Quality over quantity: data curation for automated angle of progression estimation in intrapartum ultrasound

Mareen Kalis[1,2], Nora Andelfinger[4], Christian Baumgartner[3], and Lisa Koch[1,2]

[1] Department of Diabetes, Endocrinology, Nutritional Medicine and Metabolism (UDEM), Inselspital, Bern University Hospital, University of Bern, Bern, Switzerland
[2] Department of Digital Medicine, University of Bern, Bern, Switzerland
[3] Faculty of Health Sciences and Medicine, University of Lucerne, Lucerne, Switzerland
[4] Department of Obstetrics and Gynecology, Haga Teaching Hospital, The Hague, Netherlands

Abstract. Assessment of fetal head descent is central to intrapartum labour management, but the clinical standard of digital vaginal examination is subjective and examiner-dependent. Transperineal ultrasound enables objective measurement of the angle of progression (AoP), yet manual measurement and automated estimation remain sensitive to image quality, anatomical visibility, and annotation consistency. This study investigates whether dataset curation improves automated AoP estimation. Segmented frames in an intrapartum ultrasound video dataset were quality-classified and their pubic symphysis and fetal head masks manually refined. Segmentation models trained on curated and original masks were compared via downstream AoP, calculated geometrically and referenced against consensus landmark annotations from two experts. Training on a small curated dataset of 356 frames led to a median AoP error of 4.59°, outperforming models trained on an order of magnitude more non-curated frames, and almost matching the inter-observer variability of 4.09°. These findings suggest that annotation quality, anatomical consistency, and frame-quality assessment are important for robust automated AoP estimation. We release the curated subset, multi-expert AoP reference, and code as a reproducible benchmark.[5].

Keywords: intrapartum ultrasound, data quality, angle of progression

1 Introduction

Labour progress is monitored by assessing fetal head descent, typically documented as fetal head station on a clinical scale from -4 to +4 cm relative to the interspinous plane, which guides labour management and decisions about

[5] Data repository: https://doi.org/10.5281/zenodo.21189540; Code repository: https://github.com/mlm-lab-research/thesis_intrapartum_ultrasound.

operative vaginal delivery. In practice this relies on digital vaginal examination, but the reference standard is subjective and examiner-dependent: in a birth-simulator study, [5] reported fetal head station errors in 50–88% of assessments by residents and 36–80% by attending physicians, indicating that incorrect station assessment occurred more frequently than correct assessment, and [4] found agreement on station level between digital examination and transperineal ultrasound in only 31% of cases.

Transperineal ultrasound (Fig. 1 a) has been proposed as a more objective alternative to digital vaginal examination. The angle of progression (AoP) – the angle between the long axis of the pubic symphysis and the line from its inferior margin tangent to the fetal skull (Fig. 1 b,c) – is a reproducible measure of head descent that is associated with mode of delivery [2], and low AoP is predictive of failed vacuum extraction [3]. Manual measurement, however, requires expertise in both image acquisition and geometric assessment, motivating automated, AI-assisted analysis.

Automatic AoP estimation typically chains standard-plane recognition, segmentation of the pubic symphysis and fetal head, landmark detection, and geometric post-processing [10]. Image-based resources such as the PSFHS Challenge [1] advanced the segmentation step, and the Maternal-Fetal Ultrasound Video (MFV) Dataset [11] moved toward end-to-end, video-based biometry that better reflects the dynamic nature of intrapartum acquisition. Reliable AoP estimation nonetheless depends critically on annotation quality, and consistent, high-quality standard-plane and segmentation labels remain scarce.

In this work we: (1) provide high-quality subset annotations for the video dataset of Niu et al. [11], including two-expert AoP annotations and refined segmentation masks with frame-quality classifications; (2) train segmentation models to examine how training data curation affects downstream AoP calculation; and (3) release the annotations and code as a reproducible benchmark for this clinically important problem.

