

# Landmark Detection Challenge for Intrapartum Ultrasound Measurement Meeting the Actual Clinical Assessment of Labor Progress: Structured description of the challenge design

## CHALLENGE ORGANIZATION

### Title

Use the title to convey the essential information on the challenge mission.

Landmark Detection Challenge for Intrapartum Ultrasound Measurement Meeting the Actual Clinical Assessment of Labor Progress

### Challenge acronym

Preferable, provide a short acronym of the challenge (if any).

IUGC2025

### Challenge abstract

Provide a summary of the challenge purpose. This should include a general introduction in the topic from both a biomedical as well as from a technical point of view and clearly state the envisioned technical and/or biomedical impact of the challenge.

In 2018, the World Health Organization (WHO) published 56 recommendations to improve the quality of intrapartum care and enhance women's childbirth experiences. In response, the WHO developed the Labour Care Guide (LCG) in 2020, a next-generation tool designed to promote evidence-based, respectful, and woman-centered care during labor and delivery. The LCG was created through expert consultations, primary research with maternity healthcare providers, and usability studies across multiple countries. It serves as a practical tool for monitoring labor progress and maternal and fetal well-being by recording key clinical parameters. When deviations from normal labor progression are detected, the LCG highlights these issues, prompting timely interventions to ensure safe and effective care. Intrapartum ultrasound for labor progression analysis is a crucial examination in labor management. The core operation in this analysis is the identification of landmarks from intrapartum ultrasound images. These landmarks serve as the basis for subsequent qualitative evaluations of angles and distances, which offer valuable diagnostic information regarding labor arrest and influence decisions about the timing and type of intervention. However, obtaining reliable landmark annotations typically demands experienced physicians, and even for proficient obstetricians, manual landmark identification is a time-consuming and labor-intensive endeavor. Consequently, the development of fully automatic and precise landmark localization techniques has been an area of significant and persistent need.

The Intrapartum Ultrasound Grand Challenge (IUGC) 2025 is a collaborative initiative involving the "Deep Learning in Intrapartum Ultrasound Image Analysis" cooperative group and prominent clinical societies such as the International Society of Ultrasound in Obstetrics & Gynecology (ISUOG), the World Association of Perinatal

Medicine (WAPM), the Perinatal Medicine Foundation (PMF), and the National Institute for Health and Care-Excellence (NICE). The objective of this partnership is to formulate and promote clinically relevant challenges, thereby maximizing the potential clinical impact of innovative algorithmic contributions from participating teams.

Since its inception at MICCAI 2023, the IUGC has advanced the Pubic Symphysis - Fetal Head Segmentation (PSFHS) by facilitating and benchmarking algorithmic progress and providing high-quality annotated image datasets. In MICCAI 2024, the IUGC expanded to incorporate multiple benchmarks: (1) The analysis objects were extended from images to videos; (2) The tasks were augmented from image segmentation to classification, segmentation, and biometric parameter measurement; (3) The quantitative parameters were increased from one (i.e., Angle of Progression (AOP)) to two (i.e., AOP and head - symphysis distance (HSD)); and (4) The data sources were broadened from being solely from Asia to include Asia, Europe, and Africa. This novel and inventive design has established a benchmarking ecosystem for the systematic comparison of algorithms across diverse tasks and clinical challenges.

The significance of the IUGC 2025 lies in its concentration on addressing the actual clinical assessment of labor progress, covering (1) end-to-end measurements (which are currently indirect measurements based on segmentation results); (2) all fetal descent stations during the childbirth process (comprising five “minus”, one “zero”, and three “plus” stations for reliable longitudinal assessment of labor progress); (3) computational tasks (such as regression, detection); and (4) learning methods (semi-supervised, weakly-supervised, and barely-supervised learning). In line with the IUGC's goal of addressing clinical requirements, authoritative and leading clinical organizations have allied with the IUGC. We have extended the IUGC 2024 Challenge from an indirect ultrasound measurement based on segmentation results to an end-to-end measurement based on landmarks. Specifically, we provide 300 labeled cases and 31,421 unlabeled cases in the training set, 100 visible cases for validation, and 501 hidden cases for testing. The targets are the coordinates of three landmarks and the corresponding biometric parameter. In addition to the typical Mean Radial Error (MRE) and the absolute difference between predicted and manually measured parameters, our evaluation metrics also emphasize inference speed.

