.6197±0.0621	0.6846±0.0240	0.7074±0.2076	0.6971±0.0553	0.7675±0.1945

provement of approximately 50%. These trends align with known properties of CT reconstruction. Variability in reconstruction kernel and slice thickness induces intensity drift and artifacts, which in turn make HD95 highly sensitive to boundary outliers. Methods that adapt without explicit shape constraints, for example GraTa and TENT, are therefore more prone to boundary shifts or over-suppression during adaptation, which elevates HD95. Fig 17 illustrates the performance of different TTA methods on the CT task, with most models achieving comparable results.

4.3.3. Ultrasound Modality

In the Ultrasound task shown in Table 13, most methods achieved improvements. SFDA-FSM achieved the most significant Dice gain, with a score of 0.8333 compared to the baseline, representing an increase of 0.3522, or 73 percent. However, its HD95 worsened to 10.6708, an 83 percent increase. Methods such as RSA also exhibited marked HD95 degradation, closely related to intrinsic ultrasound

Table 12: **Comprehensive performance comparison of TTA methods on Liver Segmentation in Computed Tomography.** Results colored in red, orange, and yellow denote the best, second-best, and third-best performances, respectively. Rows with a gray background indicate Dice values lower than that of the Target-Domain(w/o TTA). Symbols denote paradigms: ♠ for Input-level Translation, ♣ for Feature-level Alignment, ♥ for Output-level Regularization, ♦ for Prior Estimation. All improvements are statistically significant at $p < 0.05$.

Method	Dice ↑	HD95 ↓	JI ↑	Sen ↑	PPV ↑
Intra-Domain	0.8949$\pm$0.0250	1.63$\pm$1.68	0.8116$\pm$0.0388	0.9569$\pm$0.0172	0.8433$\pm$0.0382
Target-Domain (w/o TTA)	0.6595$\pm$0.2833	13.91$\pm$11.30	0.5489$\pm$0.2723	0.6376$\pm$0.0693	0.8814$\pm$0.2263
♠ RSA (2021, MICCAI) [29]	0.6973 $\pm$ 0.2583	12.39 $\pm$ 10.28	0.5857 $\pm$ 0.2579	0.6461 $\pm$ 0.0655	0.8743 $\pm$ 0.2211
♠ DL-TTA (2022, TMI) [71]	0.6752 $\pm$ 0.2183	11.58 $\pm$ 9.51	0.6031 $\pm$ 0.1947	0.6673 $\pm$ 0.0521	0.9002 $\pm$ 0.1248
♠ SFDA-FSM (2022, MIA) [49]	0.7780 $\pm$ 0.1147	6.88 $\pm$ 7.17	0.6845 $\pm$ 0.1066	0.6764 $\pm$ 0.0271	0.8711 $\pm$ 0.0894
♠ STDR (2024, TMI) [72]	0.7255 $\pm$ 0.2403	11.34 $\pm$ 9.82	0.6147 $\pm$ 0.2458	0.6543 $\pm$ 0.0626	0.8771 $\pm$ 0.2176
♠ AIF-SFDA (2025, AAAI) [73]	0.7018 $\pm$ 0.2625	9.39 $\pm$ 10.28	0.5900 $\pm$ 0.2620	0.6503 $\pm$ 0.0696	0.8786 $\pm$ 0.2252
♣ DANN (2016, JMLR) [30]	0.7082 $\pm$ 0.2205	10.99 $\pm$ 8.83	0.5866 $\pm$ 0.2285	0.6499 $\pm$ 0.0593	0.8415 $\pm$ 0.1911
♣ DeTTA (2024, WACV) [62]	0.6838 $\pm$ 0.1139	11.51 $\pm$ 9.81	0.5749 $\pm$ 0.0958	0.6570 $\pm$ 0.0257	0.8182 $\pm$ 0.1307
♣ TestFit (2024, MIA) [50]	0.6614 $\pm$ 0.2155	9.95 $\pm$ 8.27	0.5816 $\pm$ 0.1835	0.6440 $\pm$ 0.0492	0.9005 $\pm$ 0.1086
♣ UDA-MIMA (2024, CMPB) [52]	0.6919 $\pm$ 0.2354	12.21 $\pm$ 10.97	0.5756 $\pm$ 0.2036	0.6423 $\pm$ 0.0542	0.8842 $\pm$ 0.2352
♣ GraTa (2025, AAAI) [61]	0.6916 $\pm$ 0.1485	23.12 $\pm$ 15.27	0.5472 $\pm$ 0.1652	0.6519 $\pm$ 0.0334	0.7493 $\pm$ 0.1877
♥ TENT (2021, ICLR) [33]	0.6918 $\pm$ 0.1484	23.10 $\pm$ 15.25	0.5474 $\pm$ 0.1652	0.6519 $\pm$ 0.0334	0.7495 $\pm$ 0.1876
♥ UPL-SFDA (2023, TMI) [58]	0.6748 $\pm$ 0.1145	12.73 $\pm$ 10.26	0.5657 $\pm$ 0.1964	0.6471 $\pm$ 0.0526	0.8023 $\pm$ 0.1632
♥ DG-TTA (2023, arXiv) [31]	0.8500 $\pm$ 0.1037	8.61 $\pm$ 11.89	0.7518 $\pm$ 0.1408	0.7112 $\pm$ 0.0210	0.8216 $\pm$ 0.1615
♥ SaTTCA (2023, MICCAI) [57]	0.6793 $\pm$ 0.1098	12.51 $\pm$ 9.81	0.5706 $\pm$ 0.0919	0.6528 $\pm$ 0.0218	0.8139 $\pm$ 0.1268
♥ SmaRT (2025, arXiv) [70]	0.8405 $\pm$ 0.1231	8.91 $\pm$ 12.26	0.7422 $\pm$ 0.1634	0.7261 $\pm$ 0.0061	0.7534 $\pm$ 0.1649
♦ AdaMI (2022, MICCAI) [51]	0.7094 $\pm$ 0.2200	10.97 $\pm$ 8.82	0.5879 $\pm$ 0.2282	0.6503 $\pm$ 0.0592	0.8407 $\pm$ 0.1834
♦ PASS (2024, TMI) [74]	0.7088 $\pm$ 0.2204	11.52 $\pm$ 8.80	0.5873 $\pm$ 0.2284	0.6521 $\pm$ 0.0397	0.8420 $\pm$ 0.1913
♦ VPTTA (2024, CVPR) [63]	0.6792 $\pm$ 0.1187	14.73 $\pm$ 10.26	0.5700 $\pm$ 0.2005	0.6514 $\pm$ 0.0567	0.8066 $\pm$ 0.1674
♦ ExploringTTA (2025, ISBI) [59]	0.6933 $\pm$ 0.1472	20.86 $\pm$ 14.80	0.5489 $\pm$ 0.1642	0.6519 $\pm$ 0.0335	0.7525 $\pm$ 0.1841
♦ ProSFDA (2026, PR) [32]	0.6937 $\pm$ 0.2003	11.95 $\pm$ 8.54	0.5624 $\pm$ 0.2080	0.6458 $\pm$ 0.0542	0.8135 $\pm$ 0.1738

