Maximizing Domain Generalization in Automated Fetal Brain Biometry

Yijin Li[1†], Mingxuan Liu[1†], Le Chang[1†], Hongjia Yang[1], Kasidit Anmahapong[1], Xiaotian Hu[1], Ziang Wang[1], Yingqi Hao[1], Haoxiang Li[1], Yifei Chen[1], Xuesheng Li[2], Fenglin Jia[2], Yi Liao[2], Haibo Qu[2], and Qiyuan Tian[1*]

[1] Tsinghua University, Beijing, China
[2] Sichuan University, Chengdu, China
qiyuantian@tsinghua.edu.cn

Abstract. Reliable fetal brain biometry is important for prenatal neurodevelopmental evaluation and disease diagnosis, yet current clinical practice relies largely on manual annotation, which is time-consuming and susceptible to inter-observer variability. While deep learning has been applied to automated biometry, pre-trained models tend to underperform when deployed in unseen clinical environments due to domain shifts. Motivated by the lack of domain adaptation methods specifically tailored to biometry tasks, we propose `BioTTA`, a backbone-agnostic, source-free unsupervised test-time adaptation (TTA) framework that leverages geometric consistency between landmark heatmaps and regressed biometric scalars as a powerful self-supervisory signal for online adaptation. Specifically, `BioTTA` jointly optimizes three complementary self-supervised objectives at test time, namely length consistency loss, entropy minimization loss, and boundary-gradient loss, all further constrained by atlas-guided anatomical priors, to enable robust adaptation without target labels or access to source data. Extensive experiments on multi-center datasets demonstrate that `BioTTA` achieves measurement errors on target domains comparable to or lower than those of supervised fine-tuning, while maintaining expert-level diagnostic performance. Code: `https://github.com/birthlab/BioTTA`.

Keywords: Automated biometry· Test-time adaptation · Fetal brain.

1 Introduction

Fetal MRI has been increasingly used as a complement to US imaging for confirming or ruling out equivocal findings [11,17]. Reliable fetal brain biometry using MRI enables precision assessment of fetal brain maturation and development, supports standardized diagnosis of brain anomalies such as ventriculomegaly (VM), cerebellar hypoplasia, and germinal matrix hemorrhage (GMH), and facilitates gestational age (GA) estimation [9, 20]. However, biometry on clinical

[†] Equal contribution. [*] Corresponding author.

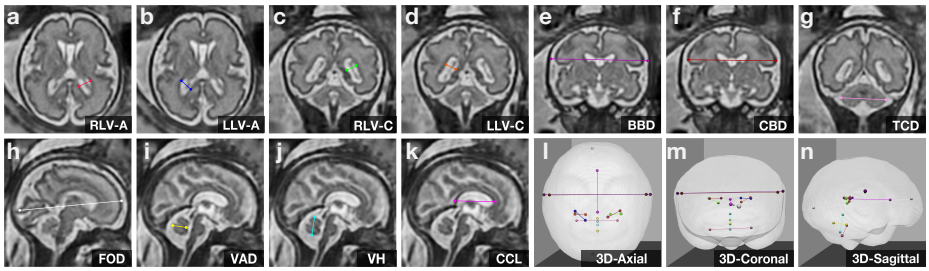

Fig. 1. Anatomical landmarks and biometric measurements used in this study. (a–k) RLV-A, LLV-A, RLV-C, LLV-C, BBD, CBD, TCD, FOD, VAD, VH, and CCL (see Section 3.1 for definitions). (l–n) Axial, coronal, and sagittal views.

2D thick-slice MRI scans is inherently limited by the difficulty of acquiring truly orthogonal reference planes in the presence of unpredictable fetal and maternal motion, the subjectivity of slice selection and anatomical landmark placement, and partial volume effects [12].

Recent studies show that radiologists manually annotate reconstructed 3D MRI volumes generated via slice-to-volume reconstruction (SVR), improving both accuracy and reproducibility [15]. However, manual measurements on 3D volumes remain labor-intensive and prone to inter- and intra-rater variability due to subjective landmark placement. With the advancement of deep learning, automated landmark detection and brain biometry methods have been introduced to solve the problem, which can be categorized into three paradigms: (1) shape computation [2, 20] or direct measurement prediction [25] based on automatic segmentation; (2) localization and measurement via coordinate heatmap regression [1, 7, 12, 13]; and (3) reinforcement learning based on the Markov decision processes [18].