In summary, the IUGC 2025 challenge exhibits three primary characteristics:

- (1) Task: Employing semi-supervised landmark detection.
- (2) Dataset: Curating a large-scale and diverse fetal ultrasound dataset that accounts for all fetal descent stations during the childbirth process. It comprises 28,919 ultrasound images from over 20 medical groups.
- (3) Evaluation measures: Focusing on detection accuracy.
- (4) Multiple raters independently annotate a subset of test cases to compare algorithmic performance against human expert inter-rater variability.

### Challenge keywords

List the primary keywords that characterize the challenge.

Fetal Ultrasound, Fetal Biometry, Intrapartum Ultrasound, Artificial Intelligence, Angle of Progression, ISUOG, Perinatal Medicine, Intrapartum Care, Landmark detection, foundation models, Semi-supervised

### Year

2025

### Novelty of the challenge

Briefly describe the novelty of the challenge.

The novelty of the IUGC 2025 challenge lies in several aspects. Firstly, it adopts semi-supervised landmark detection for end-to-end measurements as the task, which is a departure from previous fully supervised segmentation for indirect measurements. Secondly, it curates a large and diverse fetal ultrasound dataset considering all fetal descent stations in the childbirth process, consisting of 32,000 ultrasound images from over 20 medical groups. Thirdly, the evaluation measures focus on detection accuracy. Additionally, multiple raters independently annotate a subset of test cases to enable a comparison of algorithmic performance against human expert inter-rater variability, providing a more comprehensive assessment framework.

### Task description and application scenarios

Briefly describe the application scenarios for the tasks in the challenge.

Ultrasound in labor can be performed using a transabdominal approach, mainly to determine head and spine position, or a transperineal approach, for assessment of head station and position at low stations. The core operation in this analysis is the identification of landmarks from intrapartum ultrasound images. These landmarks serve as the basis for subsequent qualitative evaluations of angles and distances, which offer valuable diagnostic information regarding labor arrest and influence decisions about the timing and type of intervention.

Our proposed algorithm is designed to complement and enhance the Labour Care Guide (LCG) by providing real-time, automated monitoring of labor progress. By integrating the algorithm's outputs—such as precise landmark detection and biometric measurements—into the LCG framework, we aim to empower healthcare providers with accurate, real-time data on fetal descent and other critical parameters. This integration will enable more personalized and evidence-based labor management, such as identifying when interventions are necessary or confirming that labor is progressing normally.

## FURTHER INFORMATION FOR CONFERENCE ORGANIZERS

### Workshop

If the challenge is part of a workshop, please indicate the workshop.

ASMUS 2025: The 6th International Workshop on Advances in Simplifying Medical UltraSound

### Duration

How long does the challenge take?

Half day

In case you selected half or full day, please explain why you need a long slot for your challenge.

The challenge is part of the ASMUS2025 workshop. Last year, 7 teams reported their methods and workshop accepted 22 articles.

### Expected number of participants

Please explain the basis of your estimate (e.g. numbers from previous challenges) and/or provide a list of potential participants and indicate if they have already confirmed their willingness to contribute.

Our first PSFHS challenge (MICCAI 2023) attracted 187 participants, and the second IUGC challenge (MICCAI 2024) had 125 team registrations from 18 countries. Based on the historical data of these previous challenges, we hope to have more than 100 teams participating.

### **Publication and future plans**

Please indicate if you plan to coordinate a publication of the challenge results.

We intend to coordinate 2 specific publication plans immediately after the challenge.

(1): Coordination of the IUGC proceedings allows the participants to publish their methods in the associated Springer LNCS post-conference proceedings. We have already been doing this for IUGC since 2024.

(2): We will coordinate journal manuscripts focusing on publishing and summarizing the results of each IUGC 2025 challenge, making a comprehensive meta-analysis for each to inform the community about the obtained results, findings, and insights.

### **Space and hardware requirements**

Organizers of on-site challenges must provide a fair computing environment for all participants. For instance, algorithms should run on the same computing platform provided to all.

For algorithm implementation and training, algorithms are trained using the participants' computing infrastructure.

For testing as part of the challenge, we would use the platform CodaLab.

# TASK 1: Landmark Detection for Intrapartum Ultrasound Measurement

## SUMMARY

### Abstract

Provide a summary of the challenge purpose. This should include a general introduction in the topic from both a biomedical as well as from a technical point of view and clearly state the envisioned technical and/or biomedical impact of the challenge.