characteristics, including speckle noise, low inter-tissue contrast, acoustic shadowing, and occlusions, which cause boundary ambiguity and discontinuity. When Input-level Transformation applies strong style remapping during inference, it can induce local edge truncation or expansion, amplifying HD95 sensitivity. Fig. 18 illustrates the effects of various TTA methods on the US task and their influence on the encoder, where most models exhibit beneficial effects.

In summary, experimental results indicate that input-level transformation methods offer plug-and-play adaptability and strong cross-modality and cross-center generalization, whereas feature-level and output-level strategies primarily enhance boundary consistency and 3D adaptation. Each paradigm exhibits distinct advantages, suggesting that future work should focus on their synergistic integration to achieve more robust and widely generalizable TTA.

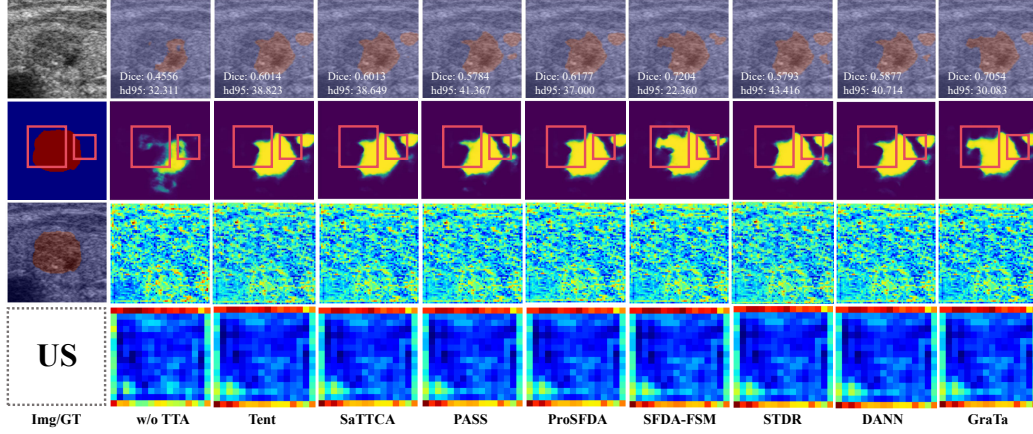


Figure 18: **Visual comparison of TTA efficacy across different methods on US datasets.** Rows 1 and 2 illustrate the segmentation outcomes. Rows 3 and 4 depict changes in features extracted by the Encoder after TTA. GT denotes the ground-truth.

5. Challenges and Prospects

5.1. Stable Adaptation and Streaming Forgetting Prevention

Clinical inference often proceeds in small batches or even single instances, where noise in early predictions can be easily amplified by entropy-driven online updates, leading to parameter drift and catastrophic forgetting. This issue is particularly pronounced in cases with gradually shifting distributions in sequential clinical settings [33]. Therefore, real-world deployment should not only emphasize immediate gains but also address three critical questions. The update target refers to which parameters are permitted to be adjusted. The update magnitude concerns how the intensity of adaptation is controlled. The rollback mechanism determines when and how to trigger reversion. To address these challenges, future studies should explore more robust sample selection strategies and anti-forgetting mechanisms to achieve stable adaptation in streaming scenarios [34].

