Topology-Inspired Backward-Free Framework for Test-Time Adaptation in Medical Detection

Bin Pu[1], Xingguo Lv[1], Jiewen Yang[2], Kai Xu[4], Lei Zhao[1], Zuozhu Liu[3], Kenli Li[1]*

[1]Hunan University, Changsha, China
[2]The Hong Kong University of Science and Technology, HKSAR, China
[3]Zhejiang University, Hangzhou, China
[4]Yunnan University, Kunming, China
{pubin, lvxg, zhaolei, lkl}@hnu.edu.cn, jyangcu@connect.ust.hk, zuozhuliu@intl.zju.edu.cn

Abstract

Recently, Test-Time Adaptation (TTA) has gained increasing attention in medical imaging due to its ability to improve model generalization under domain shifts without retraining. In particular, directly applying a well-trained model across various medical centers faces significant performance degradation caused by variations in equipment, operators, imaging conditions, and scanning skill levels of sonographers. Existing TTA methods either rely on parameter adaptation that increases computational cost or apply simple prediction fusion that ignores anatomical structure knowledge. To address these limitations, we propose a novel backward-free ***T*opology-aware *TTA*** framework named $\underline{T^3A}$ that integrates Structural Perception Modeling (SPM) and Box Regression Adaptation (BRA). SPM is implemented through an organ space heatmap generated via Gaussian kernel superposition. This heatmap encodes anatomical topology without requiring additional training or source data. BRA further improves localization and classification by fusing detection outputs based on the contribution of detected results to anatomically meaningful peak points from the heatmaps. Extensive experiments were conducted across six cross-domain scenarios, and the results demonstrate that our method achieves state-of-the-art cross-domain detection performance while maintaining high efficiency, offering a practical and robust solution for real-world medical diagnostic applications.

Code — https://github.com/Yore0/T3A

Introduction

Test-Time Adaptation (TTA) (Chen et al. 2022) has become an essential technique in medical image analysis, aiming to improve model robustness and generalization during inference. In clinical settings, medical images often exhibit variations in acquisition protocols, patient positioning, and imaging equipment, leading to domain shifts that may compromise model performance (Pu et al. 2025; Zhao et al. 2022). TTA addresses this challenge by applying diverse transformations to input images and aggregating predictions across augmented versions. This strategy helps mitigate the impact of data variability and enhances prediction reliability, making it particularly valuable in medical imaging, where diagnostic accuracy is critical.

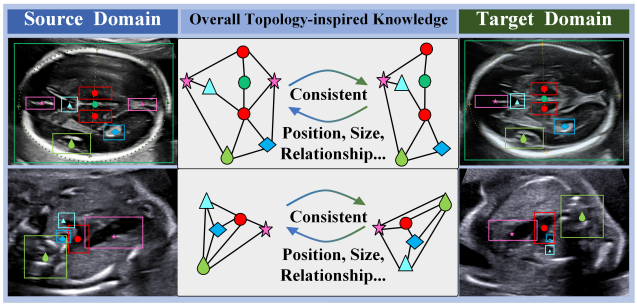

Figure 1: Motivation of our work. Medical images exhibit domain-invariant priors, as substructures within the same anatomical view remain consistent in terms of topology (e.g., position, size, and spatial relationships). This inherent consistency offers valuable insights for TTA, enabling more effective mitigation of domain shifts.

In medical image analysis, TTA offers specific advantages due to the high sensitivity of various data to acquisition conditions. For example, in fetal ultrasound imaging, fetus position, probe angle, maternal anatomy, and operator-dependent factors introduce significant variability, increasing the risk of inconsistent model behavior. By applying spatial augmentations (Shanmugam et al. 2021), such as rotations, flips, and scaling, during inference and fusing the resulting predictions, TTA effectively improves model invariance to such variations. This leads to more accurate localization and assessment of fetal structures, enhancing the robustness and clinical applicability of automated diagnostic systems in prenatal care.

Existing TTA methods have been proposed for both natural and medical image domains, and can be broadly classified into two paradigms: backpropagation-based (Chen et al. 2025b; Wang et al. 2021a) and hybrid (prior knowledge) approaches (Yang et al. 2022; Zhang et al. 2023a; Lv et al. 2025). Backpropagation-based methods, such as Tent (Wang et al. 2021a) and Gradient Alignment (Chen et al. 2025b), perform test-time optimization by minimizing predictive en-

*Corresponding author.

tropy or aligning gradients to dynamically adapt model parameters to the target domain. Hybrid approaches, on the other hand, leverage shared prior knowledge between the source and target domains to align invariant features for domain adaptation. While these methods offer strong adaptability, they often require careful regularization to prevent overfitting, and typically adjust the entire network or batch normalization (BN) layer weights during each test sample input via backpropagation. This results in inefficiencies in practical applications, where inference times are excessively prolonged. Moreover, a significant limitation of current approaches is the absence of structural perception modeling, the ability to capture and exploit the underlying spatial organization of anatomical structures. As illustrated in Figure 1, these methods often fail to model the topological consistency inherent in medical images, such as the relative positions, shapes, and spatial relationships of organs or substructures, which remain largely invariant across domains. When confronted with substantial domain shifts common in clinical imaging, these TTA methods struggle with both effectiveness and efficiency. This omission limits their capacity for anatomically consistent adaptation, especially when appearance-level cues degrade under severe domain shifts. Incorporating structural priors and topological constraints into TTA represents a promising yet underexplored direction for enhancing both robustness and clinical reliability.

To address the aforementioned limitations, we propose a novel TTA framework termed T^3A, which introduces two key components: the **Structural Perception Modeling (SPM)** tailored for medical imaging characteristics, and the **Box Regression Adaptation (BRA)** that simultaneously enhances localization precision and classification accuracy. Our method operates without acquring additional training or source data, enabling efficient and effective domain adaptation in real-time clinical settings. Specifically, in SPM, our approach incorporates topology-aware modeling of medical images by generating an Organ Space Heatmap through the superposition of Gaussian kernel functions. This design explicitly encodes anatomical spatial relationships without relying on extra data or backpropagation to learn topological structures, making it both parameter-free and highly interpretable. Building upon this heatmap, T^3A performs heatmap-guided BRA, where the peak points extracted from the aggregated heatmaps serve as anatomically informed anchors for object localization. Simultaneously, the peak points integrate with detection outputs to refine classification confidence by aggregating prediction probabilities across candidate boxes, resulting in more precise organ-aware bounding box regression. This feature is often overlooked in existing TTA methods that focus solely on entropy or confidence thresholds.