Intrapartum ultrasound for labor progression analysis is a crucial examination in labor management. The core operation in this analysis is the identification of landmarks from intrapartum ultrasound images. These landmarks serve as the basis for subsequent qualitative evaluations of angles and distances, which offer valuable diagnostic information regarding labor arrest and influence decisions about the timing and type of intervention. However, obtaining reliable landmark annotations typically demands experienced physicians, and even for proficient obstetricians, manual landmark identification is a time-consuming and labor-intensive endeavor. Consequently, the development of fully automatic and precise landmark localization techniques has been an area of significant and persistent need.

### Keywords

List the primary keywords that characterize the task.

Fetal Ultrasound, Fetal Biometry, Intrapartum Ultrasound, Artificial Intelligence, Angle of Progression, ISUOG, Perinatal Medicine, Intrapartum Care, Landmark detection, foundation models, Semi-supervised

## ORGANIZATION

### Organizers

a) Provide information on the organizing team (names and affiliations).

Jieyun Bai (Jinan University/University of Auckland)

Isaac Khobo (University of Cape Town)

Saad Slimani (Hassan II University)

Yaosheng Lu (Jinan University)

Dong Ni (Shenzhen University)

Mohammad Yaqub (Mohamed bin Zayed University of Artificial Intelligence)

Karim Lekadir (Universitat de Barcelona)

Jun Ma (University of Toronto)

Shuo Li (Case Western Reserve University)

b) Provide information on the primary contact person.

Jieyun Bai (fugc.isbi25@gmail.com)

### Life cycle type

Define the intended submission cycle of the challenge. Include information on whether/how the challenge will be continued after the challenge has taken place. Not every challenge closes after the submission deadline (one-time event). Sometimes it is possible to submit results after the deadline (open call) or the challenge is repeated with some

modifications (repeated event).

Examples:

- One-time event with fixed conference submission deadline
- Open call (challenge opens for new submissions after conference deadline)
- Repeated event with annual fixed conference submission deadline

Open call (challenge opens for new submissions after conference deadline)

### Challenge venue and platform

a) Report the event (e.g. conference) that is associated with the challenge (if any).

ASMUS 2025: The 6th International Workshop on Advances in Simplifying Medical UltraSound(a half-day session)

b) Report the platform (e.g. grand-challenge.org) used to run the challenge.

<https://competitions.codalab.org/>

c) Provide the URL for the challenge website (if any).

None at this moment.

### Participation policies

a) Define the allowed user interaction of the algorithms assessed (e.g. only (semi-) automatic methods allowed).

Fully automatic

b) Define the policy on the usage of training data. The data used to train algorithms may, for example, be restricted to the data provided by the challenge or to publicly available data including (open) pre-trained nets.

Publicly available data is allowed

c) Define the participation policy for members of the organizers' institutes. For example, members of the organizers' institutes may participate in the challenge but are not eligible for awards.

May participate but not eligible for awards and not listed in leaderboard

d) Define the award policy. In particular, provide details with respect to challenge prizes.

The top 10 teams will be awarded prizes. The prizes will consist of both cash and a certificate. The cash prize amount for IUGC 2025 is set at 10,000 RMB, which is the same as that of IUGC 2023 and 2024. The specific breakdown of how the cash prize will be distributed among the top 10 teams and any additional details regarding the certificates will be announced on our challenge homepage at a later date.

e) Define the policy for result announcement.

Examples:

- Top 3 performing methods will be announced publicly.
- Participating teams can choose whether the performance results will be made public.

All the performance results will be announced publicly.

f) Define the publication policy. In particular, provide details on ...

- ... who of the participating teams/the participating teams' members qualifies as author
- ... whether the participating teams may publish their own results separately, and (if so)
- ... whether an embargo time is defined (so that challenge organizers can publish a challenge paper first).

Participating teams are required to submit a manuscript to describe their methods, and these papers will be publicly available to the community. If the method has already been published, participants can submit the URL to the existing manuscript.

### Submission method

a) Describe the method used for result submission. Preferably, provide a link to the submission instructions.

Examples:

- Docker container on the Synapse platform. Link to submission instructions: <URL>
- Algorithm output was sent to organizers via e-mail. Submission instructions were sent by e-mail.

For the purpose of result verification and to encourage reproducibility and transparency, all entries must submit the following:

(1) Docker containers.

(2) A paper highlighting the contribution of the submission, but not limited to, the method, experimental results and analysis, prepared according to the format stipulated by MICCAI 2025. All challenge entries should be accompanied by a description of the method.

(3) GitHub repository URL containing all source codes for their implemented method and all other relevant files such as feature/parameter data. To help publicize our workshop and domain area, please mention IUGC 2025. The participants may provide this URL in a simple text file while submitting.