5.2. Style-content Disentanglement

During the TTA stage, features extracted by pretrained models often encode both style and content simultaneously. This makes it difficult to disentangle which dimensions correspond to appearance style and which to anatomical semantics, thereby limiting the effectiveness of style alignment and content modeling. Existing studies suggest that shallow features tend to capture texture and style, whereas deeper features encode shape and content [40, 39]. However, the validity of this

Table 13: **Comprehensive performance comparison of TTA methods on Ultrasound Thyroid Nodule Segmentation.** Results colored in **red**, **orange**, and **yellow** denote the best, second-best, and third-best performances, respectively. Rows with a **gray** background indicate Dice values lower than that of the Target-Domain(w/o TTA). Symbols denote paradigms: ♠ for Input-level Translation, ♣ for Feature-level Alignment, ♥ for Output-level Regularization, ♦ for Prior Estimation. All improvements are statistically significant at $p < 0.05$.

Method	Dice ↑	HD95 ↓	JI ↑	Sen ↑	PPV ↑
Intra-Domain	0.7215$\pm$0.0266	31.01$\pm$8.30	0.6066$\pm$0.0313	0.7812$\pm$0.0351	0.7541$\pm$0.0328
Target-Domain (w/o TTA)	0.4811$\pm$0.2708	62.78$\pm$68.88	0.3591$\pm$0.2393	0.4288$\pm$0.2958	0.7531$\pm$0.2801
♠ RSA (2021, MICCAI) [29]	0.4879 $\pm$ 0.3011	71.24 $\pm$ 92.52	0.3752 $\pm$ 0.2677	0.4322 $\pm$ 0.3201	0.7740 $\pm$ 0.3008
♠ DL-TTA (2022, TMI) [71]	0.4930 $\pm$ 0.3193	60.47 $\pm$ 119.30	0.3895 $\pm$ 0.2779	0.4631 $\pm$ 0.3365	0.7752 $\pm$ 0.3643
♠ SFDA-FSM (2022, MIA) [49]	0.8333 $\pm$ 0.1139	10.67 $\pm$ 9.25	0.7275 $\pm$ 0.1385	0.7731 $\pm$ 0.1462	0.9303 $\pm$ 0.0853
♠ STDR (2024, TMI) [72]	0.4949 $\pm$ 0.2537	61.55 $\pm$ 58.31	0.3807 $\pm$ 0.2146	0.5233 $\pm$ 0.1997	0.7612 $\pm$ 0.2362
♠ AIF-SFDA (2025, AAAI) [73]	0.5692 $\pm$ 0.2525	42.39 $\pm$ 59.86	0.4368 $\pm$ 0.2331	0.5373 $\pm$ 0.2935	0.7671 $\pm$ 0.2458
♣ DANN (2016, JMLR) [30]	0.5646 $\pm$ 0.2482	50.38 $\pm$ 59.86	0.4324 $\pm$ 0.2289	0.5330 $\pm$ 0.2893	0.7628 $\pm$ 0.2416
♣ DeTTA (2024, WACV) [62]	0.5049 $\pm$ 0.2735	59.50 $\pm$ 65.51	0.3807 $\pm$ 0.2453	0.4577 $\pm$ 0.3037	0.7637 $\pm$ 0.2748
♣ TestFit (2024, MIA) [50]	0.5004 $\pm$ 0.2693	58.50 $\pm$ 65.51	0.3764 $\pm$ 0.2411	0.4534 $\pm$ 0.2995	0.7594 $\pm$ 0.2706
♣ UDA-MIMA (2024, CMPB) [52]	0.6359 $\pm$ 0.1406	35.15 $\pm$ 17.99	0.4809 $\pm$ 0.1445	0.5715 $\pm$ 0.1672	0.7811 $\pm$ 0.2014
♣ GraTa (2025, AAAI)	0.5636 $\pm$ 0.2554	50.62 $\pm$ 62.30	0.4337 $\pm$ 0.2361	0.5336 $\pm$ 0.2998	0.7675 $\pm$ 0.2463
♥ TENT (2021, ICLR) [33]	0.5546 $\pm$ 0.2550	51.80 $\pm$ 62.28	0.4244 $\pm$ 0.2342	0.5254 $\pm$ 0.2992	0.7613 $\pm$ 0.2474
♥ UPL-SFDA (2023, TMI) [58]	0.4950 $\pm$ 0.2972	67.90 $\pm$ 88.33	0.3806 $\pm$ 0.2652	0.4933 $\pm$ 0.3183	0.7784 $\pm$ 0.2929
♥ DG-TTA (2023, arXiv) [31]	0.4805 $\pm$ 0.2712	61.96 $\pm$ 69.93	0.3583 $\pm$ 0.2393	0.4270 $\pm$ 0.2953	0.7560 $\pm$ 0.2818
♥ SaTTCA (2023, MICCAI) [57]	0.5640 $\pm$ 0.0711	50.45 $\pm$ 18.69	0.4319 $\pm$ 0.0644	0.5326 $\pm$ 0.0596	0.7630 $\pm$ 0.0762
♥ SmaRT (2025, arXiv) [70]	0.7957 $\pm$ 0.1038	16.39 $\pm$ 13.10	0.6722 $\pm$ 0.1347	0.6796 $\pm$ 0.0957	0.7583 $\pm$ 0.1722
♦ AdaMI (2022, MICCAI) [51]	0.5573 $\pm$ 0.2505	53.01 $\pm$ 65.82	0.4256 $\pm$ 0.2297	0.5240 $\pm$ 0.2903	0.7554 $\pm$ 0.2507
♦ PASS (2024, TMI) [74]	0.6533 $\pm$ 0.1775	31.70 $\pm$ 23.92	0.5092 $\pm$ 0.1848	0.6590 $\pm$ 0.2357	0.7563 $\pm$ 0.2227
♦ VPTTA (2024, CVPR) [63]	0.5681 $\pm$ 0.2596	46.62 $\pm$ 62.30	0.4380 $\pm$ 0.2402	0.5379 $\pm$ 0.3039	0.7718 $\pm$ 0.2504
♦ ExploringTTA (2025, ISBI) [59]	0.5640 $\pm$ 0.0711	50.45 $\pm$ 18.68	0.4319 $\pm$ 0.0644	0.5326 $\pm$ 0.0596	0.7630 $\pm$ 0.0762
♦ ProSFDA (2026, PR) [32]	0.5451 $\pm$ 0.2618	52.98 $\pm$ 63.91	0.4173 $\pm$ 0.2404	0.5065 $\pm$ 0.3016	0.7717 $\pm$ 0.2497