Despite achieving reliable performance comparable to human experts [2], existing brain biometry tools exhibit substantial performance degradation on cross-domain datasets due to variations in data acquisition centers, scanning equipment (e.g., Philips, Siemens, and GE), imaging protocols (e.g., TSE, FSE, HASTE, SSFSE, TrueFISP, FIESTA, and BTFE sequences, as well as field strengths ranging from 0.55T to 3T), and SVR methods (e.g., NiftyMic and NeSVoR), along with the broad span of gestational ages (20 to 38 weeks) and the diversity of brain morphological anomalies [10]. Although domain generalization has advanced fetal brain segmentation [22, 24, 26], it remains largely unexplored in biometry. To bridge this gap, we propose `BioTTA`, a backbone-agnostic, source-free unsupervised test-time adaptation (TTA) framework that operates in two stages: supervised pretraining on a labeled source domain, followed by sample-wise adaptation on unlabeled target data without access to source data. Specifically, three complementary TTA objectives are introduced at test time: (1) a length consistency loss, enabled by a differentiable sparse-weighted landmark detector that permits gradient flow from biometric measurements back through heatmap predictions, to distill global geometric knowledge into the heatmap

branch; (2) an entropy minimization loss that sharpens spatial confidence; and (3) a boundary-gradient loss that anchors landmarks onto anatomically salient edges, all further constrained by atlas-guided anatomical priors.

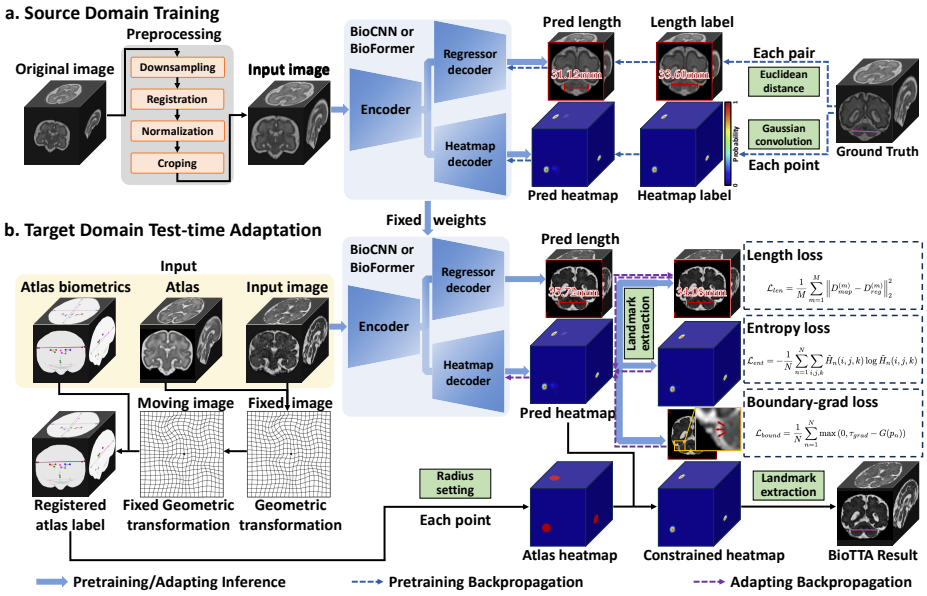

Fig. 2. Overview of `BioTTA` . (a) Source domain training: a flexible backbone (BioCNN or BioFormer) is trained to predict biometric measurements and generate landmark heatmaps. (b) Target domain test-time adaptation: the pretrained backbone is adapted via self-supervised optimization of geometric and uncertainty constraints, incorporating atlas-informed anatomical priors to produce the final landmark heatmaps.

2 Method

2.1 Source Domain Training

On the labeled source domain, a shared encoder is trained in conjunction with two task-specific decoders, one for landmark heatmap prediction and the other for direct biometry regression (Fig. 2a). To evaluate backbone generality, the architecture is instantiated with both a CNN backbone [4] (BioCNN) and a Transformer backbone [23] (BioFormer). The heatmap decoder generates Gaussian probability maps supervised by an MSE loss $\mathcal{L}_{hm}$, while the regression decoder maps deep features to scalar lengths via global pooling, optimized with an L1 loss $\mathcal{L}_{reg}$:

$$\mathcal{L}_{hm} = \frac{1}{N}\sum_{n=1}^{N}\left\|\hat{H}_n - H_n^{gt}\right\|_2^2, \quad \mathcal{L}_{reg} = \frac{1}{M}\sum_{m=1}^{M}\left|\hat{l}_m - l_m^{gt}\right|, \tag{1}$$

where $\hat{H}_n$ and H_n^{gt} denote the predicted and ground truth heatmaps, respectively. $\hat{l}_m$ and l_m^{gt} denote the predicted and ground truth scalar lengths. N denotes the number of landmarks, and M denotes the number of biometric measurements, with $N = 22$ and $M = 11$, as shown in Fig. 1.

