DeformTTA: Morphology-Aware Controllable Deformation for Test-Time Adaptation in Fetal Cardiac Ultrasound Image Segmentation

Zhihao Li[1,2,3,4], Zelong Liu[1,2,3,4], Juhua Liu[1,2,3,4], Yongchao Xu[1,2,3,4⋆], and Bo Du[1,2,3,4∗]

[1] School of Computer Science, Wuhan University, Hubei, China
[2] Institute of Artificial Intelligence, Wuhan University, Hubei, China
[3] Hubei Key Laboratory of Multimedia and Network Communication Engineering, Hubei, China
[4] National Engineering Research Center for Multimedia Software, Hubei, China
(dubo@whu.edu.cn)

Abstract. Fetal cardiac segmentation frequently suffers from performance degradation under domain shifts induced by heterogeneous scanners and rapid physiological motion. Standard test-time adaptation (TTA) methods typically rely on generic consistency or entropy minimization, which lack explicit anatomical constraints and fail to preserve biological plausibility under complex cardiac deformations. We propose DeformTTA, a physiology-driven TTA framework that bolsters segmentation robustness by explicitly modeling cardiac dynamics. Unlike domain-agnostic augmentations, DeformTTA leverages a deformation network to simulate realistic fetal heart movements, integrated with a generative model for authentic texture synthesis. Furthermore, a morphology-aware shape prior is introduced to regularize the adaptation process, ensuring the structural integrity of the predicted segmentations. By optimizing for motion-conditioned consistency and morphological alignment, DeformTTA effectively generalizes to target domains during inference. Multi-center experiments demonstrate that DeformTTA significantly outperforms state-of-the-art baselines, particularly in segmenting complex pathological cases. Code is available at https://github.com/SplinterLi/DeformTTA.

Keywords: Fetal Cardiac Segmentation · Test-Time Adaptation · Motion Prior · Morphological Integrity.

1 Introduction

Fetal echocardiography is paramount for the early detection of congenital heart defects (CHDs), a leading cause of infant mortality affecting approximately 1%

⋆ Co-corresponding authors.

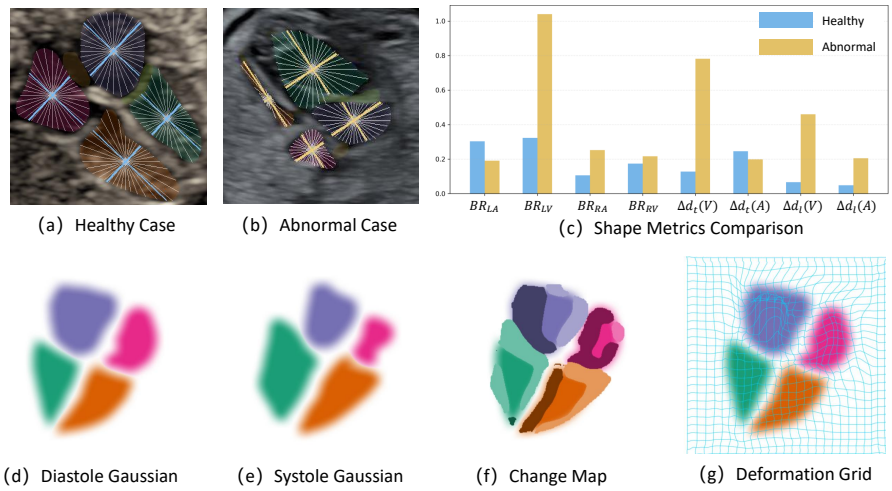

Fig. 1. Motivation of DeformTTA. (a–b) Fetal cardiac anatomy in healthy and pathological cases, the centroid-traversing diameters are used to calculate Boundary Regularity (BR) coefficients, while bold lines indicate transverse (d_t)/longitudinal (d_l) axes. (c) Quantitative morphological metrics, including BR coefficients and RA/LA, RV/LV inter-chamber disparities (e.g., $\Delta d_l(A)$ denotes the difference in longitudinal diameter between RA and LA.) for samples (a–b). (d–e) Diastolic and systolic Gaussian distance maps. (f) Change map of diastolic and systolic capturing inherent motion patterns. (g) Deformation grid modeling the cardiac cycle from diastole to systole.

of live births. Precise anatomical assessment, particularly within the apical four-chamber view (4CV), is essential for evaluating cardiac symmetry and identifying conditions such as ventricular hypoplasia. While deep learning architectures [19,4] have significantly advanced fetal ultrasound segmentation [1,25,11], their clinical utility is often circumscribed by the *i.i.d.* assumption. Consequently, these models frequently exhibit diminished robustness across heterogeneous real-world clinical environments where data distributions shift unpredictably.