hierarchical assumption in medical segmentation encoders remains insufficiently verified. Future research should develop interpretable disentanglement frameworks that can quantitatively assess the extent of style alteration and the degree of content preservation, thereby enabling the separate modeling and independent optimization of style and content.

5.3. Test-Time Hyperparameter Adaptation

In the absence of annotations, adaptation performance is highly sensitive to hyperparameters such as the learning rate, update steps, and the number of updatable layers; however, the lack of a validation set renders grid search infeasible. Existing approaches typically rely on entropy or consistency as surrogate signals for hyperparameter adjustment [33, 41]. However, these signals fluctuate substantially under domain shifts and noise perturbations, leading to unstable performance across different centers. Future efforts should advance in two di-

rections: first, hyperparameter robustness, namely designing objective functions and optimization schemes that remain stable across broad hyperparameter ranges. Second, hyperparameter self-adaptation, namely leveraging surrogate signals at the case level to automatically determine whether to update, the magnitude of update, and the set of updatable layers, thereby reducing manual tuning costs and alleviating cross-domain variability.

5.4. Representation Alignment and Class Prototypes

In unlabeled target domains, pseudo-label noise and class imbalance distort the alignment direction, making it particularly prone to boundary drift across institutions and scanners. The introduction of class prototypes provides a stable anchor to mitigate this problem. In 3D organ segmentation, prototype constraints have been employed to alleviate spatial deformations during cross-institutional transfer, demonstrating robustness [46]. In scenarios that more closely approximate domain shifts, prototypes can further map target-domain features into a source-style decodable space, thereby reducing decoders’ sensitivity to distributional discrepancies [47]. Key challenges for future research include developing structure-aware prototype estimation under unlabeled streaming inputs, quantifying prototype uncertainty, and maintaining stable pixel-to-prototype associations in cases involving small objects or ambiguous boundaries.

5.5. Prior Estimation and Weak Prompts

When priors do not match the target domain, they can cause excessive smoothing or incorrect segmentation, while organ deformation and lesion heterogeneity further undermine their reliability across cases [74]. In practical TTA scenarios, the fundamental problem is how to extract case-level priors on intensity distribution, scale, and coarse shape with associated confidence scores, and how to dynamically attenuate their influence when they conflict with observations while retaining traceable records, thereby enabling systematic evaluation of heterogeneous effects across tissues and modalities [14].

5.6. Semantic Anchors and Retrieval-Enhanced Adaptive Closed Loop

With the expansion of unlabeled testing scenarios, large-scale medical-domain VLMs such as BioViL and BiomedCLIP, owing to their stronger generalization across institutions and scanners, can provide robust semantic representations for segmentation TTA [44, 45]. Meanwhile, general-purpose VLMs such as CLIP and BLIP-2 possess programmable prompting interfaces that, through appropriately designed prompts, can simultaneously capture both specific and standard features

of the target domain, thereby facilitating source-target semantic alignment [43]. A promising direction is to utilize VLMs as intelligent agents that dynamically retrieve and aggregate priors, such as shape, scale, and contextual information, for each incoming case. These priors are then injected into pretrained segmentation models in a controlled manner, creating a self-adaptive closed loop that delivers stronger performance at test time.

6. Use of Large Language Models and Research Integrity

This manuscript used ChatGPT GPT-5 Thinking solely for language polishing and clarity improvement. The model was not used to generate or edit scientific content, analysis code, results, references, or citations. All references and citations were compiled by the authors from primary sources and verified line by line. Numerical values, method descriptions, and conclusions were checked against the original data and code to ensure accuracy and traceability.

7. Conclusion

MedSeg-TTA provides a unified, comprehensive assessment of TTA methods for medical image segmentation across seven medical imaging modalities and multiple clinically relevant tasks. By rigorously controlling preprocessing procedures, backbone settings, and evaluation protocols, the benchmark reveals the heterogeneous behavior of four major adaptation paradigms and exposes their strengths, limitations, and stability boundaries under both mild and severe domain shifts. The findings demonstrate that adaptation performance is strongly modality-dependent and that several widely used methods may yield unreliable outcomes when confronted with significant distribution gaps. These insights underscore the importance of systematic evaluation and careful methodological consideration when implementing adaptive models in clinical settings. By releasing standardized datasets, reproducible implementations, and an open leaderboard, MedSeg-TTA establishes a reference point for future research, encouraging the development of more robust, interpretable, and clinically reliable adaptation strategies.