2.2 Target Domain Test-Time Adaptation

During inference on unlabeled target data, all convolutional and fully connected layers remain frozen, and only the affine parameters (γ, β) of the batch normalization layers present in the convolutional modules are updated online following the TENT paradigm [21]. Three complementary self-supervised objectives are proposed for TTA (Fig. 2b):

Length Consistency Loss ($\mathcal{L}_{len}$). The regression branch, which integrates global features holistically, is inherently more robust to domain shift than the spatially sensitive heatmap branch. $\mathcal{L}_{len}$ distills its geometric knowledge into the heatmap branch by enforcing consistency between pairwise landmark distances and the regressed scalar measurements. However, the argmax operation for extracting landmark coordinates from heatmaps is non-differentiable, blocking gradient backpropagation from the measurement space to the heatmap predictions. A sparse-weighted differentiable landmark detector is introduced to solve the problem. Specifically, for each predicted heatmap channel $\hat{H}_n (n = 1, \ldots, N)$, let $\mathcal{K}_n$ denote the set of spatial coordinates (i, j, k) of the top k voxels ($k = 50$) with the highest activations. A weighting coefficient for each voxel is computed as the softmax-normalized activation:

$$w_n(i,j,k) = \frac{\exp\left(\hat{H}_n(i,j,k)/\tau\right)}{\sum_{(i',j',k')\in\mathcal{K}_n} \exp\left(\hat{H}_n(i',j',k')/\tau\right)}, \tag{2}$$

where τ=0.05 controls the sharpness of the weighting. The landmark coordinate $p_n \in \mathbb{R}^3$ is computed as a weighted sum:

$$p_n = \sum_{(i,j,k)\in\mathcal{K}_n} w_n(i,j,k) \cdot (i,j,k). \tag{3}$$

The Euclidean distance $D_{map}^{(m)}$ between each pair of landmark endpoints p_{2m-1} and p_{2m} is then aligned with the pseudo label $D_{reg}^{(m)}$ from the regression decoder:

$$\mathcal{L}_{len} = \frac{1}{M} \sum_{m=1}^{M} \left\| D_{map}^{(m)} - D_{reg}^{(m)} \right\|_2^2. \tag{4}$$

Entropy Minimization Loss ($\mathcal{L}_{ent}$). To suppress background noise and drive the model toward sharp, high-confidence peaks at the anatomical landmarks, the Shannon entropy of the spatial probability distribution is minimized. Specifically, for each channel of the predicted heatmap $\hat{H}_n$, softmax normalization along the spatial dimensions yields a probability distribution $\tilde{H}_n$, from

which the entropy loss is computed as:

$$\mathcal{L}_{ent} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{i,j,k}\tilde{H}_n(i,j,k)\log\tilde{H}_n(i,j,k). \tag{5}$$

Boundary-Gradient Loss ($\mathcal{L}_{bound}$). To leverage the prior that biometric landmarks inherently reside on structural edges marked by intensity discontinuities, image gradient magnitude is employed as a self-supervised signal to drive predictions toward high-contrast anatomical boundaries. Specifically, the input image x is first smoothed with the Gaussian kernel ($\sigma = 1.0$) to reduce noise, and the Scharr operator is applied to compute the 3D gradient magnitude map $G = \|\nabla x\|_2$. The Hinge loss is employed to penalize predicted landmarks residing in low-gradient regions:

$$\mathcal{L}_{bound} = \frac{1}{N}\sum_{n=1}^{N}\max\left(0, \tau_{grad} - G(p_n)\right), \tag{6}$$

where $G(p_n)$ is the gradient magnitude sampled at the predicted coordinate p_n via trilinear interpolation, and τ_{grad} is an adaptive gradient threshold (i.e., the median of the gradient map).

Final Loss. $\mathcal{L}_{total} = \lambda_{len}\mathcal{L}_{len} + \lambda_{ent}\mathcal{L}_{ent} + \lambda_{bound}\mathcal{L}_{bound}$. The hyperparameters λ_{len}, λ_{ent} and λ_{bound} were empirically set to balance the loss terms and kept constant across all experiments.

Atlas-guided Anatomical Priors. To enforce anatomical plausibility and mitigate outliers, a gestational age-matched spatio-temporal fetal brain atlas [6] is introduced as a structural prior. Specifically, ANTs (SyNAggro) is utilized to non-linearly register the standard atlas to the target image space. The registered atlas labels are propagated to the target space, yielding an anatomical mask ($10\times 10\times 10$) that constrains the predicted heatmaps, thereby ensuring landmarks are located within valid structures. Final localizations and biometry measurements are subsequently obtained via the differentiable detector.

3 Experiments and Results

3.1 Data Acquisition and Annotation

A total of 259 fetal brain MRI images from five centers were included. Specifically, source domain: WCT dataset (West China Second University Hospital, TSE, 1.5T, SVR with NiftyMIC, n=81). Target domain for quantitative evaluation: FeTA dataset (publicly available [16], FSE, 1.5T or 3T, SVR with [14] or [19], n=50) and LFC dataset (publicly available [7], FSE, 1.5T, SVR with NeSVoR, n=50). Target domain for diagnostic evaluation: FeTA dataset (19 normal, 21 VM) and OoD-GMH dataset (Sichuan Provincial Woman's and Children's Hospital and Chengdu Women's and Children's Central Hospital, SS or TSE, 1.5T or 3T, SVR with NeSVoR, 66 normal, 12 GMH-IVH). All brain