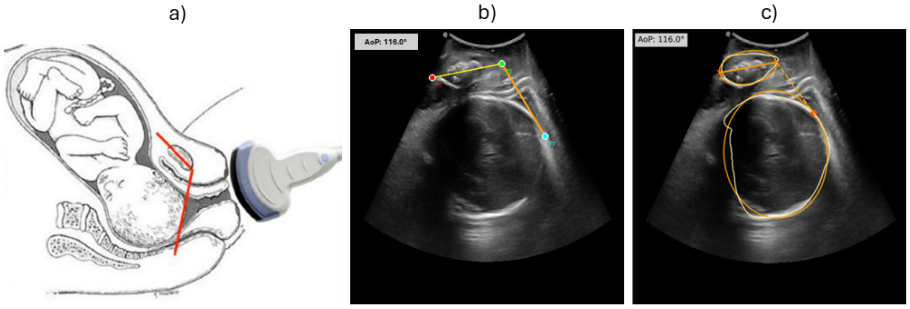

Fig. 1: Angle of progression measurement and annotation examples. (a) ISUOG guideline illustration of AoP measurement, reproduced from Ghi et al. [6] with permission from John Wiley and Sons. (b) Representative ultrasound frame with human reference annotation. (c) Representative ultrasound frame with predicted segmentation masks and algorithmic AoP estimation.

2 Data

2.1 Original dataset

Our work is based on the MFV-Dataset provided by [11]. The dataset was collected retrospectively from 774 pregnant women with singleton pregnancies at or beyond 37 weeks of gestation and fetuses in cephalic presentation. Data were acquired at three medical institutions in China using different ultrasound systems. The original MFV dataset comprises 774 videos, divided into 434 training, 40 validation, and 300 test videos. For the present study, frames from the training split were curated and model performance was evaluated on the test split; the original validation split was not used. In the training split, 266 videos were labelled as "positive", i.e. containing a standard plane. On these, the pubic symphysis and fetal head were segmented on every 10th frame irrespective of the quality and plane of the individual frames. In the original MFV test split, one frame from each of the 300 videos had been selected and manually segmented. We used exactly these 300 preselected frames for the present evaluation; no additional frame selection was performed.

2.2 Data curation and manual re-annotation

Building on the dataset of Niu et al. [11], we investigated whether task-specific data curation improves segmentation-based AoP estimation. In particular, we tested whether refined anatomical segmentation masks and exclusion of frames with inadequate visibility or quality for AoP calculation enhance downstream model performance. We therefore curated and refined a subset of the original training and test data, including frame-quality classification, manual review and editing of segmentation masks, and selection of frames for model training and

evaluation. Frame curation and segmentation-mask refinement were performed using a custom Python application based on OpenCV. The application supported frame review, frame-quality classification, brush-based editing of segmentation masks, and mask export [9].

Frame quality assessment During data curation and mask refinement, the first author, a resident physician with two years of experience in obstetrics, classified all screened segmented frames as `Standard Plane`, `Acceptable`, or `Inadequate` according to the following criteria:

a. Standard-plane: The symphysis contours are clearly visible in full, and the fetal skull below the symphysis, in probe-wise direction and in the tangent-point region is clearly visible.
b. Acceptable: The symphysis contours are mostly clearly visible, meaning more than two-thirds are visible, and the fetal skull in the tangent-point area of the symphysis is clearly visible.
c. Inadequate: Criteria a and b are not met because of insufficient image quality, incomplete visualization of the relevant anatomical structures, or pronounced artefacts.

For the test set, these frame-quality categories assigned by the first author during mask refinement were used for all category-stratified analyses reported in this study. According to this classification, the 300 test frames comprised 51 `Standard Plane`, 196 `Acceptable`, and 53 `Inadequate` frames.

Manual segmentation refinement We next manually refined the segmentation masks according to the following criteria:

a. Previous fetal head and pubic symphysis boundaries were manually deleted and repainted where necessary using a mouse-brush tool, focusing on clear and consistent boundary selection. For the fetal head, only the area surrounded by visible skull contour distinguishable from the background or acoustic shadow was annotated; the distal ventral and caudal ends of the visible skull were connected directly rather than completed by fitting a circular shape (Fig. 2).
b. Post-processing: Connected-component filtering was applied by retaining only the largest connected component per class and removing small isolated components. Small internal holes within the remaining connected regions were filled. Mild boundary smoothing was applied using binary closing followed by binary opening with disk-shaped structuring elements.

As many of the training frames were highly redundant and manual segmentation was time consuming, we selected up to three visually suitable frames (`Standard Plane` or `Acceptable`) per video for refinement. In the test set, the single frame previously selected and segmented for each video in the original MFV dataset was manually refined. All 300 test frames were retained, including those classified as `Inadequate`, to enable evaluation across the full range of frame quality.