CRedit authorship contribution statement

Wenjing Yu: Methodology, Visualization, Writing - original draft. **Shuo Jiang:** Methodology, Validation, Writing - original draft. **Yifei Chen:** Methodology, Conceptualization, Writing - original draft. **Shuo Chang:** Validation, Visualization. **Yuanhan Wang:** Validation, Visualization. **Beining Wu:** Visualization.

Jie Dong: Visualization. **Mingxuan Liu:** Validation, Visualization. **Shenghao Zhu:** Validation, Visualization. **Feiwei Qin:** Supervision, Writing - review & editing. **Changmiao Wang:** Writing - review & editing. **Qiyuan Tian:** Project administration, Writing - review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 82302166), Tsinghua University Startup Fund, Fundamental Research Funds for the Provincial Universities of Zhejiang (No. GK259909299001-006), Anhui Provincial Joint Construction Key Laboratory of Intelligent Education Equipment and Technology (No. IEET202401), and the State Key Lab of CAD&CG, Zhejiang University (A2510).

References

- [1] X. You, J. He, J. Yang, and Y. Gu, “Learning with explicit shape priors for medical image segmentation,” *IEEE Transactions on Medical Imaging*, vol. 44, no. 2, pp. 927–940, 2025.
- [2] Y. Meng, Y. Zhang, J. Xie, J. Duan, M. Joddrell, S. Madhusudhan, T. Peto, Y. Zhao, and Y. Zheng, “Multi-granularity learning of explicit geometric constraint and contrast for label-efficient medical image segmentation and differentiable clinical function assessment,” *Medical Image Analysis*, vol. 95, p. 103183, 2024.
- [3] Z. Wang, Y. Luo, L. Zheng, Z. Chen, S. Wang, and Z. Huang, “In search of lost online test-time adaptation: A survey,” *International Journal of Computer Vision*, vol. 133, no. 3, pp. 1106–1139, 2025.
- [4] C. Yi, H. Chen, Y. Zhang, Y. Xu, Y. Zhou, and L. Cui, “From question to exploration: Can classic test-time adaptation strategies be effectively applied in semantic segmentation?” in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 10 085–10 094.

- [5] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, “On calibration of modern neural networks,” in *International Conference on Machine Learning*, 2017, pp. 1321–1330.
- [6] S. Niu, J. Wu, Y. Zhang, Y. Chen, S. Zheng, P. Zhao, and M. Tan, “Efficient test-time model adaptation without forgetting,” in *International Conference on Machine Learning*, 2022, pp. 16 888–16 905.
- [7] V. Sandfort, K. Yan, P. J. Pickhardt, and R. M. Summers, “Data augmentation using generative adversarial networks (CycleGAN) to improve generalizability in CT segmentation tasks,” *Scientific Reports*, vol. 9, no. 1, p. 16884, 2019.
- [8] Z. Li, W. Han, Y. Zhang, Q. Fu, J. Li, L. Qin, R. Dong, H. Sun, Y. Deng, and L. Yang, “Learning spatiotemporal dynamics with a pretrained generative model,” *Nature Machine Intelligence*, vol. 6, no. 12, pp. 1566–1579, 2024.
- [9] Z. Liu, X. Yang, R. Gao, S. Liu, H. Dou, S. He, Y. Huang, Y. Huang, H. Luo, Y. Zhang *et al.*, “Remove appearance shift for ultrasound image segmentation via fast and universal style transfer,” in *IEEE 17th International Symposium on Biomedical Imaging*, 2020, pp. 1824–1828.
- [10] B. Zheng, R. Zhang, S. Diao, J. Zhu, Y. Yuan, J. Cai, L. Shao, S. Li, and W. Qin, “Dual domain distribution disruption with semantics preservation: Unsupervised domain adaptation for medical image segmentation,” *Medical Image Analysis*, vol. 97, p. 103275, 2024.
- [11] M. Zenk, D. Zimmerer, F. Isensee, J. Traub, T. Norajitra, P. F. Jäger, and K. Maier-Hein, “Comparative benchmarking of failure detection methods in medical image segmentation: unveiling the role of confidence aggregation,” *Medical Image Analysis*, vol. 101, p. 103392, 2025.
- [12] W. Liu, Z. Ni, X. Zhu, Q. Chen, L. Ni, and P. Xia, “Spectrum intervention based invariant causal representation learning for single-domain generalizable medical image segmentation,” *Medical Image Analysis*, p. 103741, 2025.
- [13] H. Li, M. Ou, H. Li, Z. Qiu, K. Niu, H. Fu, and J. Liu, “Multi-view test-time adaptation for semantic segmentation in clinical cataract surgery,” *IEEE Transactions on Medical Imaging*, vol. 44, no. 5, pp. 2307–2318, 2025.