In practice, supervised segmentation frameworks [22] encounter pronounced performance fluctuations due to domain shifts stemming from hardware variability and operator-dependent protocols. This challenge is further compounded by high-frequency cardiac motion (Fig. 1 d–g) and structural distortions in pathological cases (Fig. 1 a–c). Although prior research has utilized deformable models and flow-based networks to capture complex spatial transformations [2,8,26], maintaining anatomical consistency under these dynamic conditions remains a non-trivial hurdle. To mitigate these shifts, researchers have explored Domain Generalization (DG) [30,3] and Test-Time Adaptation (TTA) [23,24,17] to implement dynamic update mechanisms during inference.

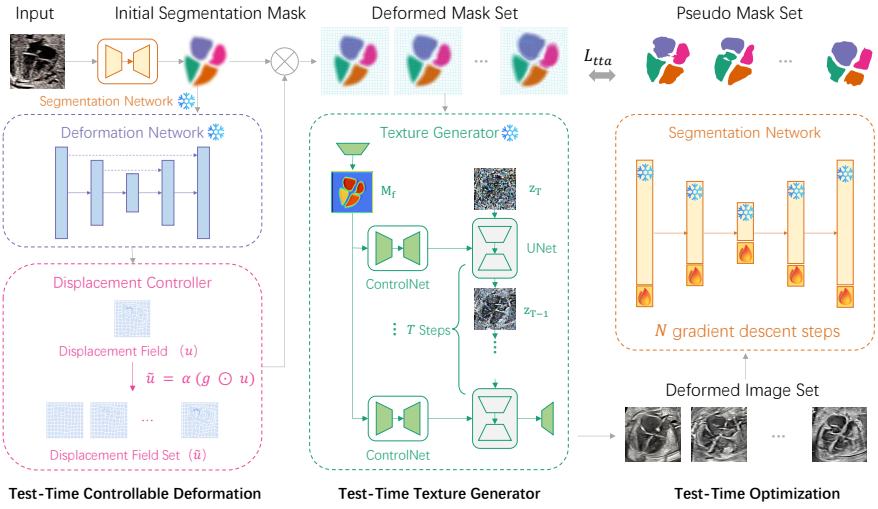

Fig. 2. Overview of DeformTTA. An initial segmentation mask is first predicted from the target ultrasound image, then controllably deformed via a displacement field $\tilde{u} = \alpha(g \odot u)$ to form a deformed mask set. Conditioned on these masks, a ControlNet-based generator synthesizes target-style images, and the resulting synthetic image–mask pairs are used for test-time optimization of the segmentation network.

Despite significant advancements in TTA regarding mechanism design [16,29,27], computational efficiency [20,9,6], reliability [17], and practical deployment [10,12,13], these approaches often face considerable limitations in fetal imaging. Standard TTA objectives often demonstrate a vulnerability to the low-contrast boundaries and high acoustic noise inherent to ultrasound. The structural complexity of the fetal heart poses substantial barriers even for foundation-model-driven approaches [21,18] to provide reliable semantic guidance.

To address these challenges, we propose DeformTTA, a physiology-driven TTA framework that integrates morphology-aware priors with controllable deformations to ensure structural integrity. Our contribution is threefold: First, we introduce a **Boundary Regularity (BR) coefficient** to explicitly quantify anatomical morphology, providing a standardized shape prior to regularize the adaptation process. Second, we develop a **Controllable Morphological Deformation** module operating on structure-centric Gaussian-smoothed fields. By utilizing an anatomical gate, this module simulates realistic cardiac motion while mitigating texture-based overfitting. Third, we employ a ControlNet-based generative model[28,14] to synthesize high-fidelity, motion-conditioned images, facilitating a TTA process that optimizes for cross-view consistency and morphological alignment. Multi-center experiments demonstrate that DeformTTA offers superior generalization compared to state-of-the-art baselines, particularly in complex pathological cases.