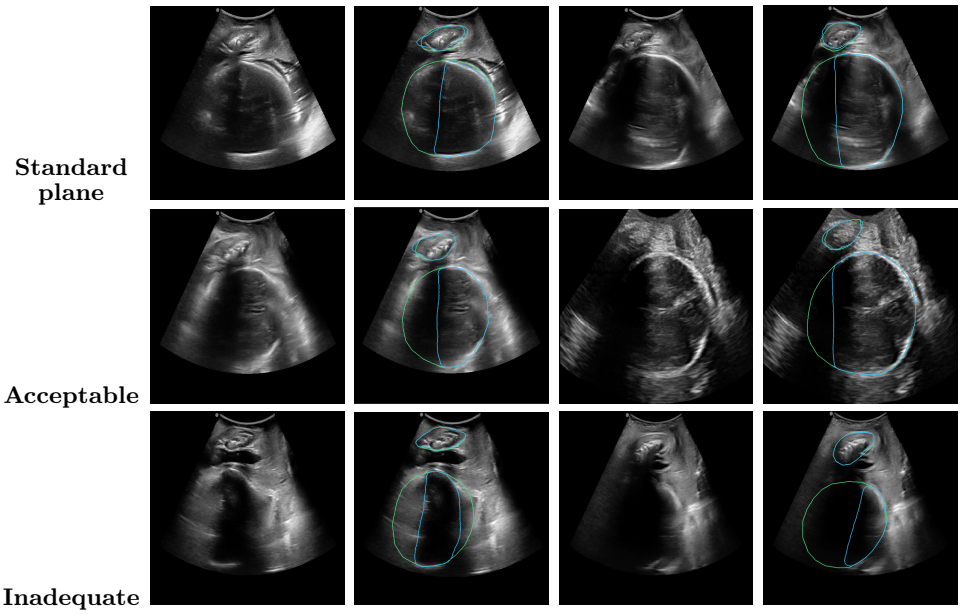

Fig. 2: Two example ultrasound original image frames shown next to corresponding mask overlay per frame-quality category. Rows show standard-plane, acceptable, and inadequate frames. Refined segmentation boundaries are shown in blue, original segmentation boundaries are shown in green.

Curated segmentation dataset In the training split, 2,561 segmented frames from 266 ultrasound videos were screened. The final curated training dataset comprised 356 `Standard Plane` or `Acceptable` frames from 137 videos, with a maximum of three frames per video. The test set comprised the fixed set of 300 frames provided with the original MFV test split, with one manually segmented frame per video. These frames were manually refined and evaluated irrespective of the frame-quality category assigned by the first author.

Multi-expert landmark annotation as a gold-standard AoP reference In the original dataset, AoP reference values were calculated by fitting ellipses to the pubic symphysis and fetal head segmentations, algorithmically identifying landmarks and calculating the angle of progression [11]. To additionally provide gold-standard human reference AoP measurements, we developed a custom annotation tool [9]. For each measurable frame, the annotators marked the two pubic-symphysis endpoints and the fetal-head tangency point.

Two resident physicians, each with two years of experience in obstetrics, independently annotated the AoP on the test frames by drawing two connected lines, following the measurement principle commonly used in clinical ultrasound software (Fig. 1 b). Within the annotation tool, each annotator independently assigned a frame-quality category and only annotated landmarks on frames they classified as at least `Acceptable`. These two reader-specific classifications were separate from the frame-quality classification assigned by the first author during mask refinement. They were used only to determine whether the respective annotator provided landmark coordinates and were not used for the frame-quality-stratified model evaluation. The final consensus AoP was calculated from the mean landmark positions of both annotators. If only one annotator accepted a frame, the AoP was calculated from a single landmark annotation. A gold-standard consensus AoP was available for 278 of 300 frames (92.67%), which at least one annotator considered of sufficient quality for reliable AoP measurement (Table 1). Both the consensus AoP as well as the individual annotator data are publicly available through Zenodo [8].

Table 1: Availability of gold-standard AoP reference from $n = 300$ test frames.