(4) For all files, participants should submit a single zip file and upload it to the submission system as supplementary material.

The submission link will be made available starting 01/08/2025

(5) Winners of the challenge are required to submit their source code and a detailed method description. We will rigorously review and execute their code locally to verify compliance with the challenge's learning objectives and ensure that the methods align with the intended semi-supervised approaches.

b) Provide information on the possibility for participating teams to evaluate their algorithms before submitting final results. For example, many challenges allow submission of multiple results, and only the last run is officially counted to compute challenge results.

The participating teams will be able to validate their results based on the validation set provided by the organizers. Submissions to IUGC 2025 are issued a validation score. This is to provide a sanity check of the submission (ensure the submission is in the correct format) and is not intended to be used for algorithm ranking or evaluation.

### Challenge schedule

Provide a timetable for the challenge. Preferably, this should include

- the release date(s) of the training cases (if any)

- the registration date/period
- the release date(s) of the test cases and validation cases (if any)
- the submission date(s)
- associated workshop days (if any)
- the release date(s) of the results

Training data release: 15/03/2025

Validation data release: 01/06/2025

Evaluation: 01/07/2025

Submission deadline: 01/08/2025

Winner and invitation speakers: 01/09/2025

Announcement of results at MICCAI 2025: (subject to change depending on the MICCAI 2025 deadlines)

### **Ethics approval**

Indicate whether ethics approval is necessary for the data. If yes, provide details on the ethics approval, preferably institutional review board, location, date and number of the ethics approval (if applicable). Add the URL or a reference to the document of the ethics approval (if available).

All data is anonymized. Ethics approval by the Medical Ethics Committee of the First Affiliated Hospital of Jinan University (No. JNUKY-2022-019).

### **Data usage agreement**

Clarify how the data can be used and distributed by the teams that participate in the challenge and by others during and after the challenge. This should include the explicit listing of the license applied.

Examples:

- CC BY (Attribution)
- CC BY-SA (Attribution-ShareAlike)
- CC BY-ND (Attribution-NoDerivs)
- CC BY-NC (Attribution-NonCommercial)
- CC BY-NC-SA (Attribution-NonCommercial-ShareAlike)
- CC BY-NC-ND (Attribution-NonCommercial-NoDerivs)

CC BY-NC-SA (Attribution-NonCommercial-ShareAlike)

### **Code availability**

a) Provide information on the accessibility of the organizers' evaluation software (e.g. code to produce rankings). Preferably, provide a link to the code and add information on the supported platforms.

We will provide an online platform (<https://competitions.codalab.org/>) to evaluate the results. For transparency, we will release the source code used for calculating final scores after the closing date of the challenge.

b) In an analogous manner, provide information on the accessibility of the participating teams' code.



Algorithm code release and corresponding conference paper submission will be prerequisites for award eligibility and further use of the data. To this end, GitHub repository URL containing the source code for their implemented method and conference paper following the guidelines will need to be provided.

### Conflicts of interest

Provide information related to conflicts of interest. In particular provide information related to sponsoring/funding of the challenge. Also, state explicitly who had/will have access to the test case labels and when.

There are no conflicts of interest.

## MISSION OF THE CHALLENGE

### Field(s) of application

State the main field(s) of application that the participating algorithms target.

Examples:

- Diagnosis
- Education
- Intervention assistance
- Intervention follow-up
- Intervention planning
- Prognosis
- Research
- Screening
- Training
- Cross-phase

CAD, Research, Decision support, Diagnosis

### Task category(ies)

State the task category(ies)

Examples:

- Classification
- Detection
- Localization
- Modeling
- Prediction
- Reconstruction

- Registration
- Retrieval
- Segmentation
- Tracking

## Detection, Localization

### Cohorts

We distinguish between the target cohort and the challenge cohort. For example, a challenge could be designed around the task of medical instrument tracking in robotic kidney surgery. While the challenge could be based on ex vivo data obtained from a laparoscopic training environment with porcine organs (challenge cohort), the final biomedical application (i.e. robotic kidney surgery) would be targeted on real patients with certain characteristics defined by inclusion criteria such as restrictions regarding sex or age (target cohort).

a) Describe the target cohort, i.e. the subjects/object(s) from whom/which the data would be acquired in the final biomedical application.

**The target cohort consists of pregnant women in labor.**

b) Describe the challenge cohort, i.e. the subject(s)/object(s) from whom/which the challenge data was acquired.

**Pregnant women required to assess the fetal station.**

### Imaging modality(ies)

Specify the imaging technique(s) applied in the challenge.