2 Methodology

The proposed framework aims to predict a robust segmentation mask $\hat{\mathbf{y}}$ for a test image $\mathbf{x} \in \mathbb{R}^{H \times W}$ under potential domain shift. As illustrated in Fig. 2, the system comprises three primary modules: (i) **Morphology-Aware Shape Prior Construction** (Sec. 2.1), which quantifies anatomical integrity; (ii) **Controllable Morphological Deformation** (Sec. 2.2), which models physiological variations; and (iii) **Physiology-Driven Test-Time Adaptation** (Sec. 2.3), which optimizes the model during inference.

2.1 Morphology-Aware Shape Prior Construction

To move beyond generic shape descriptors, we introduce a *Boundary Regularity* (BR) coefficient to explicitly quantify anatomical morphology. For each class $c \in \mathcal{C}$ with a binary mask $\mathbf{M}_c$, we extract contour points $\{\mathbf{p}_{c,i}\}_{i=1}^{N_c}$ and calculate the centroid $\bar{\mathbf{p}}_c$. The radial distances $r_{c,i}$ and the mean radius $\bar{r}_c$ are defined as:

$$r_{c,i} = \|\mathbf{p}_{c,i} - \bar{\mathbf{p}}_c\|_2, \quad \bar{r}_c = \frac{1}{N_c} \sum_{i=1}^{N_c} r_{c,i}. \tag{1}$$

The BR coefficient, formulated as the radial coefficient of variation, is defined as:

$$\mathrm{BR}_c = \frac{\sqrt{\frac{1}{N_c - 1} \sum_{i=1}^{N_c} (r_{c,i} - \bar{r}_c)^2}}{\bar{r}_c + \epsilon}. \tag{2}$$

As illustrated in Fig. 1(a–c), lower BR_c values signify smooth, canonical boundaries, whereas higher values indicate irregular or pathological morphology. Using source-domain statistics $\{\mu_j, \sigma_j\}_{j=1}^{d}$ for the shape descriptor vector $\mathbf{s}$, we define the standardized deviation $z_j = (s_j - \mu_j)/(\sigma_j + \epsilon)$ and an outlier ratio $\eta = \frac{1}{d} \sum_{j=1}^{d} \mathbb{I}(|z_j| > 3)$. This ratio η serves as a quantitative criterion for monitoring morphological integrity during adaptation.

2.2 Controllable Morphological Deformation

Structure-Centric Representation. To mitigate overfitting to noise-prone ultrasound textures, the deformation network operates on Gaussian-smoothed structural fields. Given a multi-class mask $\mathbf{M}$, we derive one-hot channels $\mathbf{Q}$ and apply a Gaussian kernel $\mathcal{G}_\sigma$ to obtain the representation $\mathbf{z} = \mathcal{G}_\sigma(\mathbf{Q})$, providing stable spatial gradients for optimization.

Controllable Flow Prediction. A deformation network G_θ predicts a dense displacement field $\mathbf{u} = G_\theta(\mathbf{z})$, where $\mathbf{u} : \Omega \to \mathbb{R}^2$. To simulate diverse physiological variations, we introduce a gated field $\tilde{\mathbf{u}} = \alpha(\mathbf{g} \odot \mathbf{u})$, where α denotes deformation strength and $\mathbf{g}$ is an anatomical gate for region-specific perturbation. A Spatial Transformer Network (STN) utilizes $\tilde{\mathbf{u}}$ to generate K deformed conditions $\{\mathbf{c}^{(k)}\}_{k=1}^{K}$.