Method	Available	Missing
Annotator 1	271 (90.33%)	29 (9.67%)
Annotator 2	269 (89.67%)	31 (10.33%)
Gold-standard AoP reference	278 (92.67%)	22 (7.33%)

3 End-to-end pipeline for AoP estimation and validation

Here, we describe the end-to-end pipeline used to estimate the angle of progression from intrapartum ultrasound images. First, a segmentation model is

trained on training frames to delineate the pubic symphysis and fetal head. The trained model is then applied to segment the test frames. From the resulting predicted segmentation, anatomical landmarks are automatically identified, and the angle of progression is calculated geometrically from these landmarks. The following subsections describe the segmentation model training and the landmark-derivation and AoP-calculation strategy in detail.

3.1 Segmentation model training

We used the nnU-Net (v2) [7] as the segmentation framework for all experiments, using the standard 2D nnU-Net configuration ("2d") with default "nnUNetTrainer". The default training setup included 1000 epochs, a combined Dice and cross-entropy loss, stochastic gradient descent optimization, a polynomial learning-rate schedule, and standard nnU-Net data augmentation. Segmentation masks comprised two foreground classes, pubic symphysis and fetal head, plus background. To evaluate the influence of dataset curation and mask refinement, three nnU-Net models were trained.

Curated refined-mask model (nnUNet-Proposed): The model was trained on selected frames with manually refined segmentation masks. This model represents the curated training setting, combining frame selection with manual mask correction.

Curated original-mask model (nnUNet-OrigSmall): The model was trained on the same curated subset of frames, but using the original segmentation masks provided with the dataset. This allowed the effect of manual mask refinement to be evaluated independently of frame selection.

Full original-mask model (nnUNet-OrigLarge): The model was trained on all available segmented frames using the original masks. This model was used to assess the effect of training on a larger, non-curated dataset.

These predicted segmentation masks served as input for automated landmark derivation and downstream AoP calculation.

3.2 Automated landmark derivation and AoP calculation

To calculate the angle of progression from the predicted segmentation mask, we used the standard approach also used in the PSFHS Challenge [1] (Fig. 1 c), which fits an ellipse to each of the pubic symphysis and fetal head segmentations and derives the AoP from the resulting ellipse axes. While the approach assumes that both structures are reasonably well approximated by an ellipse (an assumption that is frequently violated), we found in preliminary analyses that this method was more robust than fitting the landmarks for AoP calculation to the predicted segmentations directly.

4 Results

4.1 Segmentation performance

Overall, fetal head segmentation achieved higher Dice scores than pubic symphysis segmentation, with a median Dice of 0.92 compared with 0.90 for the pubic symphysis (Fig. 3). Despite strong overlap-based performance, the 95th percentile Hausdorff distance (HD95) was higher for the fetal head than for the pubic symphysis. This finding should be considered in relation to the substantially larger size and longer boundary of the fetal head. When stratified by frame-quality category, fetal head Dice showed a moderate decrease from standard-plane frames to inadequate frames (Fig. 3). The corresponding HD95 values increased across these categories, suggesting reduced boundary accuracy in lower-quality frames. For the pubic symphysis, segmentation performance was comparatively stable across frame categories, with no clear deterioration in Dice or HD95 in inadequate frames (Fig. 3).

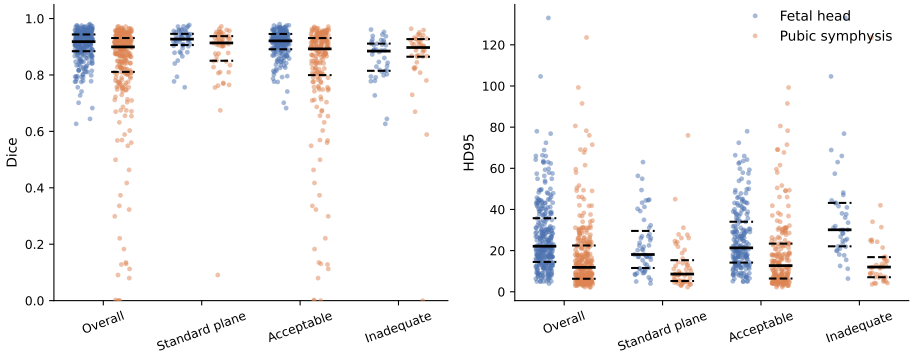

Fig. 3: Segmentation performance (left: Dice coefficient, right: Hausdorff distance) for fetal head and pubic symphysis overall and stratified by frame quality. Horizontal lines denote the Median and interquartile range.