- [14] J. Zhu, B. Bolsterlee, Y. Song, and E. Meijering, “Improving cross-domain generalizability of medical image segmentation using uncertainty and shape-aware continual test-time domain adaptation,” *Medical Image Analysis*, vol. 101, p. 103422, 2025.
- [15] M. C. de Verdier, R. Saluja, L. Gagnon, D. LaBella, U. Baid, N. H. Tahon, M. Foltyn-Dumitru, J. Zhang, M. Alafif, S. Baig, M. others, S. Baig, M. others, J. Zhang, M. Alafif, S. Baig *et al.*, “The 2024 brain tumor segmentation (BraTS) challenge: glioma segmentation on post-treatment MRI,” *arXiv:2405.18368*, 2024.
- [16] M. Adewole, J. D. Rudie, A. Gbdamosi, O. Toyobo, C. Raymond, D. Zhang, O. Omidiji, R. Akinola, M. A. Suwaid, A. Emegoakor *et al.*, “The brain tumor segmentation (BraTS) challenge 2023: Glioma segmentation in sub-saharan africa patient population (BraTS-Africa),” *arXiv: 2305.17033 v7*, 2024.
- [17] P. Bilic, P. Christ, H. B. Li, E. Vorontsov, A. Ben-Cohen, G. Kaissis, A. Sze-skin, C. Jacobs, G. E. H. Mamani, G. Chartrand *et al.*, “The liver tumor segmentation benchmark (LiTS),” *Medical Image Analysis*, vol. 84, p. 102680, 2023.
- [18] L. Soler, A. Hostettler, V. Agnus, A. Charnoz, J.-B. Fasquel, J. Moreau, A.-B. Osswald, M. Bouhadjar, and J. Marescaux, “3D image reconstruction for comparison of algorithm database,” *URL: <https://www.ircad.fr/research/data-sets/liver-segmentation-3d-ircadb-01>*, vol. 13, 2010.
- [19] M. G. K. K. W. R. A. C. Nicholas Kurtansky, Veronica Rotemberg, “ISIC 2024 - skin cancer detection with 3D-TBP,” 2024. [Online]. Available: <https://kaggle.com/competitions/isic-2024-challenge>
- [20] T. Mendonça, M. Celebi, T. Mendonca, and J. Marques, “PH2: A public database for the analysis of dermoscopic images,” *Dermoscopy Image Analysis*, pp. 419–439, 2015.
- [21] S. Hu, Z. Liao, and Y. Xia, “Domain specific convolution and high frequency reconstruction based unsupervised domain adaptation for medical image segmentation,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2022.

- [22] S. Graham, H. Chen, J. Gamper, Q. Dou, P.-A. Heng, D. Snead, Y. W. Tsang, and N. Rajpoot, “MILD-Net: Minimal information loss dilated network for gland instance segmentation in colon histology images,” *Medical Image Analysis*, vol. 52, pp. 199–211, 2019.
- [23] K. Sirinukunwattana, J. P. Pluim, H. Chen, X. Qi, P.-A. Heng, Y. B. Guo, L. Y. Wang, B. J. Matuszewski, E. Bruni, U. Sanchez, A. Böhm, O. Ronneberger, B. B. Cheikh, D. Racoceanu, P. Kainz, M. Pfeiffer, M. Urschler, D. R. Snead, and N. M. Rajpoot, “Gland segmentation in colon histology images: The glas challenge contest,” *Medical Image Analysis*, vol. 35, pp. 489–502, 2017.
- [24] S. Jaeger, S. Candemir, S. Antani, Y.-X. J. Wang, P.-X. Lu, and G. Thoma, “Two public chest X-ray datasets for computer-aided screening of pulmonary diseases,” *Quantitative Imaging in Medicine and Surgery*, vol. 4, no. 6, p. 475, 2014.
- [25] S. Stirenko, Y. Kochura, O. Alienin, O. Rokovyi, Y. Gordienko, P. Gang, and W. Zeng, “Chest X-ray analysis of tuberculosis by deep learning with segmentation and augmentation,” in *IEEE 38th International Conference on Electronics and Nanotechnology*, 2018, pp. 422–428.
- [26] H. Gong, G. Chen, R. Wang, X. Xie, M. Mao, Y. Yu, F. Chen, and G. Li, “Multi-task learning for thyroid nodule segmentation with thyroid region prior,” in *IEEE 18th International Symposium on Biomedical Imaging*, 2021, pp. 257–261.
- [27] L. Pedraza, C. Vargas, F. Narváez, O. Durán, E. Muñoz, and E. Romero, “An open access thyroid ultrasound image database,” in *10th International Symposium on Medical Information Processing and Analysis*, vol. 9287, 2015, pp. 188–193.
- [28] J. Liang, R. He, and T. Tan, “A comprehensive survey on test-time adaptation under distribution shifts,” *International Journal of Computer Vision*, vol. 133, no. 1, pp. 31–64, 2025.
- [29] H. Zeng, K. Zou, Z. Chen, R. Zheng, and H. Fu, “Reliable source approximation: Source-free unsupervised domain adaptation for vestibular schwannoma MRI segmentation,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2024, pp. 622–632.