Our method results in more anatomically plausible and spatially coherent detection outcomes, particularly beneficial in medical imaging scenarios where multiple overlapping predictions may correspond to the same structure. The proposed framework thereby achieves superior robustness under severe domain shifts while maintaining computational efficiency and clinical applicability. Our primary contributions are summarized as follows:

- We propose a novel TTA framework for medical object detection that explicitly captures the topological consistency of anatomical structures across augmented views through Structural Perception Modeling. This enables structural perception without the need for parameter learning or source data.
- We devise a heatmap-guided Box Regression Adaptation that improves both localization accuracy and classification confidence via contribution-weighted fusion, addressing limitations of conventional NMS and existing TTA strategies.
- Extensive experiments are conducted under six different domain transfer scenarios, and our method achieves state-of-the-art adaptive performance, showing significant improvements over existing TTA methods while maintaining low computational overhead suitable for clinical deployment.

Related Work

TTA in Natural Image Scenarios

Test-time adaptation (TTA) in the natural image domain has evolved from simple entropy-based approaches to more complex mechanisms involving memory, augmentations, and structured perception. One of the earliest and most influential methods is entropy minimization (Wang et al. 2021a), which adjusts BatchNorm statistics during inference. Follow-up works (Niu et al. 2023; Zhang et al. 2023b; Cho et al. 2024) aimed to enhance robustness through sharpness-aware optimization and feature perturbation. Memory-based methods (Zhang et al. 2023b) leverage prior test instances to guide personalized adaptation, while augmentation-based frameworks (Cho et al. 2024) reuse high-entropy features to improve generalization. Recent research also extends TTA to structured tasks, such as slot-based object perception (Prabhudesai et al. 2023) and object detection via IoU-guided pseudo-labeling (Wang et al. 2021b), eliminating the need for explicit labels. As TTA matured, security and stability became key concerns. Adversarial inputs (Wu et al. 2023) demonstrated severe vulnerabilities, prompting solutions like temporal memory regularization (Yuan, Xie, and Li 2023). Meanwhile, researchers began exploring cross-modal scenarios such as image-text retrieval (Yang et al. 2024; Li et al. 2024a), applying confidence-driven query mechanisms.

TTA in Medical Image Scenarios

TTA in medical imaging emerged due to the scarcity of annotated data and domain shifts from heterogeneous acquisition protocols. Early works focused on adaptive optimization (Yang et al. 2022; Chen et al. 2025b), including learning rate scheduling and gradient-based stability control. Normalization-based methods (Yuan et al. 2024; Zhang et al. 2023a; Dong et al. 2025) adapted source-target statistics or applied pseudo-label regularization in 3D tasks. Graph structures and morphological priors were used for domain generalization during testing (Lv et al. 2025). Benchmark platforms like TTAbase (Zhang et al. 2025) support reproducible evaluations across datasets. Dong *et al.* (Dong

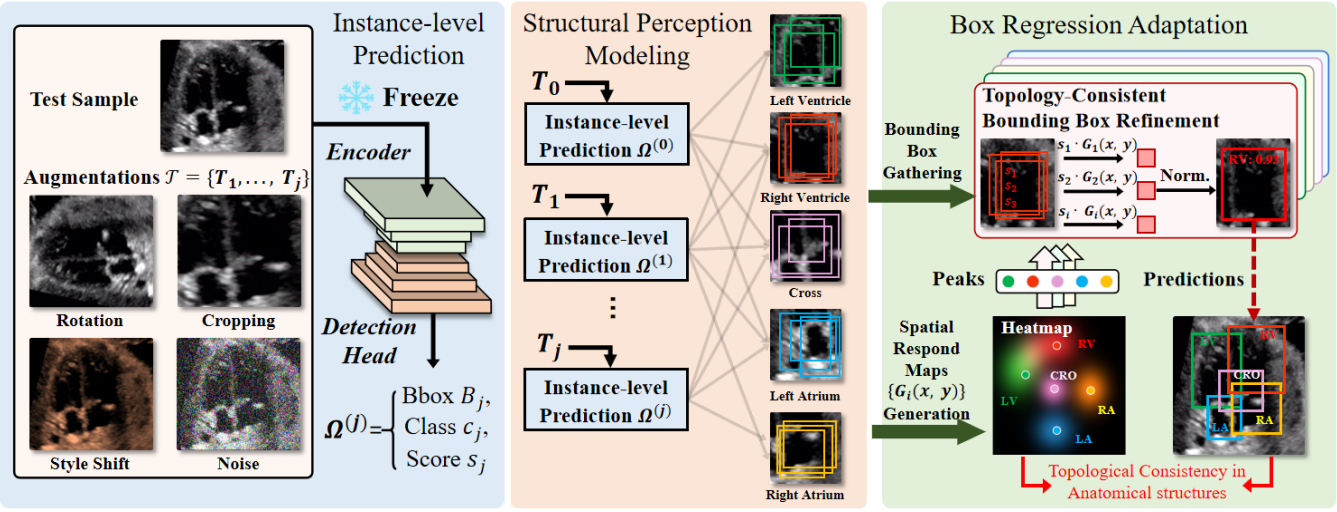

Figure 2: Overview of T^3A. During inference, each test sample undergoes multiple randomized augmentations and is fed into the pretrained model to generate instance-level predictions. Structural Perception Modeling captures the spatial distribution of organ centers across augmented views and learns topological consistency of anatomical structures. These structural priors guide the final bounding box regression adaptation via peak responses on the aggregated heatmap.

et al. 2024) proposed Integrated Entropy Weighting (InTEnt) for single-image adaptation. For streaming scenarios, VPTTA (Chen et al. 2024) aligns statistics via visual prompts. Zhang *et al.* (Zhang et al. 2024) introduced the PASS framework to adapt both styles and semantics. Similarly, Basak and Yin (Basak and Yin 2024) designed QUEST to build "source-like clones" without source data. Li *et al.* (Li et al. 2024b) enhanced slice consistency via cached feature regularization. Valanarasu *et al.* (Valanarasu et al. 2024) proposed a backpropagation-free TTA for efficient deployment. Human-in-the-loop strategies (Hu et al. 2024) incorporated clinician feedback. For foundation models, Chen *et al.* (Chen et al. 2025a) proposed non-parametric adaptation, and Aleem *et al.* (Aleem et al. 2024) introduced SaLIP for zero-shot segmentation via prompt-based reasoning.