Registration-Guided Optimization. We supervise G_θ using a warm-start loss against classical B-spline registration: $\mathcal{L}_{\text{init}} = \|\mathbf{u} - \mathbf{u}_{\text{init}}\|_1$. The total deformation loss is:

$$\mathcal{L}_{\text{def}} = \lambda_{\text{sim}}\mathcal{L}_{\text{sim}} + \lambda_{\text{reg}}\mathcal{L}_{\text{smooth}} + \lambda_{\text{init}}(t)\mathcal{L}_{\text{init}}, \tag{3}$$

where $\mathcal{L}_{\text{sim}}$ minimizes the MSE between warped and target distance maps to enforce structural fidelity. $\mathcal{L}_{\text{smooth}}$ incorporates first-order gradients and second-order bending energy to prevent unrealistic folding. The weighting $\lambda_{\text{init}}(t)$ decays over time, enabling a transition from imitation to autonomous adaptation.

2.3 Physiology-Driven Test-Time Adaptation

To address interpolation artifacts in geometric warping, we employ a generative model $\mathcal{G}_\psi$ to synthesize realistic target-domain imagery.

Motion-Conditioned Synthesis. For each deformed condition $\mathbf{c}^{(k)}$, $\mathcal{G}_\psi$ synthesizes an augmented view $\tilde{\mathbf{x}}^{(k)} = \mathcal{G}_\psi(\mathbf{c}^{(k)}, \epsilon^{(k)}, \mathbf{p})$, where $\epsilon^{(k)}$ represents latent noise and $\mathbf{p}$ is a text prompt. The segmentation model F_ϕ then yields cross-view predictions $\hat{\mathbf{y}}^{(k)} = F_\phi(\tilde{\mathbf{x}}^{(k)})$.

Adaptive Optimization. Following standard TTA protocols [23], we perform parameter-efficient updates by only optimizing the Group Normalization (GN) layers. For each test sample, we iteratively refine F_ϕ by minimizing:

$$\mathcal{L}_{\text{tta}} = \beta_1\mathcal{L}_{\text{ent}} + \beta_2\mathcal{L}_{\text{cons}} + \beta_3\mathcal{L}_{\text{seg}}, \tag{4}$$

where $\mathcal{L}_{\text{ent}}$ is the entropy loss, $\mathcal{L}_{\text{cons}}$ enforces consistency across augmented views $\{\tilde{\mathbf{x}}^{(k)}\}$, and $\mathcal{L}_{\text{seg}}$ combines the Dice loss and cross-entropy loss.

3 Experiments

3.1 Experimental Setup

Dataset. We evaluated our framework on a multi-center dataset comprising 650 fetal four-chamber view (4CV) ultrasound scans. Each scan consists of **paired end-diastolic (ED) and end-systolic (ES) frames**, capturing the primary cardiac cycle phases. The cohort includes 200 normal controls from Renmin Hospital of Wuhan University and 450 pathological cases from the Maternal and Child Health Hospital of Hubei Province, the latter covering structural abnormalities such as cardiomegaly and ventricular hypoplasia (e.g., Hypoplastic Left Heart Syndrome, HLHS; and Hypoplastic Right Heart Syndrome, HRHS). To evaluate cross-domain robustness, we benchmarked the models under both **Domain Generalization (DG)** and **Test-Time Adaptation (TTA)** protocols. In all experiments, models are pre-trained on source domain(s) and adapted or evaluated on unseen target domain(s) without access to target-domain annotations.

Table 1. Cross-domain 4CV segmentation results. LA/RA: left/right atrium, LV/RV: left/right ventricle. Values are Dice (%)/HD95(mm).

Method	LA	LV	RA	RV	Mean
w/o TTA	69.08/14.48	68.18/13.00	75.05/9.273	69.22/9.599	70.38/11.61
Training-time DG Methods					
ERM [5]	69.77/13.56	69.70/12.18	75.99/9.409	69.93/9.095	71.35/11.06
MixStyle [30]	70.02/12.77	70.03/11.95	75.84/9.290	69.79/8.842	71.42/10.71
SWAD [3]	70.71/12.50	70.85/11.70	76.24/9.051	70.81/8.665	72.15/10.48
Test-time Adaptation Methods					
TENT [23]	69.10/14.46	68.18/13.00	75.07/9.275	69.25/9.536	70.40/11.57
SAR [17]	69.13/14.47	68.21/12.96	75.13/9.271	69.27/9.533	70.43/11.57
CoTTA [24]	70.90/12.95	72.41/10.20	78.95/6.973	72.71/7.894	73.74/9.505
TEGDA [31]	73.07/7.268	72.22/6.631	79.08/4.652	73.24/6.070	74.40/6.155
Ours	**72.98/4.830**	**75.67/4.710**	**82.45/4.190**	**74.40/5.824**	**76.38/4.888**