4.2 Inter-rater variability of AoP measurements

The median interobserver absolute angular difference in expert AoP measurements was 4.09° (IQR 1.80–7.22°) (Tab. 2) corresponding approximately to a fetal head station difference of 0.4 cm (0.2–0.7 cm) according to the ISUOG conversion between AoP and fetal head station [6]. This provides important context for the human AoP baseline, as digital vaginal examination, the conventional clinical assessment of fetal head station, is known to be subjective and examiner-dependent, with substantial reported station errors [5]. Comparison with the original MFV landmarks showed a similar magnitude of deviation. The median

absolute AoP difference was 4.47° (2.11–8.17°) for Annotator 1 versus the MFV landmark and 4.69° (2.14–7.95°) for Annotator 2 versus the MFV landmark, corresponding approximately to 0.4 cm and 0.5 cm, respectively.

4.3 Training data curation reduces downstream AoP error

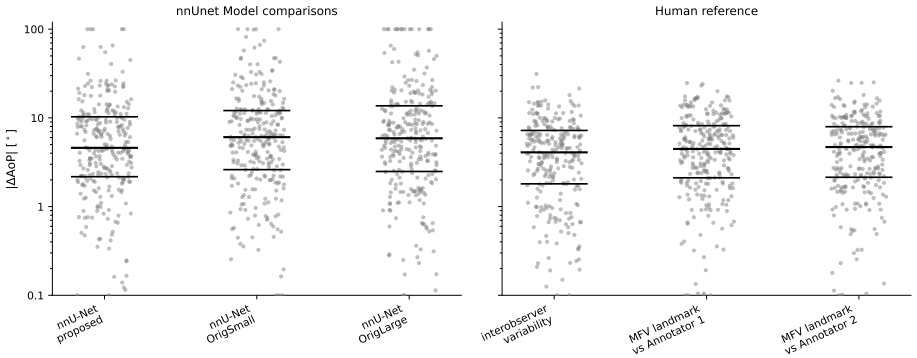

Fig. 4: $|\Delta\text{AoP}|$ [°] of models trained with different datasets (left) and human interobserver variability and comparison with MFV landmarks (right). Horizontal lines denote the Median AoP error and interquartile range.

Data curation and refinement led to more accurate downstream AoP calculation (Table 2, Fig. 4). Using the selected AoP calculation strategy, downstream AoP performance was primarily evaluated by comparing nnUnet-Proposed with the MFV-based reference models (nnUnet-OrigSmall and nnUnet-OrigLarge). nnUnet-Proposed achieved the lowest median absolute difference to the human baseline, with 4.59° [2.17–10.28°].

nnUnet-OrigSmall, trained on selected positive frames with original masks ($n = 356$), showed a higher median absolute difference of 6.06° [2.61–12.09°]. Although nnUnet-OrigLarge showed a slightly lower median error of 5.89° [2.49–13.68°] compared with nnUnet-OrigSmall, both MFV-based models showed higher median errors than nnUnet-Proposed.

Pairwise two-sided Wilcoxon signed-rank tests with Holm correction showed that the absolute AoP error was significantly lower for nnUnet-Proposed than

Table 2: $|\Delta\text{AoP}|$ [°] to human baseline and manual reference comparison. Failed model AoP calculations were set to 100° when baseline AoP was available.

Comparison	Median [Q1–Q3]	Failed AoP calculations
nnUnet-Proposed	4.59 [2.17–10.28]	5
nnUnet-OrigSmall	6.06 [2.61–12.09]	4
nnUnet-OrigLarge	5.89 [2.49–13.68]	15
Interobserver variability	4.09 [1.80–7.22]	–

for nnUnet-OrigLarge ($p = 0.006$). In contrast, the difference between nnUnet-Proposed and nnUnet-OrigSmall was not statistically significant ($p = 0.117$). Failed AoP calculations were included as errors of 100°.

Overall, training on the curated dataset with manually refined masks and frame-quality filtering resulted in the lowest median downstream AoP error. The improvement was statistically significant compared with nnUnet-OrigLarge, but not compared with nnUnet-OrigSmall.

