Automated Fetal Head Circumference Measurement via Multi-Task Deep Learning

Nguyen The Khai - 23BI14205 - DS

I. Introduction

Precise fetal head circumference (HC) quantification from ultrasound imagery remains crucial for obstetric assessment. We present a multi-task architecture combining semantic segmentation with direct regression for automated HC measurement on the HC18 challenge dataset. Our approach simultaneously learns anatomical boundary delineation and metric prediction through shared encoder representations.

II. Methodology

A. Data Preprocessing

Initial dataset examination revealed several quality concerns requiring systematic filtration. Among 999 training instances, we identified empty annotations, undersized masks (< 100 pixels), and samples exhibiting HC calculation anomalies. Samples demonstrating $> 50\%$ discrepancy between computed and ground-truth measurements were excluded. This curation process yielded a refined corpus suitable for robust model development.

Figure 1 illustrates the error distribution characteristics post-cleaning, revealing a concentrated distribution with mean absolute deviation of 0.71mm and median of 0.66mm.

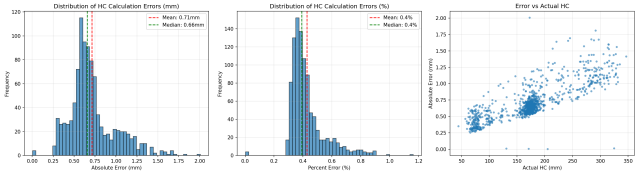

Fig. 1. Error distribution analysis across validation samples showing concentrated performance around ground truth values.

B. Network Architecture

We deployed a multi-task U-Net configuration leveraging ResNet-34 as the encoder backbone, pre-initialized with ImageNet weights. The architecture bifurcates into dual prediction branches: (i) a segmentation decoder generating pixel-wise probability maps through sigmoid activation, and (ii) a regression head employing global average pooling followed by fully-connected layers for direct HC estimation. This design enables simultaneous learning of spatial and metric representations.

C. Training Protocol

We implemented 5-fold cross-validation with stratified partitioning to ensure generalization assessment. Training employed combined loss formulation $\mathcal{L} = \mathcal{L}_{\text{seg}} + 0.1\mathcal{L}_{\text{reg}}$, where segmentation loss integrates Dice coefficient and binary cross-entropy, while regression utilizes mean absolute error. The AdamW optimizer with initial learning rate 10^{-4} and ReduceLROnPlateau scheduling managed optimization. Input dimensions were standardized to 256×256 pixels, with augmentation strategies including horizontal flipping, affine transformations, and elastic deformations.

III. Results

A. Quantitative Performance

Cross-validation evaluation yielded consistent performance across folds, as detailed in Table I. The ensemble achieved mean MAE of 12.13 ± 0.81mm, demonstrating stable generalization. Fold-wise variance remained minimal, indicating robust architecture design independent of data partitioning.

TABLE I
Cross-validation performance metrics across five folds

Fold	MAE (mm)	Rel. Performance
1	12.49	+2.9%
2	11.12	-8.3%
3	11.88	-2.0%
4	13.49	+11.2%
5	11.66	-3.9%
Mean $\pm$ Std	**12.13 $\pm$ 0.81**	–

Figure 2 demonstrates strong correlation between predictions and ground truth annotations, with prediction clusters tightly distributed along the identity diagonal. The residual analysis reveals relatively uniform error distribution across the HC measurement spectrum, though slight heteroscedasticity appears at extreme values.

B. Qualitative Assessment

Visual inspection of segmentation outputs across validation folds reveals varying degrees of anatomical boundary adherence. Figure 3 presents representative examples from each cross-validation partition. Fold 1 demonstrates precise elliptical contour extraction with 5.8mm regression error, while Folds 2-4 exhibit fragmentary segmentation despite reasonable regression performance. Fold 5 maintains geometric consistency with moderate accuracy.

High-quality samples from Figure 4 demonstrate optimal performance scenarios where both segmentation and regression converge successfully, achieving sub-millimeter precision.

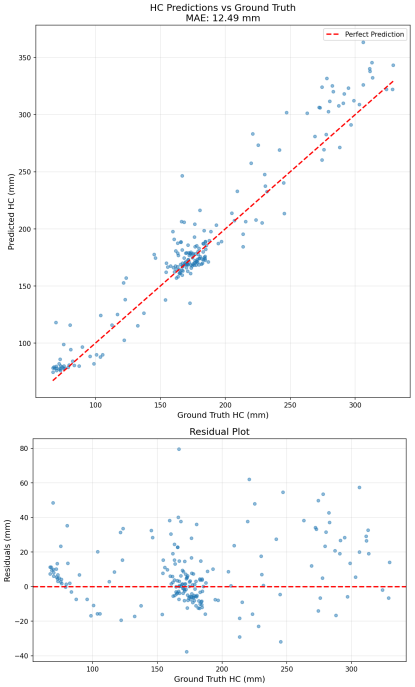

Fig. 2. Up: Predictions versus ground truth with identity line. Down: Residual distribution showing homoscedastic error characteristics.