Implementation Details. Preliminary benchmarks comparing nnU-Net, TransUnet, and fine-tuned MedSAM [15] with LoRA [7] identified nnU-Net as the most robust architecture for this specific task. Consequently, we adopted nnU-Net as the primary backbone for all subsequent comparisons to ensure a rigorous and standardized evaluation. In our DeformTTA implementation, deformation conditions were derived from Gaussian-smoothed distance fields to provide stable structural priors. A fine-tuned ControlNet was then employed to synthesize target-style images conditioned on these deformed masks, maintaining texture consistency. For the adaptation phase, we utilized $K = 4$ generated samples and $T = 2$ optimization steps per test instance. All experiments were implemented in PyTorch and executed on a single NVIDIA GeForce RTX 3090 GPU (24GB RAM). Performance was quantified using the Dice Similarity Coefficient (DSC) and HD95 as the primary evaluation metric.

3.2 Comparative Experiments

To ensure a rigorous comparison across diverse adaptation regimes, we benchmarked both training-time Domain Generalization (DG) and test-time adaptation (TTA) methods under a unified protocol. All methods utilize the same backbone architecture, initialization, and data augmentation policies. For DG approaches, we maintained a consistent optimization budget on the source domain, while TTA methods were evaluated using identical target-domain stream settings. This setup isolates the performance gains attributable to each method's core mechanism, providing a controlled evaluation of cross-domain robustness.

Analysis of Standard TTA and DG Methods. As reported in Table 1, conventional DG and TTA methods yield varying improvements over the 70.38% Mean Dice baseline. Entropy-driven TTA, such as TENT and SAR, achieves

only 70.43%, suggesting that entropy minimization has limited efficacy against the low-contrast boundaries and acoustic noise of ultrasound. Training-time DG methods like MixStyle and SWAD reach up to 72.15% by diversifying feature statistics; however, they remain constrained by source-domain distributions and overlook the structural variations in pathological cases. More recent frameworks like TEGDA [31] demonstrate improved adaptability by incorporating sophisticated temporal-geometric constraints to preserve anatomical structures during inference. While TEGDA reaches 74.40% Mean Dice and reduces HD95 to 6.155 mm, it still encounters challenges in maintaining precise anatomical topology under the extreme morphological distortions typical of CHD cases.

Performance of DeformTTA. Our proposed **DeformTTA** achieves the highest performance with a Mean Dice of **76.38%**, an improvement of approximately **2.0%** over the strongest recent baseline (TEGDA) and **4.2%** over the best DG method (SWAD). Notably, DeformTTA achieves the lowest HD95 (4.888 mm), suggesting a superior capability in preserving boundary integrity. This performance gain is largely attributable to the modeling of intrinsic cardiac motion characteristics. By explicitly incorporating rhythmic and geometric deformations through the controllable flow network, DeformTTA effectively adapts to target-specific morphologies during the inference phase. Qualitative results in Fig. 3 further indicate that the method is capable of refining incomplete segmentation boundaries and maintaining structural consistency. These observations suggest that a deformation-driven strategy is well-suited for scenarios involving significant anatomical alterations, such as cardiomegaly or hypoplastic left heart syndrome, where traditional statistical or generic geometric adaptation may encounter difficulties.

3.3 Ablation Study

To rigorously evaluate the contribution of each component within our framework, we perform an incremental ablation study as summarized in Table 2.

Efficacy of Learned Deformation and Generative Synthesis. The experimental results demonstrate that the deformation network is the most critical driver of performance; its removal causes the Mean Dice to plummet from 76.38% to 69.42%, indicating that simple stochastic noise cannot capture the complex morphological shifts of the fetal heart. Furthermore, substituting ControlNet with direct warping results in a substantial increase in HD95 (from 4.888 to 11.92 mm), confirming that our generative approach preserves boundary integrity far better than spatial interpolation, which often produces unrealistic artifacts.