Qualitative examples illustrate that AoP accuracy was closely related to the quality of pubic-symphysis segmentation (Fig. 5). In frames with anatomically plausible pubic-symphysis masks, all models produced comparable AoP estimates with low error (Fig. 5 a). In contrast, incomplete or geometrically inaccurate pubic-symphysis segmentations resulted in larger AoP deviations, particularly when the estimated symphysis axis and endpoints no longer reflected the underlying anatomy (Fig. 5 b, c).

Algorithmic failures occurred more frequently for nnUNet-OrigLarge despite the larger training dataset. In 10 frames, nnUnet-OrigLarge did not segment the pubic symphysis, resulting in zero pubic-symphysis pixels and subsequent failure of the AoP algorithm (Fig. 5 b). In these frames, nnUnet-OrigSmall and nnUnet-Proposed still produced pubic-symphysis segmentations and corresponding AoP values, although the segmentations were often of limited quality. This indicates that training-set size alone did not determine downstream AoP robustness.

Reliable endpoint estimation of the pubic symphysis by ellipse fitting requires a sufficient number of segmented pubic-symphysis pixels and an anatomically plausible mask geometry. Low pubic-symphysis pixel counts were associated with larger AoP deviations in several cases. Therefore, the predicted pubic-symphysis pixel count may provide a simple image-level quality-control measure and could be used as a deferral criterion to identify cases in which automated AoP estimation is likely to be unreliable.

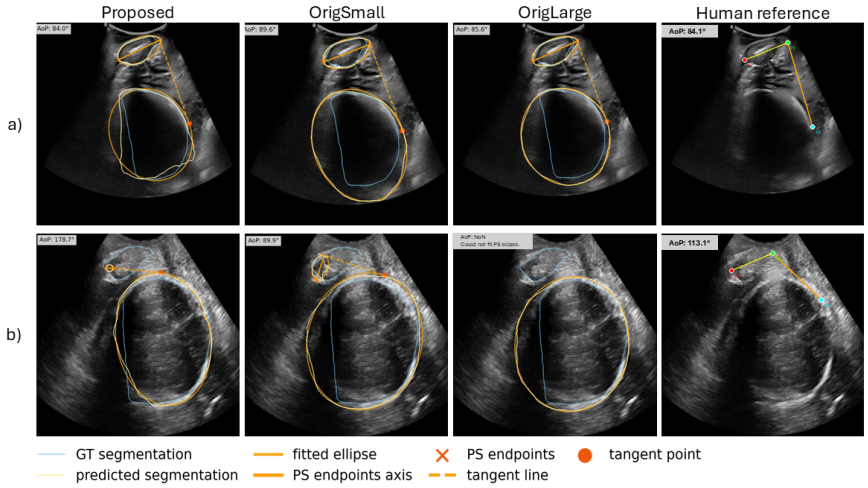

Fig. 5: Qualitative examples of AoP estimation using the three segmentation models Proposed, OrigSmall, OrigLarge (left to right), and Human reference. **a)** top row: low AoP errors, **b)** bottom row: AoP failure.

5 Discussion

We investigated whether careful dataset curation improves automated angle-of-progression estimation from intrapartum ultrasound. Our main finding is that a small, curated training set of quality-selected frames exceeded the downstream AoP performance of a segmentation model trained on the full, non-curated dataset, despite using an order of magnitude fewer frames. Manual refinement of the segmentation masks contributed an additional but more moderate improvement. Together these results suggest that, for this task, the quality and anatomical consistency of the training frames matter more than dataset size alone, a relevant consideration given that annotation effort, is typically the limiting factor for medical AI models. Our contributions are threefold, and we release all of them to support further work. First, we provide a curated subset of the Maternal-Fetal Ultrasound Video Dataset [11], comprising quality-classified frames and manually refined pubic symphysis and fetal head masks. Second, we provide an independent multi-expert reference for AoP, with landmark annotations from two physicians on the test frames, together with the resulting consensus values and a quantification of interobserver variability. Third, we release the segmentation and AoP-calculation code as a reproducible benchmark. The multi-expert reference is a particular strength, as it situates automated performance relative to human measurement variability. The principal limitation concerns robustness. While median agreement with the human reference approached interobserver variability, a subset of frames produced large AoP errors, driven primarily by inadequate pubic symphysis segmentation rather than by frame quality as classified. Because the pubic symphysis defines two of the landmarks

used for the angle, local segmentation failures propagate disproportionately into the final measurement. This points clearly to the next step: automated methods to flag unreliable predictions before an angle is reported. Such safeguards would be a prerequisite for clinical use, where a silently wrong angle is more harmful than a flagged missing one.