Methodology

As shown in Figure 2, we propose a backward-free test-time adaptation framework specifically designed for medical image detection called T^3A. This framework consists of Structural Perception Modeling (SPM) for medical data and Box Refinement Adaptation (BRA) to improve the precision of the bounding box and the accuracy of the classification simultaneously, without requiring additional training. Specifically, SPM captures the spatial distribution of organ centers across augmented views and learns the topological consistency of anatomical structures, generating corresponding heatmaps. These heatmaps guide the final BRA through peak responses, which are then used to regress the coordinates of the bounding boxes for each organ. Unlike previous methods that rely on training-based updates or backpropagation, our framework operates solely during the inference phase, making it computationally efficient and eliminating the risk of error propagation or model instability.

Test-time Adaptation with Non-Maximum Suppression

Given a target domain sample X, a pre-trained object detection model θ, trained on the source domain, generates a set of instance-level predictions $\Omega = \{(B_i, c_i, s_i)\}_{i=1}^{N}$, where $B_i \in R^4$ denotes the bounding box, $c_i \in \{1, \ldots, C\}$ is the category label, and $s_i \in [0, 1]$ is the confidence score for the i-th detected object. Through extensive empirical analysis, we observe that exclusively refining or filtering category labels $\{c_i\}$ provides limited performance gain in the detection context. In fact, due to a significant amount of false positives and false negatives, the overall detection robustness of the model is considerably degraded.

A widely adopted technique to improve robustness is Test-Time Augmentation, which applies a set of geometric and photometric transformations $\mathcal{T} = \{T_j\}$ (e.g., horizontal/vertical flipping, rotation, cropping, noise injection, saturation shift) to the input sample X, yielding augmented samples $\{X_j = T_j(X)\}$. Each augmented sample is then passed through the detector θ, generating transformed predictions Ω_j. These predictions are subsequently inverse-transformed $\Omega'_j = T_j^{-1}(\Omega_j)$ to align with the original input coordinates. Finally, all $\{\Omega'_j\}$ are fused, usually via non-maximum suppression (NMS) (Girshick et al. 2014; Girshick 2015), to produce the final result of the inference. We refer to this method as TTA-NMS.

While TTA-NMS has demonstrated substantial improvements in general object detection benchmarks, it exhibits limitations when applied to medical image detection tasks. This is primarily due to the augmentation and fusion process being agnostic to anatomical topology and contextual constraints, thereby failing to preserve the inherent anatomical consistency and spatial regularity characteristic of medical images. We define this property as **Topological Consistency**

of Organ Space. For instance, in the four-chamber view of the heart, the spatial pattern—left ventricle (LV), right ventricle (RV), left atrium (LA), right atrium (RA)—remains structurally consistent regardless of photometric or moderate geometric transformations. This characteristic offers a unique opportunity to design TTA strategies that respect such anatomical priors, rather than treating predictions as independent and identically distributed across augmentations.

Structural Perception Modeling with Topological Consistency

Medical images exhibit domain-invariant priors, as substructures within the same anatomical view consistently preserve topological properties such as position, size, and spatial relationships. Recent studies have explored learning-based methods to model anatomical priors (Pu et al. 2024a; Lv et al. 2025), such as using graph structures to encode inter-organ spatial relationships (Pu et al. 2025), explicit shape templates for individual organs (Liu et al. 2022), or spatial relation modules and organ layout predictors (Pu et al. 2024a). While effective, these approaches typically require large-scale annotated datasets and extensive training, which contradicts the source-free and parameter-efficient paradigm of test-time adaptation. To overcome this limitation, we propose a parameter-free and backward-free framework that enforces topological consistency via Gaussian Kernel Fusion (GKF). GKF models the spatial distribution of organ centers across augmented views of a test sample using kernel density estimation, capturing stable anatomical layouts without additional learning.

Given a test sample X and a set of its augmented variants $\{X_j\}_{j=1}^{M}$, the source-trained detector θ performs independent inference on each version to produce a set of predicted instances: $\Omega^{(0)} = \{(B_i, c_i, s_i)\}_{i=1}^{N_0}$ for the original sample and $\Omega^{(j)} = \{(B_i^{(j)}, c_i^{(j)}, s_i^{(j)})\}_{i=1}^{N_j}$ for each augmented sample X_j. These predictions are then aggregated into a unified instance set $\mathbf{\Omega} = \bigcup_{j=0}^{M} \Omega^{(j)} = \{(B_i, c_i, s_i)\}_{i=1}^{N}$, where each box B_i provides a center location $(\mu_{x,i}, \mu_{y,i})$, width w_i, and height h_i. We generate a spatial response map $G_i(x, y)$ for each box by placing an anisotropic 2D Gaussian kernel centered at its geometric center:

$$G_i(x,y) = \exp\left(-\left[\frac{(x-\mu_{x,i})^2}{2\sigma_{x,i}^2} + \frac{(y-\mu_{y,i})^2}{2\sigma_{y,i}^2}\right]\right), \quad (1)$$

where the kernel bandwidths are adaptive to the box size, given by $\sigma_{x,i} = \alpha \cdot w_i, \sigma_{y,i} = \alpha \cdot h_i$, and α is a scaling hyperparameter controlling the kernel spread and shape. To aggregate detection confidence for organ classes C, we define a organ-specific heatmap by summing the score-weighted kernels over all boxes with predicted label κ:

$$H_\kappa(x,y) = \sum_{i:c_i=\kappa} s_i \cdot G_i(x,y). \quad (2)$$

The resulting map $H_\kappa(x, y)$ models the spatial density of each anatomical structure by fusing confident and topologically aligned detections. Regions supported by multiple bounding boxes exhibit higher consensus and are thus assigned stronger responses. Misaligned or spurious boxes are naturally down-weighted due to lower overlap in the response space, while anatomically consistent structures are amplified, leading to more robust detections.

Topology-Consistent Box Refinement Adaptation

Existing TTA-based detection method (Yuan et al. 2024) attempts to refine bounding boxes at the test time by adapting the regression layers through gradient-based optimization. However, in medical imaging scenarios, test samples frequently exhibit significant domain discrepancies due to variations in imaging devices, acquisition protocols, or clinical practices across institutions. This leads to a heightened risk of catastrophic forgetting, model collapse, and instability, particularly when model parameters are updated independently per test image. To avoid these pitfalls, we propose a topology-consistent refinement that operates without altering the network parameters.