Impact of Structural Representation and Geometric Priors. The introduction of distance-Gaussian inputs provides a significant gain of approximately 6% in Dice, as the smoothed representation prevents the model from overfitting to high-frequency ultrasound noise and local texture distractions. Finally, the shape prior and ITK warm-start serve as essential geometric constraints; their inclusion

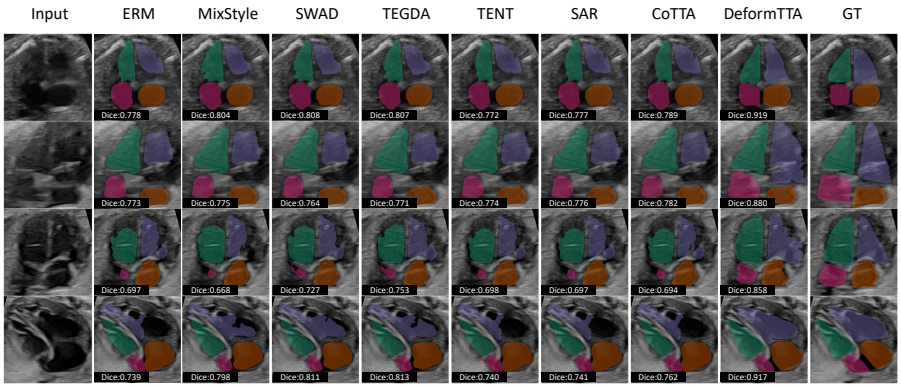

Fig. 3. Qualitative comparison of fetal cardiac ultrasound cases. The first three rows demonstrate cases of cardiomegaly presenting with global cardiac enlargement and structural displacement. The bottom row illustrates hypoplastic left heart syndrome (HLHS), a pathological manifestation characterized by a significantly underdeveloped left ventricle and associated valvular hypoplasia.

Table 2. Ablation study of the key components in DeformTTA. Mean Dice (%) and HD95 (mm) are reported as the evaluation metrics.

Variant	Mean Dice (%) ↑	HD95 (mm) ↓
Full Model	**76.38**	**4.888**
w/o Deformation Net (Stochastic Deformation)	69.42	15.86
w/o Distance-Gaussian Input (Raw Image/Mask)	70.39	14.64
w/o ControlNet (Direct Warp only)	72.31	11.92
w/o Shape Prior Regularization	74.16	7.420
w/o ITK Warm-start	74.98	5.548

reduces boundary outliers and ensures that the model converges to a structurally consistent solution. This is particularly evident in the HD95 metric, where these constraints help maintain the global topology of the four chambers even under significant domain shifts.

4 Conclusion

In this work, we presented DeformTTA, a specialized test-time adaptation framework for fetal cardiac ultrasound. By integrating a learned deformation network with ControlNet-based synthesis, our method effectively bridges the domain gap caused by significant morphological anomalies like cardiomegaly and HLHS. Our results demonstrate that modeling intrinsic cardiac motion and structural priors—via distance-Gaussian fields and shape regularization—surpasses traditional entropy-based TTA, achieving a Mean Dice of 76.38%. This approach

offers a robust solution for deploying high-precision segmentation models across diverse clinical centers with varying scanning protocols and equipment, significantly improving diagnostic reliability.

Despite its performance, our method has certain limitations. The iterative adaptation process introduces a measurable computational overhead; utilizing four augmented samples per inference adds approximately 3–5 seconds to the processing time per case, with slight fluctuations depending on GPU performance. Furthermore, the framework remains partially dependent on the quality of the baseline's initial segmentation. While our shape priors provide critical stabilization, a severely inaccurate initial prediction can limit the effectiveness of the subsequent deformation and adaptation. Future work will focus on optimizing the generative pipeline for real-time efficiency and developing self-correction mechanisms to reduce the dependency on initial "cold-start" results.

Acknowledgement This work was partially supported by the National Key Research and Development Program of China (2023YFC2705701), the National Natural Science Foundation of China under Grant 62222112, and Grant 62176186, the Innovative Research Group Project of Hubei Province under Grant 2024AFA017, and the New Cornerstone Science Foundation through the XPLORER PRIZE. This work was also supported by the WHU-Kingsoft Joint Lab. The numerical calculations in this paper have been performed on the supercomputing system at the Supercomputing Center of Wuhan University.