References

1. Bai, J., Zhou, Z., Ou, Z., Koehler, G., Stock, R., Maier-Hein, K., Elbatel, M., Martí, R., Li, X., Qiu, Y., Gou, P., Chen, G., Zhao, L., Zhang, J., Dai, Y., Wang, F., Silvestre, G., Curran, K., Sun, H., Xu, J., Cai, P., Jiang, L., Lan, L., Ni, D., Zhong, M., Chen, G., Campello, V.M., Lu, Y., Lekadir, K.: PSFHS challenge report: Pubic symphysis and fetal head segmentation from intrapartum ultrasound images. Medical Image Analysis **99**, 103353 (Jan 2025)
2. Barbera, A.F., Pombar, X., Perugino, G., Lezotte, D.C., Hobbins, J.C.: A new method to assess fetal head descent in labor with transperineal ultrasound. Ultrasound in Obstetrics & Gynecology **33**(3), 313–319 (2009)
3. Bultez, T., Quibel, T., Bouhanna, P., Popowski, T., Resche-Rigon, M., Rozenberg, P.: Angle of fetal head progression measured using transperineal ultrasound as a predictive factor of vacuum extraction failure: Angle of fetal head progression and vacuum extraction failure. Ultrasound in Obstetrics & Gynecology **48**(1) (2016)
4. Dall'Asta, A., Angeli, L., Kiener, A., Degennaro, V.A., Di Pasquo, E., Morganelli, G., Falcone, V., Salluce, M., Bontempo, P., Melito, C., Corno, E., Serio, M.D., Fieni, S., Ghi, T.: Correlation between clinical and sonographic assessment of the fetal head station in the second stage of labor: a prospective observational study. European Journal of Obstetrics & Gynecology and Reproductive Biology **313** (2025)
5. Dupuis, O., Silveira, R., Zentner, A., Dittmar, A., Gaucherand, P., Cucherat, M., Redarce, T., Rudigoz, R.C.: Birth simulator: Reliability of transvaginal assessment of fetal head station as defined by the American College of Obstetricians and Gynecologists classification. Am. J. Obstet. Gynecol. **192**(3) (2005)
6. Ghi, T., Eggebø, T., Lees, C., Kalache, K., Rozenberg, P., Youssef, A., Salomon, L.J., Tutschek, B.: ISUOG Practice Guidelines: intrapartum ultrasound. Ultrasound in Obstetrics & Gynecology **52**(1), 128–139 (Jul 2018)
7. Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J., Maier-Hein, K.H.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods **18**(2), 203–211 (Feb 2021)
8. Kalis, M.: Maternal-Fetal Edited Annotations. https://doi.org/10.5281/zenodo.21189540 (2026). https://doi.org/10.5281/zenodo.21189540, dataset
9. Kalis, M.: Software tools for landmark-based angle of progression annotation and segmentation-mask refinement in intrapartum ultrasound. https://github.com/mlm-lab-research/thesis_intrapartum_ultrasound (2026), gitHub repository, commit d7899262af260f5cd7f8b6e80dc3c98a26ab362e, accessed 4 August 2026
10. Lu, Y., Zhi, D., Zhou, M., Lai, F., Chen, G., Ou, Z., Zeng, R., Long, S., Qiu, R., Zhou, M., Jiang, X., Wang, H., Bai, J.: Multitask Deep Neural Network for the Fully Automatic Measurement of the Angle of Progression. Computational and Mathematical Methods in Medicine **2022**, 1–14 (Sep 2022)
11. Niu, M., Bai, J., Gao, Y., Tang, Y., Lu, Y., Han, Z., Hou, H., Huang, Y.: Maternal-Fetal Ultrasouno Video Dataset for End-to-end Intrapartum Biometry and Multitask Learning. Scientific Data **13**(1), 327 (Feb 2026)