Leveraging the anatomical heatmaps $H_\kappa(x, y)$ generated in the previous subsection, which encode spatial densities of organ presence. We refine bounding boxes by explicitly aligning them to spatial anatomical priors. Specifically, we identify candidate anatomical centers by extracting organ-wise local maxima from the heatmaps. The set of spatial peaks for organ κ is defined as:

$$\mathcal{P}_\kappa = \left\{ \begin{matrix} (x_k, y_k) = \arg \max\limits_{(x,y)\in\mathcal{N}_\epsilon(x_k,y_k)} H_\kappa(x,y); \\ H_\kappa(x,y) > \tau \end{matrix} \right\}, \quad (3)$$

where $\mathcal{N}_\epsilon(x_k, y_k)$ denotes a local neighborhood of kernel ϵ centered at (x_k, y_k), and τ is a peak activation threshold. Eq. (3) ensures that peaks correspond to locally maximal activations likely to represent organ centers. The peak detection is performed organ-wise to avoid inter-class interference in spatial proximity.

For each peak $(x_k, y_k) \in \mathcal{P}_\kappa$, we compute the spatial affinity between the peak and each predicted box B_i of organ class κ using an anisotropic Gaussian kernel centered at the box location $(\mu_{x,i}, \mu_{y,i})$, weighted by the predicted detection confidence s_i:

$$w_i^{(k)} = s_i \cdot \exp\left(-\left[\frac{(x_k-\mu_{x,i})^2}{2\sigma_{x,i}^2} + \frac{(y_k-\mu_{y,i})^2}{2\sigma_{y,i}^2}\right]\right), \quad (4)$$

where $\sigma_{x,i}$ and $\sigma_{y,i}$ are defined in Eq. (1). Boxes with weights $w_i^{(k)}$ exceeding a threshold η are considered contributing boxes and collected into the set $\mathcal{C}_k$, corresponding to peak (x_k, y_k).

Instead of applying conventional NMS, which discards overlapping predictions based solely on confidence, we adopt a soft fusion strategy inspired by weighted boxes fusion (Solovyev, Wang, and Gabruseva 2021). The rationale is that NMS is blind to anatomical topology and often removes valid boxes near anatomical boundaries, especially in low-contrast or multi-instance regions. In contrast, weighted boxes fusion enables joint integration of multiple overlapping detections, yielding spatially coherent predictions. The refined box $\hat{B}_k$ associated with peak (x_k, y_k) is computed as

Method	Center 1→2										Center 2→1									
	LA	RA	LV	RV	CR	R	VS	SP	DAO	**mAP**	LA	RA	LV	RV	CR	R	VS	SP	DAO	**mAP**
No Adapt	59.59	45.88	53.04	52.30	64.54	57.89	62.84	57.57	61.43	57.20	70.56	49.82	54.26	61.88	73.04	64.48	74.25	77.32	53.28	64.32
TENT (ICLR'21) (Wang et al. 2021a)	37.22	41.75	43.61	35.13	43.31	32.59	59.89	50.37	76.44	46.70	83.65	87.57	85.66	85.09	87.89	86.02	85.17	82.24	67.91	83.47
DLTTA (TMI'22) (Yang et al. 2022)	19.70	29.47	32.38	28.19	28.70	38.93	34.94	36.83	36.09	32.13	47.01	49.63	47.93	47.04	47.10	39.28	48.56	48.52	47.94	47.05
DomainAdaptor (ICCV'23) (Zhang et al. 2023a)	31.09	44.21	26.76	25.78	28.42	28.92	31.59	50.36	55.07	35.80	84.59	87.95	84.23	82.53	85.48	**90.51**	82.07	87.08	68.95	83.71
MonoTTA (ECCV'24) (Lin et al. 2024)	30.50	34.60	31.04	31.70	29.59	38.07	44.28	48.70	45.64	37.12	81.30	91.09	89.82	81.77	82.61	90.44	84.84	84.91	63.51	83.37
VPTTA (CVPR'24) (Chen et al. 2024)	53.71	57.68	68.64	60.42	66.29	59.61	75.04	65.70	80.68	65.31	87.08	89.22	86.78	86.83	87.95	87.75	85.77	90.59	67.36	85.48
TTDG-MGM (CVPR'25) (Lv et al. 2025)	56.98	56.59	70.21	61.07	70.44	59.50	70.71	69.76	74.90	65.57	**88.67**	89.53	87.31	87.27	87.91	87.46	86.85	**91.51**	70.86	86.37
GraTa (AAAI'25) (Chen et al. 2025b)	52.26	56.15	68.31	65.48	69.96	62.61	75.35	67.18	80.10	66.38	86.39	89.08	86.10	86.76	87.93	86.77	85.72	89.72	65.70	84.91
T^3A (Ours)	**61.20**	**68.45**	**79.47**	**67.46**	**84.53**	**68.74**	**86.71**	**82.15**	**81.31**	**75.57**	82.62	**92.78**	**93.63**	**90.61**	**94.81**	87.14	**86.88**	88.17	**87.71**	**89.37**

Table 1: Performance comparison of various methods in the heart of $FUSH^2$ dataset. *No Adapt* denotes the baseline that was trained only in the source domain without using any adaptation method. The best results in each column are **bold**.

the weighted average of its contributing boxes:

$$\hat{B}_k = \frac{1}{\sum_{j \in \mathcal{C}_k} w_j^{(k)}} \sum_{i \in \mathcal{C}_k} w_i^{(k)} \cdot B_i. \tag{5}$$

For a given organ class κ, the refined detection results are given by the set $\{\hat{B}_k\}_{k=1}^{|\mathcal{P}_\kappa|}$, which serves as the final output of the source model θ on the test sample. This process ensures that each anatomical region is represented by a consensus box derived from topologically consistent spatial evidence, enhancing both geometric fidelity and semantic stability.

Result

Datasets

$FUSH^2$ (Pu et al. 2024a) is a publicly accessible resource comprising 3,369 fetal ultrasound images of cardiac and cranial views. These images were acquired across two distinct medical centers using diverse ultrasound systems, including devices from Samsung, Sonoscape, and Philips. The dataset features 16 anatomical regions annotated with bounding boxes and semantic labels: nine structures for cardiac views (e.g., ventricles, atria) and seven for cranial views (e.g., brain ventricles, skull landmarks). In this study, we evaluate TTA performance between the two centers, with experiments conducted in both directions (center 1→2 and center 2←1). This cross-center setup enables rigorous assessment of model robustness under domain shifts caused by scanner heterogeneity and imaging protocol variations.