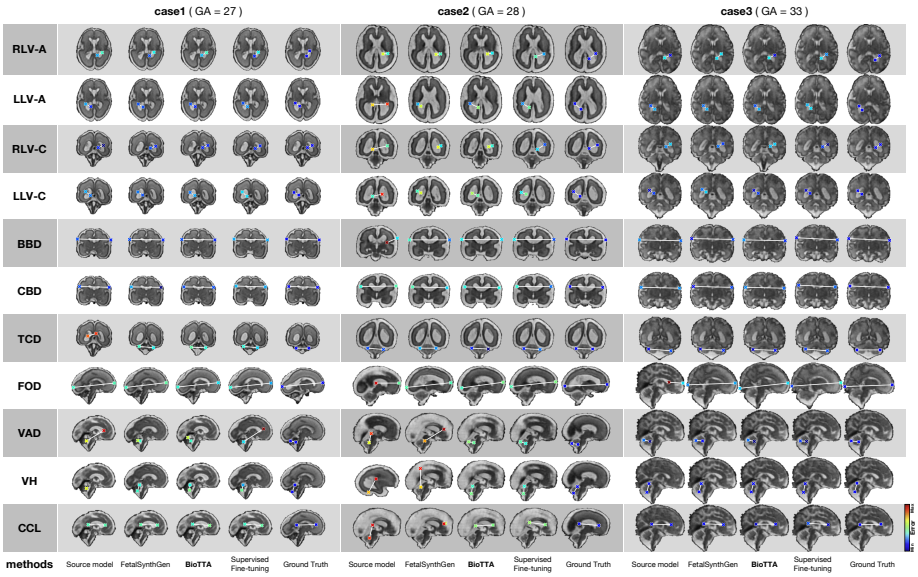

Fig. 3. Comparison of fetal brain biometry on the FeTA dataset. The color of each landmark indicates the Euclidean distance error from the ground truth.

volumes in WCT and FeTA were annotated following a reproducible biometry protocol [5] (Fig. 1): Right Lateral Ventricle-Axial (RLV-A), Left Lateral Ventricle-Axial (LLV-A), Right Lateral Ventricle-Coronal (RLV-C), Left Lateral Ventricle-Coronal (LLV-C), Bone Biparietal Diameter (BBD), Cerebral Biparietal Diameter (CBD), Transverse Cerebellar Diameter (TCD), Fronto Occipital Diameter (FOD), Vermis Anteroposterior Diameter (VAD), Vermis Height (VH), and Corpus Callosum Length (CCL). The LFC dataset was only provided with landmark annotations for TCD, VAD, and VH. The OoD-GMH and FeTA datasets were annotated by an expert, with clinical diagnoses derived according to internationally established criteria. Specifically, VM is diagnosed when atrial width $\geq$10 mm and graded as mild (10–12 mm), moderate (12.1–14.9 mm), or severe ($\geq$15 mm) [20]. GMH is graded into three severity tiers according to the modified Papile classification [3] based on hemorrhage extent and ventricular size. VM diagnosis is therefore evaluated on both datasets.

3.2 Experimental Settings

For quantitative evaluation, `BioTTA` was compared against two target-label-free (a,b) and two target-supervised (c,d) methods: (a) Source Model, which directly applies the pretrained model to quantify domain shift; (b) FetalSynthGen [26], a domain randomization method via synthetic data augmentation; (c) DGMLD [8], a boundary-aware domain generalization method only for cerebellar biometry, fine-tuned with 10 labeled target samples; and (d) Supervised Fine-tuning of the

Table 1. Performance comparison of biometry measurements on two OoD target datasets (FeTA and LFC). Measurement errors are reported as $\text{Mean}_{\pm\text{Std}}$ (in mm). The average row presents the mean error and the relative improvement percentage (in subscript) compared to the source model baseline. PCM@d denotes the percentage of measurements where the prediction error is strictly less than d ($d = 4\,\text{mm}$). Methods are grouped into target-label-free and target-supervised categories. **Bold**: best result among target-label-free methods.