- [30] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. March, and V. Lempitsky, “Domain-adversarial training of neural networks,” *Journal of Machine Learning Research*, vol. 17, no. 59, pp. 1–35, 2016.
- [31] C. Weihsbach, C. N. Kruse, A. Bigalke, and M. P. Heinrich, “DG-TTA: Out-of-domain medical image segmentation through domain generalization and test-time adaptation,” *arXiv preprint arXiv:2312.06275*, 2023.
- [32] S. Hu, Z. Liao, and Y. Xia, “Source-free domain adaptation using prompt learning for medical image segmentation,” *Pattern Recognition*, p. 112290, 2025.
- [33] D. Wang, E. Shelhamer, S. Liu, B. Olshausen, and T. Darrell, “Tent: Fully test-time adaptation by entropy minimization,” in *International Conference on Learning Representations*, 2021.
- [34] S. Niu, J. Wu, Y. Zhang, Y. Chen, S. Zheng, P. Zhao, and M. Tan, “Efficient test-time model adaptation without forgetting,” in *Proceedings of the 39th International Conference on Machine Learning*, 2022, pp. 16 888–16 905.
- [35] Q. Wang *et al.*, “Continual test-time domain adaptation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 7201–7211.
- [36] A. Tarvainen and H. Valpola, “Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results,” in *Advances in Neural Information Processing Systems*, vol. 30, 2017.
- [37] Y. Yang and S. Soatto, “FDA: Fourier domain adaptation for semantic segmentation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4085–4095.
- [38] X. Huang and S. Belongie, “Arbitrary style transfer in real-time with adaptive instance normalization,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1501–1510.
- [39] L. A. Gatys, A. S. Ecker, and M. Bethge, “Image style transfer using convolutional neural networks,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2414–2423.

- [40] R. Geirhos *et al.*, “Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness,” in *International Conference on Learning Representations*, 2019.
- [41] Y. Sun, X. Wang, Z. Liu, J. Miller, A. A. Efros, and M. Hardt, “Test-time training with self-supervision for generalization under distribution shifts,” in *Proceedings of the 37th International Conference on Machine Learning*, 2020, pp. 9229–9248.
- [42] J. Liang, D. Hu, and J. Feng, “Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation,” in *Proceedings of the 37th International Conference on Machine Learning*, 2020, pp. 6028–6039.
- [43] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh *et al.*, “Learning transferable visual models from natural language supervision,” *arXiv preprint arXiv:2103.00020*, 2021.
- [44] B. Boecking *et al.*, “Making the most of text semantics to improve biomedical vision–language processing,” in *European Conference on Computer Vision*, 2022, pp. 1–21.
- [45] S. Zhang, Y. Xu, N. Usuyama *et al.*, “BiomedCLIP: A multimodal biomedical foundation model pretrained on PMC-15M,” *arXiv preprint arXiv:2303.00915*, 2023.
- [46] Y. Li, Y. Fu, I. Gayo, Q. Yang, Z. Min, S. Saeed, W. Yan, Y. Wang, J. A. Noble, M. Emberton, M. J. Clarkson, H. Huisman, D. Barratt, V. A. Prisacariu, and Y. Hu, “Prototypical few-shot segmentation for cross-institution male pelvic structures with spatial registration,” *arXiv preprint arXiv:2209.05160*, 2022.
- [47] R. Wang and G. Zheng, “PFMNet: Prototype-based feature mapping network for few-shot domain adaptation in medical image segmentation,” *Computerized Medical Imaging and Graphics*, vol. 116, p. 102406, 2024.
- [48] S. Tang, S. Yan, X. Qi, J. Gao, M. Ye, J. Zhang, and X. Zhu, “Few-shot medical image segmentation with high-fidelity prototypes,” *arXiv preprint arXiv:2406.18074*, 2024.

- [49] S. Yang, Y. Wang, J. van de Weijer, L. Herranz, and S. Jui, “Source free domain adaptation for medical image segmentation with fourier style mining,” *Medical Image Analysis*, vol. 79, p. 102457, 2022.
- [50] Y. Zhang, T. Zhou, Y. Tao, S. Wang, Y. Wu, B. Liu, P. Gu, Q. Chen, and D. Z. Chen, “TestFit: A plug-and-play one-pass test time method for medical image segmentation,” *Medical Image Analysis*, vol. 92, p. 103069, 2024.
- [51] M. Bateson, H. Lombaert, and I. Ben Ayed, “Test-time adaptation with shape moments for image segmentation,” in *Medical Image Computing and Computer-Assisted Intervention*, 2022, pp. 736–745.
- [52] S. Hu, Z. Liao, Y. Zhang, and Y. Xia, “Unsupervised domain adaptation for brain structure segmentation via mutual information maximization alignment,” *Computer Methods and Programs in Biomedicine*, vol. 243, p. 107924, 2024.
- [53] Z. Zhang, W. Liu *et al.*, “Pass-GAN: Anatomy-aware style and shape guided domain adaptation for medical image segmentation,” *IEEE Transactions on Medical Imaging*, vol. 44, no. 5, pp. 1164–1173, 2025.
- [54] Q. Zhu, Z. Guo *et al.*, “CT-to-MRI cross-modality segmentation via multi-scale generative networks,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 2345–2354.
- [55] Y. Wu, X. Zhou *et al.*, “Contextual diffusion for medical image synthesis and domain adaptation,” *Medical Image Analysis*, vol. 85, p. 101452, 2024.
- [56] Y. Liu, P. Kothari, B. Van Delft, B. Bellot-Gurlet, T. Mordan, and A. Alahi, “TTT++: When does self-supervised test-time training fail or thrive?” *Advances in Neural Information Processing Systems*, vol. 34, pp. 21 808–21 820, 2021.
- [57] Z. Li, J. Yang, Y. Xu, L. Zhang, W. Dong, and B. Du, “Scale-aware test-time click adaptation for pulmonary nodule and mass segmentation,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2023, pp. 681–691.
- [58] J. Wu, G. Wang, R. Gu, T. Lu, Y. Chen, W. Zhu, T. Vercauteren, S. Ourselin, and S. Zhang, “UPL-SFDA: Uncertainty-aware pseudo label guided source-free domain adaptation for medical image segmentation,” *IEEE Transactions on Medical Imaging*, vol. 42, no. 12, pp. 3932–3943, 2023.