FCS (Pu et al. 2024b) comprises two fetal cardiac views, 4-Chamber Cardiac View (4CC) and 3-Vessel Trachea View (3VT), collected from two medical centers (A and B) for fetal congenital heart defect screening. It includes four cross-domain adaptation experiments: center A→B and B→A on both FCS-4CC and FCS-3VT. The ultrasound images are acquired using diverse ultrasound devices (e.g., Samsung, Sonoscape, Philips), span 20–34 weeks of gestation. Experienced ultrasonographers annotated anatomical structures and view categories. While standard 4CC images typically contain nine anatomical labels, some samples in the dataset exhibit fewer identified structures due to variability in imaging quality or fetal positioning.

Experimental Setting

Training. To ensure a fair comparison across all baseline methods, we conducted experiments using the same source model weights. We adopted ResNet-50 as the feature extractor and employed Faster R-CNN as the detection head. During training, we utilized the Stochastic Gradient Descent (SGD) optimizer with an initial learning rate of 0.01, a batch size of 8, and a total of 10,000 iterations. We applied multi-scale training by randomly resizing the short side of each image to one of the following values: $\{640, 672, 704, 736, 768, 800\}$, while proportionally scaling the long side. In addition, random horizontal flipping was applied as part of data augmentation.

Inference. During inference, all input data were obtained from medical centers distinct from those used for training the source model, thereby ensuring that the model encountered entirely unseen data. Within the proposed T^3A framework, inference was performed with a batch size of 1. Each test sample was subjected to a sequence of image scaling and transformation operations. Specifically, the short side was sequentially resized to values from the range $\{500, 600, ..., 1200\}$, while the long side was limited to a maximum of 4000 pixels. Based on these resized versions, further transformations were applied, such as horizontal flipping, rotation, noise injection, and cropping. For ensemble prediction, both the purely resized images and their transformed counterparts were preserved.

Comparison with State-of-the-arts

Result on $FUSH^2$. As illustrated in Table 1, the quantitative analysis of the $FUSH^2$ reveals significant advantages of our proposed method over existing approaches in fetal heart structure detection tasks. In the Center 1→2, our method achieves an mAP of 89.37%, outperforming all other methods. Similarly, in the Center 2→1, our method attains an mAP of 87.71%, surpassing competitors such as TTDG-MGM (86.37%), GraTa (84.91%), and DomainAdaptor (85.48%). Our method excels in detecting specific heart structures as well. For example, in Center 1→2, it achieves higher precision for LA (61.20%), RA (68.45%), and LV (79.47%) compared to other methods. These results highlight the robustness and adaptability of our method, making it a promising solution for cross-center fetal ultrasound image analysis.

Result on FCS. As listed in Table 2, the quantitative evaluation of the FCS dataset demonstrates the distinct superiority of our method in addressing domain adaptation challenges in heart structure detection. Across the two

Method	Center A→B										Center B→A									
	LA	RA	LV	RV	CR	R	VS	SP	DAO	**mAP**	LA	RA	LV	RV	CR	R	VS	SP	DAO	**mAP**
No Adapt	56.57	62.95	71.84	62.67	62.11	77.57	72.83	72.02	73.84	68.04	65.00	75.78	81.58	74.95	70.66	83.30	83.60	77.81	81.75	77.16
TENT (ICLR'21) (Wang et al. 2021a)	59.76	68.86	71.60	62.77	69.84	63.34	69.50	62.89	82.06	67.85	70.33	86.48	85.09	74.28	92.31	79.96	89.00	72.76	**72.26**	80.28
DLTTA (TMI'22) (Yang et al. 2022)	30.85	40.91	31.91	29.89	35.73	43.61	33.87	33.25	31.48	34.38	46.52	46.43	45.00	40.27	46.66	37.46	48.54	47.32	46.12	44.79
DomainAdaptor (ICCV'23) (Zhang et al. 2023a)	60.34	71.28	65.54	61.88	62.37	60.77	58.12	64.54	72.20	64.12	87.27	89.29	90.00	80.44	84.84	86.13	89.51	82.41	64.79	83.85
MonoTTA (ECCV'24) (Lin et al. 2024)	48.62	61.44	56.77	48.71	48.84	54.24	48.37	51.61	57.23	52.87	22.97	55.57	16.78	9.57	17.49	64.04	46.91	59.81	13.65	34.08
VPTTA (CVPR'24) (Chen et al. 2024)	60.91	71.12	70.43	66.15	72.22	64.07	72.04	66.41	84.66	69.78	85.10	90.82	87.98	83.11	93.38	88.42	**91.42**	80.21	68.99	85.49
TTDG-MGM (CVPR'25) (Lv et al. 2025)	70.48	80.39	75.43	73.66	77.11	73.48	73.68	70.32	82.91	75.27	84.29	86.82	86.14	83.96	87.56	86.73	90.31	80.98	70.68	84.16
GraTa (AAAI'25) (Chen et al. 2025b)	59.49	71.30	70.87	66.51	72.31	64.94	73.73	65.47	84.59	69.91	84.83	91.20	88.22	81.63	**93.97**	88.50	91.39	79.03	70.00	85.42
T^3A (Ours)	**77.94**	**81.07**	**81.41**	**78.12**	**85.92**	**84.75**	**91.05**	**82.90**	**90.19**	**83.71**	**91.12**	**93.59**	**90.57**	**85.67**	92.81	**89.79**	90.94	**86.44**	71.10	**88.00**

Table 2: Performance comparison of various methods in the heart of FCS dataset.

Method	Center 2→B										Center B→2									
	LA	RA	LV	RV	CR	R	VS	SP	DAO	**mAP**	LA	RA	LV	RV	CR	R	VS	SP	DAO	**mAP**
No Adapt	75.74	62.25	81.02	77.46	78.80	78.39	76.13	73.58	78.23	75.83	48.86	75.43	66.74	63.38	67.00	57.61	65.95	69.58	55.53	63.34
TENT (ICLR'21) (Wang et al. 2021a)	59.52	74.17	72.40	68.33	71.19	69.26	80.80	59.92	81.60	70.79	63.99	52.42	70.48	64.30	54.05	56.87	53.47	69.78	65.91	67.01
DLTTA (TMI'22) (Yang et al. 2022)	42.39	43.81	39.07	40.03	44.64	45.71	40.38	46.66	39.23	42.07	29.10	32.94	33.23	28.10	34.93	40.52	38.05	38.73	38.41	34.90
DomainAdaptor (ICCV'23) (Zhang et al. 2023a)	80.38	83.64	79.03	78.44	81.85	86.24	79.03	83.11	85.97	81.96	38.69	46.00	43.75	36.11	42.25	40.76	63.33	68.36	65.65	49.43
MonoTTA (ECCV'24) (Lin et al. 2024)	65.19	78.86	67.51	60.68	65.62	79.06	78.22	71.06	70.18	70.71	14.24	38.19	11.78	16.75	28.94	42.65	44.04	61.38	25.55	31.50
VPTTA (CVPR'24) (Chen et al. 2024)	83.56	85.76	84.47	82.31	86.86	88.65	87.28	81.06	86.47	85.15	54.60	59.45	54.64	54.78	64.59	59.96	76.52	75.26	81.45	64.58
TTDG-MGM (CVPR'25) (Lv et al. 2025)	85.61	89.96	**88.81**	87.00	92.91	90.22	92.52	86.92	90.31	89.36	59.64	65.17	63.71	61.10	71.70	63.63	77.47	78.69	82.96	69.34
GraTa (AAAI'25) (Chen et al. 2025b)	82.51	85.23	84.96	83.43	86.97	88.58	88.37	81.47	87.43	85.43	52.78	60.28	53.94	52.74	64.31	59.03	74.53	73.84	81.54	63.66
T^3A (Ours)	**91.87**	**91.60**	88.17	**90.09**	**95.84**	**92.01**	**93.48**	**89.86**	**92.75**	**91.74**	**68.04**	**75.87**	**74.27**	**68.38**	**80.93**	**75.30**	**84.67**	**88.29**	**89.60**	**78.37**