Biometry (mm)	Target-label-free			Target-supervised	
	Source Model	FetalSynthGen [26]	BioTTA (Ours)	DGMLD [8]	Supervised Fine-tuning
Target Domain: FeTA Dataset *Backbone: BioCNN*					
RLV-A	$2.342_{\pm 3.175}$	$2.124_{\pm 2.903}$	$\mathbf{1.665_{\pm 1.724}}$	—	$0.882_{\pm 0.694}$
LLV-A	$2.133_{\pm 2.797}$	$1.749_{\pm 1.527}$	$\mathbf{1.331_{\pm 0.923}}$	—	$1.287_{\pm 0.840}$
RLV-C	$1.909_{\pm 3.087}$	$2.972_{\pm 9.725}$	$\mathbf{1.812_{\pm 2.288}}$	—	$1.086_{\pm 0.940}$
LLV-C	$1.725_{\pm 3.125}$	$1.647_{\pm 1.992}$	$\mathbf{1.632_{\pm 1.364}}$	—	$1.452_{\pm 1.291}$
BBD	$4.165_{\pm 8.673}$	$2.157_{\pm 1.770}$	$\mathbf{1.226_{\pm 1.205}}$	—	$1.445_{\pm 1.224}$
CBD	$2.557_{\pm 4.277}$	$1.456_{\pm 1.399}$	$\mathbf{1.395_{\pm 1.428}}$	—	$1.368_{\pm 1.180}$
TCD	$2.166_{\pm 3.743}$	$\mathbf{1.375_{\pm 1.909}}$	$1.613_{\pm 1.641}$	$1.091_{\pm 1.328}$	$2.295_{\pm 3.857}$
FOD	$6.203_{\pm 13.372}$	$4.678_{\pm 15.227}$	$\mathbf{0.989_{\pm 0.901}}$	—	$1.526_{\pm 1.197}$
VAD	$2.558_{\pm 4.968}$	$\mathbf{2.495_{\pm 4.786}}$	$1.418_{\pm 1.106}$	$2.422_{\pm 1.855}$	$1.938_{\pm 5.271}$
VH	$2.450_{\pm 3.885}$	$3.572_{\pm 6.768}$	$\mathbf{1.309_{\pm 1.099}}$	$1.805_{\pm 1.170}$	$1.345_{\pm 1.444}$
CCL	$2.163_{\pm 4.221}$	$2.956_{\pm 5.541}$	$\mathbf{1.667_{\pm 1.598}}$	—	$1.826_{\pm 1.725}$
Average	2.761	$2.471_{(-10.50\%)}$	$\mathbf{1.460_{(-47.12\%)}}$	$1.773_{(-35.78\%)}$	$1.495_{(-45.85\%)}$
PCM@4mm	87.73%	88.64%	**94.77%**	89.17%	95.68%
Target Domain: FeTA Dataset *Backbone: BioFormer*					
RLV-A	$2.973_{\pm 2.987}$	$2.569_{\pm 2.256}$	$\mathbf{2.475_{\pm 2.558}}$	—	$2.837_{\pm 3.069}$
LLV-A	$2.921_{\pm 4.755}$	$2.972_{\pm 3.377}$	$\mathbf{2.259_{\pm 1.674}}$	—	$3.057_{\pm 4.412}$
RLV-C	$4.092_{\pm 4.320}$	$2.649_{\pm 3.254}$	$\mathbf{2.359_{\pm 2.449}}$	—	$3.751_{\pm 7.593}$
LLV-C	$2.361_{\pm 2.038}$	$2.485_{\pm 2.118}$	$\mathbf{2.106_{\pm 1.852}}$	—	$3.228_{\pm 3.657}$
BBD	$5.741_{\pm 10.570}$	$4.414_{\pm 6.550}$	$\mathbf{2.144_{\pm 1.449}}$	—	$1.408_{\pm 1.699}$
CBD	$5.400_{\pm 6.435}$	$6.711_{\pm 7.573}$	$\mathbf{3.405_{\pm 2.438}}$	—	$2.421_{\pm 2.949}$
TCD	$3.312_{\pm 4.153}$	$4.693_{\pm 5.612}$	$\mathbf{1.435_{\pm 1.428}}$	$2.215_{\pm 1.742}$	$2.904_{\pm 5.733}$
FOD	$11.471_{\pm 16.511}$	$12.670_{\pm 21.232}$	$\mathbf{2.672_{\pm 1.709}}$	—	$2.710_{\pm 4.169}$
VAD	$8.559_{\pm 12.241}$	$\mathbf{2.575_{\pm 3.021}}$	$1.958_{\pm 1.768}$	$1.487_{\pm 1.147}$	$1.822_{\pm 2.015}$
VH	$4.374_{\pm 6.958}$	$3.361_{\pm 6.525}$	$\mathbf{1.936_{\pm 1.385}}$	$1.916_{\pm 1.252}$	$3.697_{\pm 8.406}$
CCL	$4.674_{\pm 5.883}$	$5.542_{\pm 7.505}$	$\mathbf{2.012_{\pm 1.576}}$	—	$4.591_{\pm 5.881}$
Average	5.080	$4.604_{(-9.370\%)}$	$\mathbf{2.251_{(-55.69\%)}}$	$1.873_{(-63.13\%)}$	$2.948_{(-41.97\%)}$
PCM@4mm	67.05%	72.73%	**84.09%**	94.17%	82.05%
Target Domain: LFC Dataset *Backbone: BioCNN*					
TCD	$3.625_{\pm 5.697}$	$1.866_{\pm 2.073}$	$\mathbf{1.405_{\pm 1.534}}$	$10.080_{\pm 6.193}$	$1.386_{\pm 1.514}$
VAD	$1.206_{\pm 1.681}$	$1.676_{\pm 3.058}$	$\mathbf{0.985_{\pm 1.102}}$	$4.001_{\pm 1.913}$	$0.908_{\pm 0.875}$
VH	$1.318_{\pm 1.370}$	$1.432_{\pm 2.038}$	$\mathbf{0.962_{\pm 0.985}}$	$7.701_{\pm 3.798}$	$0.941_{\pm 0.901}$
Average	2.050	$1.658_{(-19.12\%)}$	$\mathbf{1.117_{(-45.51\%)}}$	$7.261_{(+254.2\%)}$	$1.078_{(-47.41\%)}$
PCM@4mm	88.33%	91.67%	**96.67%**	25.83%	98.33%
Target Domain: LFC Dataset *Backbone: BioFormer*					
TCD	$10.945_{\pm 10.363}$	$5.728_{\pm 6.633}$	$\mathbf{2.198_{\pm 2.025}}$	$9.552_{\pm 6.668}$	$4.223_{\pm 6.089}$
VAD	$14.713_{\pm 16.853}$	$6.053_{\pm 12.845}$	$\mathbf{1.503_{\pm 1.200}}$	$4.223_{\pm 2.458}$	$1.328_{\pm 1.408}$
VH	$5.373_{\pm 6.093}$	$4.014_{\pm 4.849}$	$\mathbf{1.897_{\pm 1.359}}$	$5.991_{\pm 3.874}$	$1.970_{\pm 1.610}$
Average	10.344	$5.265_{(-49.10\%)}$	$\mathbf{1.866_{(-81.96\%)}}$	$6.589_{(-36.30\%)}$	$2.507_{(-75.76\%)}$
PCM@4mm	49.17%	72.50%	**91.67%**	34.17%	85.00%