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

References

1. Arnaout, R., et al.: Deep learning-based computer-aided fetal echocardiography. Sensors **21**, 8007 (2021)
2. Balakrishnan, G., et al.: Voxelmorph: A learning framework for deformable medical image registration. IEEE Transactions on Medical Imaging **38**, 1788–1800 (2019)
3. Cha, J., et al.: Swad: Domain generalization by seeking flat minima. Advances in Neural Information Processing Systems **34**, 22405–22418 (2021)
4. Dosovitskiy, A., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2021)
5. Gulrajani, I., et al.: In search of lost domain generalization. In: ICLR (2021)
6. He, H., et al.: Training-free test-time improvement for explainable medical image classification. In: MICCAI (2025)
7. Hu, E.J., et al.: Lora: Low-rank adaptation of large language models. In: ICLR (2022)
8. Hu, Y., et al.: Weakly-supervised convolutional neural networks for multimodal image registration. In: MICCAI (2018)
9. Joshi, S., et al.: Single image test-time adaptation via multi-view co-training. In: MICCAI (2025)

10. Kim, J., et al.: DC-TTA: divide-and-conquer framework for test-time adaptation of interactive segmentation. arXiv (2025)
11. Liang, B., et al.: Automatic segmentation of 15 critical anatomical labels and measurements of cardiac axis and cardiothoracic ratio in fetal four chambers using nnu-netv2. BMC Medical Informatics and Decision Making **24**, 128 (2024)
12. Liang, Z., et al.: Itermask3d: Unsupervised anomaly detection and segmentation with test-time iterative mask refinement in 3d brain MRI. Medical Image Anal. **107**, 103763 (2026)
13. Liu, J., et al.: Q-PART: quasi-periodic adaptive regression with test-time training for pediatric left ventricular ejection fraction regression. In: CVPR (2025)
14. Liu, M., et al.: Leveraging diffusion models for continual test-time adaptation in fundus image classification. In: MICCAI (2025)
15. Ma, J., et al.: Segment anything in medical images. Nature Communications **15**, 654 (2024)
16. Niu, S., et al.: Efficient test-time model adaptation without forgetting. In: ICML (2022)
17. Niu, S., et al.: Towards stable test-time adaptation in dynamic wild world. In: ICLR (2023)
18. Prabhudesai, et al.: Diffusion-tta: Test-time adaptation of discriminative models via generative feedback. Advances in Neural Information Processing Systems **36**, 17567–17583 (2023)
19. Ronneberger, O., et al.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI (2015)
20. Song, J., et al.: Ecotta: Memory-efficient continual test-time adaptation via self-distilled regularization. In: CVPR (2023)
21. Stock, R., et al.: Segment anything in medical images with nnunet. In: CVPR (2024)
22. Tang, F., et al.: Hi-end-mae: Hierarchical encoder-driven masked autoencoders are stronger vision learners for medical image segmentation. Medical Image Anal. **107**, 103770 (2026)
23. Wang, D., et al.: Tent: Fully test-time adaptation by entropy minimization. In: ICLR (2021)
24. Wang, Q., et al.: Continual test-time domain adaptation. In: CVPR (2022)
25. Wong, K.C., et al.: Multiview and multiclass image segmentation using deep learning in fetal echocardiography. In: Medical Imaging. vol. 11597, pp. 308–313 (2021)
26. Wu, M., et al.: Hmda: A hybrid model with multi-scale deformable attention for medical image segmentation. IEEE Journal of Biomedical and Health Informatics **29**, 1243–1255 (2024)
27. Wu, Z., et al.: Anti-forgetting test-time adaptation for robust medical image analysis under distribution shift. In: MICCAI (2025)
28. Zhang, L., et al.: Adding conditional control to text-to-image diffusion models. In: ICCV (2023)
29. Zhang, Q., et al.: COME: test-time adaption by conservatively minimizing entropy. In: ICLR (2025)
30. Zhou, K., et al.: Domain generalization with mixstyle. In: ICLR (2021)
31. Zhou, Y., et al.: TEGDA: test-time evaluation-guided dynamic adaptation for medical image segmentation. In: MICCAI (2025)