Table 3: Performance comparison of various methods in heart of the FCS and $FUSH^2$ datasets.

domains (Center A→B and B→A), our approach achieves an mAP of 88.00%, significantly outperforming the No Adapt baseline (57.20% and 64.32%) and other SOTA methods. DomainAdaptor shows moderate improvement with 72.20% (A→B) and 83.85% (B→A), yet still lags behind our method. Notably, MonoTTA underperforms significantly (57.23% and 34.08%), underscoring its instability in cross-domain settings. Hybrid approaches, VPTTA (69.78% and 75.27%) and TTDG-MGM (69.91% and 75.27%) further validate the gap between their designs and our framework. The results solidify the practicality of our approach for real-world clinical applications involving diverse data sources.

Result on FCS and $FUSH^2$. As shown in Table 3, our method also achieves the best performance compared with various advanced approaches. Compared with No Adapt method, it demonstrates improvements of 15.91% and 15.03% for 2→B and B→2, respectively. Our method outperforms the second-best (TTDG-MGM) by 2.38% and 9.03% in these two scenarios. These results solidify the practicality of our approach for real-world clinical applications involving diverse data sources.

Qualitative Analysis

Qualitative comparisons with baseline methods are shown in Fig. 4. Across all baseline methods—including those using fixed source weights, BN-based adaptation (TENT, VPTTA), backpropagation-based TTA (TTDG-MGM), and adaptive learning rate strategies (Grata)—errors such as missed detections and false positives remain prevalent. In contrast, incorporating anatomical structure-aware modeling with topological consistency significantly mitigates these issues. As shown in the PR curve (Fig. 3(a)), our method achieves higher precision at nearly all recall levels. Moreover, in Fig. 3(b–d), the bounding boxes generated by the proposed Topology-Consistent Adaptation strategy are more closely aligned with ground truth, in terms of width, height, and area distributions, further demonstrating the robustness of our approach in the TTA scenario.

Augment	GKF	Box Fusion	Center1→2	Center2→1
✗	✗	-	62.93	82.07
✓	✗	NMS	72.46	86.25
✓	✓	NMS	73.20	86.09
✓	✓	WBF	73.44	87.88
✓	✓	BRA (Ours)	**75.57**	**89.37**

Table 4: Ablation results (mAP) on augmentation, GKF, and fusion strategies on $FUSH^2$.

Flip	Rotate	Noise	Crop	Resize	mAP (%)
✗	✗	✗	✗	$\{800\}$	62.93
✓	✗	✗	✗	$\{800\}$	63.62
✓	✗	✗	✗	$\{800\} \times 2$	63.98
✓	✓	✗	✗	$\{800, 1200\}$	66.22
✓	✓	✗	✗	$\{800, 1200\} \times 2$	67.05
✓	✓	✓	✗	$\{600, 800, 1000, 1200\}$	66.89
✓	✓	✓	✗	$\{600, 800, 1000, 1200\} \times 2$	68.14
✓	✓	✓	✗	$\{500, ..., 1200\} \times 2$	69.56
✓	✓	✓	✓	$\{500, ..., 1200\}$	70.71
✓	✓	✓	✓	$\{500, ..., 1200\} \times 2$	**72.46**

Table 5: Effect of different augmentations on Center 1→2 of $FUSH^2$. Short side resized to $\{500, \ldots, 1200\}$ and "×2" denotes inclusion of both resized and transformed images.

Ablation Study

Effect of SPM. The core of Structural Perception Modeling (SPM) is Gaussian Kernel Fusion (GKF), which models the spatial distribution of organ centers across augmented views of a test sample using kernel density estimation. Ablation studies were conducted to assess the effectiveness of augment and GKF.

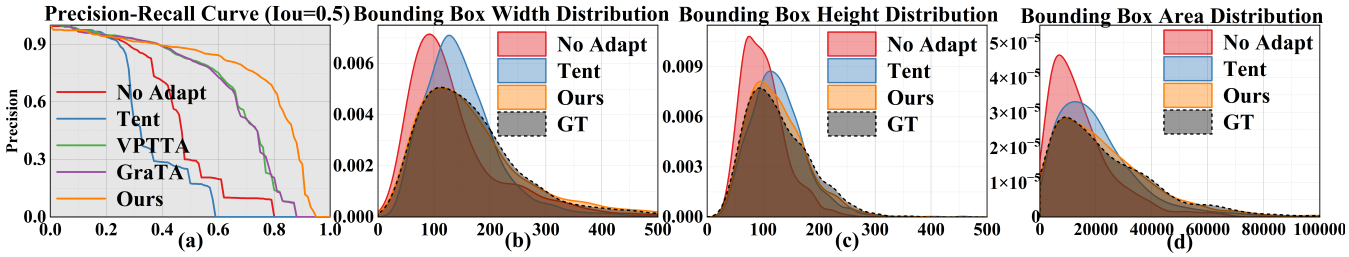

Figure 3: Quantitative comparison with baseline methods. Our method achieves the highest consistency with ground truth, as demonstrated by the Precision-Recall curve (a) and the distributions of bounding box width (b), height (c), and area (d).