source model with 10 labeled target samples, serving as a practical upper bound. All methods are implemented with both BioCNN and BioFormer backbones for fair comparison. The mean measurement error (in mm) and the Percentage of Correct Measurements (PCM) are reported, with PCM@4 mm adopted as the clinical acceptance criterion following [13]. For diagnostic evaluation, `BioTTA` is benchmarked against three radiologists of increasing seniority using accuracy, F1, and false negative rate (FNR) for disease classification and severity grading. `BioTTA` Training Setup. All experiments ran on an NVIDIA H100 GPU; source training used AdamW (lr=10^{-3}) for 100 epochs, while TTA used lr=5×10^{-4} for 5 epochs to enable rapid distribution alignment without overfitting.

Table 2. Diagnosis and grading performance of `BioTTA` versus clinical experts. Acc = Accuracy; FNR = False Negative Rate. **Bold**: best; Underline: runner-up.

Dataset	OoD-GMH (n = 78)					FeTA (n = 40)				
Tasks	VM Diagnosis			GMH Grading		VM Diagnosis			VM Grading	
Metrics	Acc	F1	FNR ↓	Acc	F1	Acc	F1	FNR ↓	Acc	F1
Medical Student	0.932	0.926	0.444	**0.932**	0.919	0.850	0.864	0.095	0.525	0.547
Junior Radiologist	0.905	0.889	0.667	0.878	0.854	0.875	0.889	**0.048**	0.625	0.649
Senior Radiologist	**0.959**	**0.960**	**0.111**	0.919	0.913	0.925	0.923	0.143	0.675	0.639
`BioTTA`	0.932	0.934	0.222	0.919	**0.922**	**0.950**	**0.952**	**0.048**	**0.725**	**0.701**

Table 3. Ablation study on the FeTA dataset. Average measurement error (mm) and relative improvement over the source model baseline are reported. **Bold**: best result.

Source Model	●	○	●	●	●	●	●	●
Atlas Guide	○	●	●	○	○	○	○	●
$\mathcal{L}_{ent}$	○	○	○	●	○	○	●	●
$\mathcal{L}_{len}$	○	○	○	○	●	○	●	●
$\mathcal{L}_{bound}$	○	○	○	○	○	●	●	●
Average	2.761	2.347	1.995	5.948	1.896	3.016	1.635	**1.460**
	—	−14.99%	−27.74%	+115.4%	−31.33%	+9.240%	−40.78%	**−47.12%**

3.3 Results

(1) Quantitative Evaluation: `BioTTA` consistently reduced measurement errors across both target datasets and backbones (Table 1, Fig. 3). With BioCNN, it achieved average error reductions of 47.12% on FeTA and 45.51% on LFC, with PCM@4 mm reaching 94.77% and 96.67%, closely matching supervised fine-tuning. With BioFormer, where baseline domain shift was more severe, gains were even more pronounced (55.69% and 81.96%), surpassing supervised fine-tuning on both datasets. In contrast, FetalSynthGen [26] yielded only modest improvements, and DGMLD [8], despite using 10 labeled target samples, degraded substantially on LFC when using BioCNN. (2) Diagnostic Evaluation:

BioTTA achieved diagnostic performance comparable to or exceeding that of the senior radiologist (Table 2). Specifically, it achieved the highest F1 on the grading task of OoD-GMH and attained the best performance across all metrics for both VM diagnosis and grading on the FeTA dataset. (3) Ablation Study (BioCNN backbone): The complete BioTTA framework achieved the lowest average error of 1.460 mm, a 47.12% reduction over the source baseline (Table 3).Notably, applying entropy minimization alone, which is equivalent to TENT [21], increased the average error to 5.948 mm (+115.4%), suggesting that entropy minimization in isolation is insufficient and can be detrimental for biometry adaptation without geometric and boundary constraints.