- [59] J. Omolegan, P. H. Yeung, M. K. Wyburd, L. Hesse, M. Haak, A. I. L. Namburete, N. K. Dinsdale, and I. 21st Consortium, “Exploring test time adaptation for subcortical segmentation of the fetal brain in 3D ultrasound,” in *IEEE International Symposium on Biomedical Imaging*, 2025, pp. 1–5.
- [60] A. Omid, A. Mohammadshahi, N. Gianchandani, R. King, L. Leijser, and R. Souza, “Unsupervised domain adaptation of MRI skull-stripping trained on adult data to newborns,” in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 7718–7727.
- [61] Z. Chen, Y. Ye, Y. Pan, and Y. Xia, “Gradient alignment improves test-time adaptation for medical image segmentation,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, 2025, pp. 2429–2437.
- [62] R. Wen, H. Yuan, D. Ni, W. Xiao, and Y. Wu, “From denoising training to test-time adaptation: Enhancing domain generalization for medical image segmentation,” in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 464–474.
- [63] Z. Chen, Y. Pan, Y. Ye, M. Lu, and Y. 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*, 2024, pp. 11 184–11 193.
- [64] P. Krähenbühl and V. Koltun, “Efficient inference in fully connected crfs with gaussian edge potentials,” in *Advances in Neural Information Processing Systems*, vol. 24, 2011.
- [65] M. C. de Verdier, R. Saluja, L. Gagnon, D. LaBella, U. Baid, N. H. Tahon, M. Foltyn-Dumitru, J. Zhang, M. Alafif, S. Baig, M. others, S. Baig, M. others, J. Zhang, M. Alafif, S. Baig *et al.*, “The 2024 brain tumor segmentation (BraTS) challenge: glioma segmentation on post-treatment MRI,” *arXiv preprint arXiv:2405.18368*, 2024.
- [66] N. C. F. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler, and A. Halpern, “Skin lesion analysis toward melanoma detection: A challenge at the 2017 International Symposium on Biomedical Imaging (ISBI), hosted by the International Skin Imaging Collaboration (ISIC),” in *IEEE 15th International Symposium on Biomedical Imaging*, 2018, pp. 168–172.

- [67] S. Zhu, Y. Chen, S. Jiang, W. Chen, C. Liu, Y. Wang, X. Chen, Y. Ke, F. Qin, C. Wang, and Z. Zhu, “XLSTM-HVED: Cross-modal brain tumor segmentation and MRI reconstruction method using vision XLSTM and heteromodal variational encoder-decoder,” in *IEEE 22nd International Symposium on Biomedical Imaging*, 2025, pp. 1–5.
- [68] S. Zhu, Y. Chen, W. Chen, Y. Wang, C. Liu, S. Jiang, F. Qin, and C. Wang, “Bridging the gap in missing modalities: Leveraging knowledge distillation and style matching for brain tumor segmentation,” in *Proceedings of Medical Image Computing and Computer Assisted Intervention*, vol. LNCS 15967, 2025, pp. 95–106.
- [69] S. Zhu, Y. Chen, W. Chen, S. Jiang, G. Zhou, Y. Wang, F. Qin, C. Wang, and Q. Tian, “No modality left behind: Adapting to missing modalities via knowledge distillation for brain tumor segmentation,” *arXiv:2509.15017*, 2025.
- [70] Y. Wang, Y. Chen, S. Jiang, W. Yu, M. Liu, B. Wu, J. Zong, F. Qin, C. Wang, and Q. Tian, “SmaRT: Style-modulated robust test-time adaptation for cross-domain brain tumor segmentation in MRI,” *arXiv:2509.17925*, 2025.
- [71] H. Yang, C. Chen, M. Jiang, Q. Liu, J. Cao, P. A. Heng, and Q. Dou, “DLTTA: Dynamic learning rate for test-time adaptation on cross-domain medical images,” *IEEE Transactions on Medical Imaging*, vol. 41, no. 12, pp. 3575–3586, 2022.
- [72] H. Wang, J. Chen, S. Zhang, Y. He, J. Xu, M. Wu, J. He, W. Liao, and X. Luo, “Dual-reference source-free active domain adaptation for nasopharyngeal carcinoma tumor segmentation across multiple hospitals,” *IEEE Transactions on Medical Imaging*, vol. 43, no. 12, pp. 4078–4090, 2024.
- [73] H. Li, H. Li, J. Chen, R. Zhong, K. Niu, H. Fu, and J. Liu, “AIF-SFDA: Autonomous information filter driven source-free domain adaptation for medical image segmentation,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, no. 5, 2025, pp. 4716–4724.
- [74] C. Zhang, H. Zheng, X. You, Y. Zheng, and Y. Gu, “PASS: test-time prompting to adapt styles and semantic shapes in medical image segmentation,” *IEEE Transactions on Medical Imaging*, 2024.