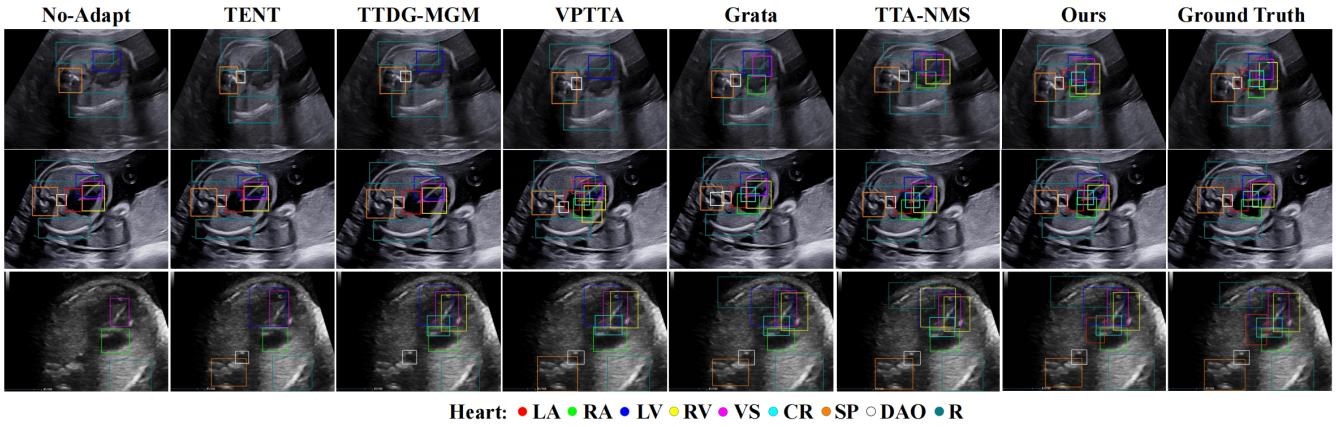

Figure 4: Qualitative result comparison on FUSH^2 and FCS datasets. The top two rows show samples from Center 1 and Center 2 of FUSH^2, respectively, while the bottom row presents samples from Center B of FCS.

Effect of η and τ			Effect of α	
η	τ	mAP (%)	α	mAP (%)
0.05	0.05	74.84	0.1	75.08
0.10	0.05	**75.08**	0.2	**75.57**
0.10	0.10	74.63	0.5	75.29
0.20	0.20	74.28	1.0	74.61

Table 6: Effect of hyperparameters η, τ, and α on Center 1→2 of FUSH^2.

As shown in Table 4, TTA significantly improved mAP by 9.53%, highlighting the role of low-level transformations in reducing domain shifts. Table 5 further shows that detection performance improves as more and diverse augmentations are applied. Notably, combining both resized and transformed variants outperforms using only transformed images, due to their complementary effects: resized images preserve structural integrity, while transformed ones enhance appearance diversity. Incorporating GKF yielded an additional 0.74% gain.

Effect of BRA. As shown in Table 4, we conducted ablation studies to compare Box Refinement Adaptation (BRA) with various box fusion strategies. Replacing standard NMS with Weighted Boxes Fusion (WBF) (Solovyev, Wang, and Gabruseva 2021) resulted in a modest 0.24% improvement; this limited gain suggests that WBF does not fully leverage domain-specific priors such as anatomical topology. BRA achieved a further 2.37% improvement over NMS.

Parameters Sensitivity. We also conducted ablation studies on the hyperparameters η, τ, and α. Specifically, α controls the bandwidth of the Gaussian kernel, τ defines the response threshold for local peak detection in Eq. (3), and η sets the minimum weight contribution for box refinement. Detailed results are provided in Table 6, we set the final hyperparameters to $\eta = 0.1$, $\tau = 0.05$, and $\alpha = 0.2$. For a more comprehensive ablation experiment and parameter analysis, please review the Appendix.

Conclusion

In this work, we propose a backward-free Test-Time Adaptation framework T^3A, which incorporates structural perception modeling through an organ space heatmap for anatomical adaptation detection in medical images. By leveraging structural perception modeling and a heatmap-guided box regression adaptation method, T^3A effectively encodes anatomical spatial relationships and enhances both localization precision and classification accuracy. This approach not only mitigates the impact of data variability but also improves the robustness and clinical applicability of automated diagnostic systems without the need for additional training data or extensive parameter tuning. Future work will focus on refining the adaptability to more complex scenarios and expanding its utility to diverse medical imaging tasks.

Acknowledgments

This work was supported in part by the National Key R&D Program of China under Grant 2025YFB3003705, in part by the National Natural Science Foundation of China under Grants 62227808, Grants 62506124, and in part by the Natural Science Foundation of Hunan Province under Grants 2025JJ60408.

References

Aleem, S.; Wang, F.; Maniparambil, M.; Arazo, E.; Dietlmeier, J.; Curran, K.; Connor, N. E.; and Little, S. 2024. Test-time adaptation with salip: A cascade of sam and clip for zero-shot medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5184–5193.

Basak, H.; and Yin, Z. 2024. Quest for Clone: Test-Time Domain Adaptation for Medical Image Segmentation by Searching the Closest Clone in Latent Space. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 555–566. Springer.

Chen, D.; Wang, D.; Darrell, T.; and Ebrahimi, S. 2022. Contrastive test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 295–305.

Chen, K.; Luo, X.; Qin, T.; Liu, J.; Liu, H.; Lee, V. H. F.; Yan, H.; and Li, H. 2025a. Test-time Adaptation for Foundation Medical Segmentation Model without Parametric Updates. *arXiv preprint arXiv:2504.02008*.

Chen, Z.; Pan, Y.; Ye, Y.; Lu, M.; and Xia, Y. 2024. Each test image deserves a specific prompt: Continual test-time adaptation for 2d medical image segmentation. In *Proc. IEEE/CVF Comput. Vis. Pattern Recog.*, 11184–11193.

Chen, Z.; Ye, Y.; Pan, Y.; and Xia, Y. 2025b. Gradient alignment improves test-time adaptation for medical image segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 2429–2437.

Cho, Y.; Kim, Y.; Yoon, J.; Hong, S.; and Lee, D. 2024. Feature Augmentation based Test-Time Adaptation. *arXiv preprint arXiv:2410.14178*.

Dong, H.; Konz, N.; Gu, H.; and Mazurowski, M. A. 2024. Medical image segmentation with intent: Integrated entropy weighting for single image test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5046–5055.

Dong, X.; Wang, L.; Lv, X.; Zhang, X.; Zhang, H.; Pu, B.; Gao, Z.; Liao, I. Y.; and Jin, Z. 2025. CertainTTA: Estimating uncertainty for test-time adaptation on medical image segmentation. *Information Fusion*, 103300.

Girshick, R. 2015. Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, 1440–1448.