4 Conclusion

We presented BioTTA, a source-free and backbone-agnostic test-time adaptation framework for automated fetal brain biometry that requires neither target labels nor access to source data. Extensive experiments on multi-center datasets demonstrated that BioTTA achieved accuracy comparable to or exceeding supervised fine-tuning, while providing diagnostic reliability comparable to that of senior radiologists on out-of-distribution target domain data.

Acknowledgments. This work was supported by the National Natural Science Foundation of China (grant number 82572198), the Natural Science Foundation of Sichuan Province (grant number 2025ZNSFSC1768), the Scientific Research Project of Sichuan Medical Association (grant number 2024HR130), the Science and Technology Department of Sichuan Province, China (grant number 25SYSX0255), the Tsinghua University Startup Fund, and the Tsinghua University Dushi Program (grant numbers 20241080026, 20251080056).

Disclosure of Interests. The authors have no competing interests to declare.

References

1. Avisdris, N., Ben Bashat, D., Ben-Sira, L., Joskowicz, L.: Fetal brain MRI measurements using a deep learning landmark network with reliability estimation. In: Uncertainty for Safe Utilization of Machine Learning in Medical Imaging and Perinatal Imaging, Placental and Preterm Image Analysis (UNSURE and PIPPI). pp. 210–220. Springer (2021). https://doi.org/10.1007/978-3-030-87735-4_20
2. Avisdris, N., Yehuda, B., Ben-Zvi, O., Link-Sourani, D., Ben-Sira, L., Miller, E., Zharkov, E., Ben Bashat, D., Joskowicz, L.: Automatic linear measurements of the fetal brain on MRI with deep neural networks. International Journal of Computer Assisted Radiology and Surgery **16**(9), 1481–1492 (2021). https://doi.org/10.1007/s11548-021-02436-8
3. Burstein, J., Papile, L.A., Burstein, R.: Intraventricular hemorrhage and hydrocephalus in premature newborns: A prospective study with CT. American Journal of Roentgenology **132**(4), 631–635 (1979). https://doi.org/10.2214/ajr.132.4.631

4. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3D U-Net: Learning dense volumetric segmentation from sparse annotation. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI). pp. 424–432. Springer (2016). https://doi.org/10.1007/978-3-319-46723-8_49
5. Garel, C.: MRI of the Fetal Brain: Normal Development and Cerebral Malformations. Springer (2004). https://doi.org/10.1007/978-3-642-18747-6
6. Gholipour, A., Rollins, C.K., Velasco-Annis, C., Ouaalam, A., Akhondi-Asl, A., et al.: A normative spatiotemporal MRI atlas of the fetal brain for automatic segmentation and analysis of early brain growth. Scientific Reports **7**(1), 476 (2017). https://doi.org/10.1038/s41598-017-00525-w
7. Gong, H., Liu, H., Wang, Y., Liu, X., Wan, X., Shi, Q., Li, H.: Fetal cerebellum landmark detection based on 3D MRI: Method and benchmark. IEEE Journal of Biomedical and Health Informatics **29**(8), 5712–5721 (2025). https://doi.org/10.1109/JBHI.2025.3559702
8. Gong, H., Lu, Y., Wan, X., Li, H.: Domain generalized medical landmark detection via robust boundary-aware pre-training. In: AAAI Conference on Artificial Intelligence (AAAI). vol. 39, pp. 3140–3148 (2025). https://doi.org/10.1609/aaai.v39i3.32323
9. Kyriakopoulou, V., Vatansever, D., Davidson, A., Patkee, P., Elkommos, S., Chew, A., et al.: Normative biometry of the fetal brain using magnetic resonance imaging. Brain Structure and Function **222**(5), 2295–2307 (2017). https://doi.org/10.1007/s00429-016-1342-6
10. Li, Y., Liu, M., Yang, H., Li, H., Bai, X., Liao, Y., Qu, H., Tian, Q.: Anatomy-guided test-time adaptation for automated fetal brain MRI morphometry. In: Medical Imaging with Deep Learning (MIDL) Short Papers (2025), https://openreview.net/forum?id=iLBipDelQu
11. Liu, M., Liao, Y., Li, H., Zhu, J., Yang, H., Hao, Y., Qu, H., Tian, Q.: Quality-label-free fetal brain MRI quality control based on image orientation recognition uncertainty. Medical Image Analysis p. 103994 (2026). https://doi.org/10.1016/j.media.2026.103994
12. Luis, A., Uus, A., Matthew, J., et al.: Towards automated fetal brain biometry reporting for 3-dimensional T2-weighted 0.55-3T magnetic resonance imaging at 20-40 weeks gestational age range. Pediatric Radiology **56**, 366–383 (2026). https://doi.org/10.1007/s00247-025-06403-2
13. Masterl, E., Parkelj, A., Vesnaver, T.V., Špiclin, Ž.: Enhancing prenatal diagnosis: Automated fetal brain MRI morphometry. In: Perinatal, Preterm and Paediatric Image Analysis (PIPPI). pp. 105–116. Springer (2024). https://doi.org/10.1007/978-3-031-73260-7_10
14. Murgasova, M., Quaghebeur, G., Rutherford, M.A., Hajnal, J.V., Schnabel, J.A.: Reconstruction of fetal brain MRI with intensity matching and complete outlier removal. Medical Image Analysis **16**(8), 1550–1564 (2012). https://doi.org/10.1016/j.media.2012.07.004
15. Nagaraj, U.D., Greer, J.S., Debnath, P., Tkach, J.A., Kline-Fath, B.M.: Slice-to-volume reconstruction of fetal brain MR imaging in clinical practice. American Journal of Neuroradiology **46**(8), 1726–1730 (2025). https://doi.org/10.3174/ajnr.A8728
16. Payette, K., de Dumast, P., Kebiri, H., Ezhov, I., Paetzold, J.C., Shit, S., et al.: An automatic multi-tissue human fetal brain segmentation benchmark using the fetal tissue annotation dataset. Scientific Data **8**(1), 167 (2021). https://doi.org/10.1038/s41597-021-00946-3