IV. Discussion

The multi-task formulation provides complementary supervision signals, though our results suggest task asymmetry. The regression pathway demonstrates greater stability than segmentation, possibly due to reduced sensitivity to minor boundary perturbations. This finding aligns with clinical practice, where HC measurement tolerates slight annotation variance.

Performance heterogeneity across folds likely stems from acoustic shadowing, fetal positioning variations, and operator-dependent acquisition quality. Future work should investigate attention mechanisms to emphasize diagnostically relevant regions and uncertainty quantification for clinical deployment confidence estimation.

V. Conclusion

We demonstrated effective automated fetal HC measurement through multi-task deep learning, achieving 12.13 ± 0.81mm mean absolute error across five-fold validation. The architecture successfully balances segmentation accuracy with direct metric prediction, offering a viable approach for obstetric ultrasound quantification. Continued refinement targeting segmentation consistency and edge case handling will advance clinical applicability.

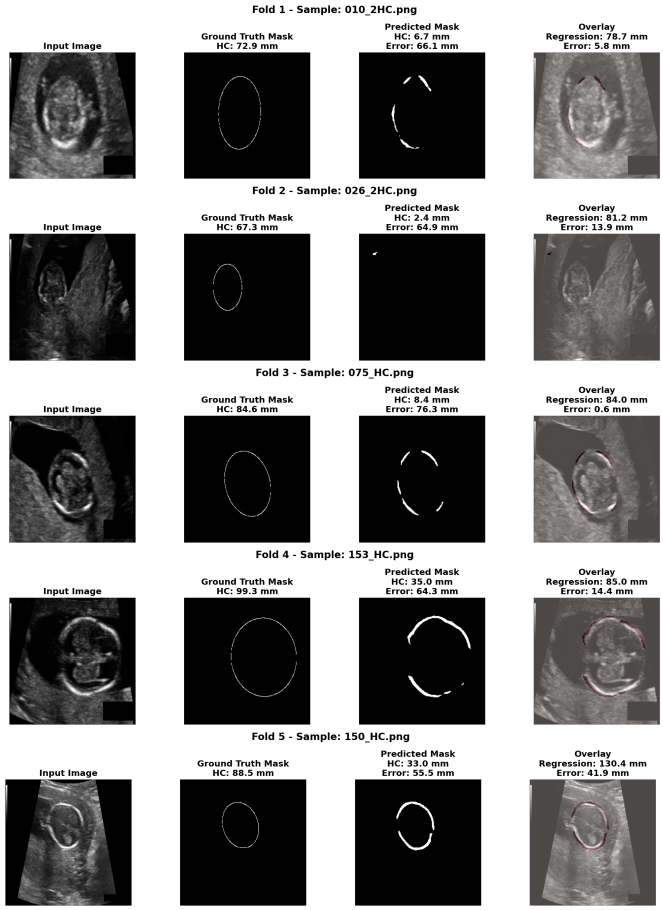

Fig. 3. Representative predictions across five validation folds demonstrating segmentation quality variance.

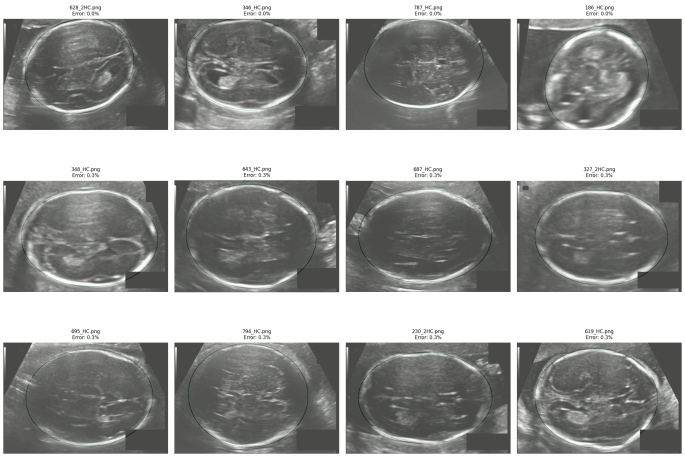

Fig. 4. Exemplary predictions demonstrating near-perfect anatomical delineation and metric estimation.