Girshick, R.; Donahue, J.; Darrell, T.; and Malik, J. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 580–587.

Hu, S.; Liao, Z.; Liu, Z.; and Xia, Y. 2024. Towards Clinician-Preferred Segmentation: Leveraging Human-in-the-Loop for Test Time Adaptation in Medical Image Segmentation. *arXiv preprint arXiv:2405.08270*.

Li, H.; Hu, P.; Zhang, Q.; Peng, X.; Liu, X.; and Yang, M. 2024a. Test-time Adaptation for Cross-modal Retrieval with Query Shift. *arXiv preprint arXiv:2410.15624*.

Li, X.; Fang, H.; Wang, C.; Liu, M.; Duan, L.; and Xu, Y. 2024b. Cache-Driven Spatial Test-Time Adaptation for Cross-Modality Medical Image Segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 146–156. Springer.

Lin, H.; Zhang, Y.; Niu, S.; Cui, S.; and Li, Z. 2024. Monotta: Fully test-time adaptation for monocular 3d object detection. In *European Conference on Computer Vision*, 96–114. Springer.

Liu, Q.; Chen, C.; Dou, Q.; and Heng, P.-A. 2022. Single-domain generalization in medical image segmentation via test-time adaptation from shape dictionary. In *AAAI*, volume 36, 1756–1764.

Lv, X.; Dong, X.; Wang, L.; Yang, J.; Zhao, L.; Pu, B.; Jin, Z.; and Li, X. 2025. Test-Time Domain Generalization via Universe Learning: A Multi-Graph Matching Approach for Medical Image Segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 15621–15631.

Niu, S.; Wu, J.; Zhang, Y.; Wen, Z.; Chen, Y.; Zhao, P.; and Tan, M. 2023. Towards stable test-time adaptation in dynamic wild world. *arXiv preprint arXiv:2302.12400*.

Prabhudesai, M.; Goyal, A.; Paul, S.; Van Steenkiste, S.; Sajjadi, M. S.; Aggarwal, G.; Kipf, T.; Pathak, D.; and Fragkiadaki, K. 2023. Test-time adaptation with slot-centric models. In *International Conference on Machine Learning*, 28151–28166. PMLR.

Pu, B.; Lv, X.; Yang, J.; Dong, X.; Lin, Y.; Li, S.; Li, K.; and Li, X. 2025. Leveraging Anatomical Consistency for Multi-Object Detection in Ultrasound Images via Source-free Unsupervised Domain Adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 6532–6540.

Pu, B.; Lv, X.; Yang, J.; Guannan, H.; Dong, X.; Lin, Y.; Shengli, L.; Ying, T.; Fei, L.; Chen, M.; et al. 2024a. Unsupervised domain adaptation for anatomical structure detection in ultrasound images. In *Forty-first International Conference on Machine Learning*.

Pu, B.; Wang, L.; Yang, J.; He, G.; Dong, X.; Li, S.; Tan, Y.; Chen, M.; Jin, Z.; Li, K.; et al. 2024b. M3-uda: a new benchmark for unsupervised domain adaptive fetal cardiac structure detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11621–11630.

Shanmugam, D.; Blalock, D.; Balakrishnan, G.; and Guttag, J. 2021. Better aggregation in test-time augmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1214–1223.

Solovyev, R.; Wang, W.; and Gabruseva, T. 2021. Weighted boxes fusion: Ensembling boxes from different object detection models. *Image and Vision Computing*, 107: 104117.

Valanarasu, J. M. J.; Guo, P.; Patel, V. M.; et al. 2024. On-the-fly test-time adaptation for medical image segmentation. In *Medical Imaging with Deep Learning*, 586–598. PMLR.

Wang, D.; Shelhamer, E.; Liu, S.; Olshausen, B.; and Darrell, T. 2021a. Tent: Fully Test-Time Adaptation by Entropy Minimization. In *ICLR*.

Wang, H.; Cong, Y.; Litany, O.; Gao, Y.; and Guibas, L. J. 2021b. 3dioumatch: Leveraging iou prediction for semi-supervised 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14615–14624.

Wu, T.; Jia, F.; Qi, X.; Wang, J. T.; Sehwag, V.; Mahloujifar, S.; and Mittal, P. 2023. Uncovering adversarial risks of test-time adaptation. *arXiv preprint arXiv:2301.12576*.

Yang, H.; Chen, C.; Jiang, M.; Liu, Q.; Cao, J.; Heng, P. A.; and Dou, Q. 2022. Dltta: Dynamic learning rate for test-time adaptation on cross-domain medical images. *IEEE Transactions on Medical Imaging*, 41(12): 3575–3586.

Yang, M.; Li, Y.; Zhang, C.; Hu, P.; and Peng, X. 2024. Test-time adaptation against multi-modal reliability bias. In *The Twelfth International Conference on Learning Representations*.

Yuan, J.; Zhang, B.; Gong, K.; Yue, X.; Shi, B.; Qiao, Y.; and Chen, T. 2024. Reg-TTA3D: Better Regression Makes Better Test-Time Adaptive 3D Object Detection. In *European Conference on Computer Vision*, 197–213. Springer.

Yuan, L.; Xie, B.; and Li, S. 2023. Robust test-time adaptation in dynamic scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15922–15932.

Zhang, C.; Zheng, H.; You, X.; Zheng, Y.; and Gu, Y. 2024. Pass: test-time prompting to adapt styles and semantic shapes in medical image segmentation. *IEEE Transactions on Medical Imaging*.

Zhang, J.; Qi, L.; Shi, Y.; and Gao, Y. 2023a. Domainadaptor: A novel approach to test-time adaptation. In *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 18971–18981.

Zhang, X.; Hong, B.-W.; Park, H.; Pak, D. H.; Rickmann, A.-M.; Staib, L. H.; Duncan, J. S.; and Wong, A. 2025. Progressive Test Time Energy Adaptation for Medical Image Segmentation. *arXiv preprint arXiv:2503.16616*.

Zhang, Y.; Wang, X.; Jin, K.; Yuan, K.; Zhang, Z.; Wang, L.; Jin, R.; and Tan, T. 2023b. Adanpc: Exploring non-parametric classifier for test-time adaptation. In *International Conference on Machine Learning*, 41647–41676. PMLR.

Zhao, Z.; Zhou, F.; Xu, K.; Zeng, Z.; Guan, C.; and Zhou, S. K. 2022. LE-UDA: Label-efficient unsupervised domain adaptation for medical image segmentation. *IEEE Transactions on Medical Imaging*, 42(3): 633–646.