17. Sanchez, T., Mihailov, A., Koob, M., Girard, N., Manchon, A., Valenzuela, I., et al.: Biometry and volumetry in multi-centric fetal brain magnetic resonance imaging: Assessing the bias of super-resolution reconstruction. Pediatric Radiology **55**(10), 2064–2075 (2025). https://doi.org/10.1007/s00247-025-06347-7
18. Sharvani, B.S., JayaLakshmi, G., Bhargavi, K.: Enhancing anatomical landmark detection in brain images through multiagent deep reinforcement learning. In: 2024 First International Conference on Electronics, Communication and Signal Processing (ICECSP). pp. 1–6 (2024). https://doi.org/10.1109/ICECSP61809.2024.10698165
19. Tourbier, S., Bresson, X., Hagmann, P., Thiran, J.P., Meuli, R., Bach Cuadra, M.: An efficient total variation algorithm for super-resolution in fetal brain MRI with adaptive regularization. NeuroImage **118**, 584–597 (2015). https://doi.org/10.1016/j.neuroimage.2015.06.018
20. Vahedifard, F., Ai, H.A., Supanich, M.P., Gauthier, S.A., Spincemaille, P., Wang, Y., Zhang, S.: Automatic ventriculomegaly detection in fetal brain MRI: A step-by-step deep learning model for novel 2D-3D linear measurements. Diagnostics **13**(14), 2355 (2023). https://doi.org/10.3390/diagnostics13142355
21. Wang, D., Shelhamer, E., Liu, S., Olshausen, B., Darrell, T.: Tent: Fully test-time adaptation by entropy minimization. In: International Conference on Learning Representations (ICLR) (2021), https://openreview.net/forum?id=uXl3bZLkr3c
22. Wang, Y., Chen, Y., Jiang, S., Yu, W., Liu, M., et al.: DR-TTA: Dynamic and robust test-time adaptation under low-quality mri conditions for brain tumor segmentation. In: 2025 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). pp. 2899–2906 (2025). https://doi.org/10.1109/BIBM66473.2025.11356381
23. Xu, Y., Zhang, J., Zhang, Q., Tao, D.: ViTPose: Simple vision transformer baselines for human pose estimation. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 35, pp. 38571–38584 (2022)
24. Yu, W., Jiang, S., Chen, Y., Chang, S., Wang, Y., Wu, B., Dong, J., Liu, M., et al.: A large scale benchmark for test time adaptation methods in medical image segmentation. arXiv preprint arXiv:2512.02497 (2025). https://doi.org/10.48550/arXiv.2512.02497
25. Zalevskyi, V., Sanchez, T., Kaandorp, M., Roulet, M., Fajardo-Rojas, D., et al.: Advances in automated fetal brain MRI segmentation and biometry: Insights from the FeTA 2024 challenge. Medical Image Analysis **109**, 103941 (2026). https://doi.org/10.1016/j.media.2026.103941
26. Zalevskyi, V., Sanchez, T., Roulet, M., Aviles Verdera, J., Hutter, J., Kebiri, H., Bach Cuadra, M.: Improving cross-domain brain tissue segmentation in fetal MRI with synthetic data. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI). pp. 437–447. Springer (2024). https://doi.org/10.1007/978-3-031-72378-0_41