**2D ultrasound**

### Context information

Provide additional information given along with the images. The information may correspond ...

a) ... directly to the image data (e.g. tumor volume).

**(1) One landmark on the fetal head and two on the pubic symphysis.**

**(2) Biometric parameters were measured with these landmarks.**

b) ... to the patient in general (e.g. sex, medical history).

**The images are acquired from pregnant woman with a variety of age (from 18-year old to 46-year old).**

### Target entity(ies)

a) Describe the data origin, i.e. the region(s)/part(s) of subject(s)/object(s) from whom/which the image data would be acquired in the final biomedical application (e.g. brain shown in computed tomography (CT) data, abdomen shown in laparoscopic video data, operating room shown in video data, thorax shown in fluoroscopy video). If necessary, differentiate between target and challenge cohort.

**The data origin is the pubic symphysis-fetal head in the 2D B mode ultrasound image.**

b) Describe the algorithm target, i.e. the structure(s)/subject(s)/object(s)/component(s) that the participating algorithms have been designed to focus on (e.g. tumor in the brain, tip of a medical instrument, nurse in an operating

theater, catheter in a fluoroscopy scan). If necessary, differentiate between target and challenge cohort.

The algorithm target is the pubic symphysis-fetal head shown in the 2D B mode ultrasound image. Each image has three landmarks used to calculate biomedical parameters.

### Assessment aim(s)

Identify the property(ies) of the algorithms to be optimized to perform well in the challenge. If multiple properties are assessed, prioritize them (if appropriate). The properties should then be reflected in the metrics applied (see below, parameter metric(s)), and the priorities should be reflected in the ranking when combining multiple metrics that assess different properties.

- Example 1: Find highly accurate liver segmentation algorithm for CT images.
- Example 2: Find lung tumor detection algorithm with high sensitivity and specificity for mammography images.

Corresponding metrics are listed below (parameter metric(s)).

Accuracy, Usability, Runtime, Robustness

## DATA SETS

### Data source(s)

a) Specify the device(s) used to acquire the challenge data. This includes details on the device(s) used to acquire the imaging data (e.g. manufacturer) as well as information on additional devices used for performance assessment (e.g. tracking system used in a surgical setting).

Transperineal ultrasound examinations were performed in standard B-mode ultrasound using systems of different vendors such as Philips-cx50, Toshiba Aplio300, Voluson P8, Esaote Mylab, Lian-med ObEye and Youkey Q7.

b) Describe relevant details on the imaging process/data acquisition for each acquisition device (e.g. image acquisition protocol(s)).

In order to obtain high-quality images, the transducer was prepped by covering it with a surgical latex glove filled with coupling gel, then the prepped transducer, after applying gel, was placed between labia below the pubic symphysis to obtain a sagittal plane, small adjustments in the form of lateral movements of the probe were made until an image obtained showed clear maternal pelvic (pubic symphysis) and fetal (fetal skull) landmarks that did not show any shadows from the pubic rami.

c) Specify the center(s)/institute(s) in which the data was acquired and/or the data providing platform/source (e.g. previous challenge). If this information is not provided (e.g. for anonymization reasons), specify why.

First Affiliated Hospital of Jinan University

Zhujiang Hospital of Southern Medical University

Nanfeng Hospital of Southern Medical University

Third Affiliated Hospital of Sun Yat-sen University

Guangzhou Women and Children's Medical Center of Guangzhou Medical University

Third Affiliated Hospital of Guangzhou Medical University

Shenzhen People's Hospital

Chengdu Women's and Children's Central Hospital of the University of Electronic Science and Technology of China

Wuxi People's Hospital  
 Department of Obstetrics and Gynecology of Fudan University  
 Lu'an City Jin'an District Maternal and Child Health Hospital  
 Lianyungang Maternal and Child Health Hospital  
 Weifang Maternal and Child Health Hospital  
 Qianjiang Central Hospital  
 Shantou Maternal and Child Health Hospital  
 Ma'anshan Maternal and Child Health Hospital  
 Zhangjiagang First People's Hospital  
 Peking University First Hospital  
 Taizhou First People's Hospital  
 Xianju County People's Hospital  
 Sanmen County People's Hospital  
 Foshan Nanhai District People's Hospital  
 Zhongshan Boai Hospital  
 Dongguan Donghua Hospital

d) Describe relevant characteristics (e.g. level of expertise) of the subjects (e.g. surgeon)/objects (e.g. robot) involved in the data acquisition process (if any).

Data were acquired by specialized teams consisting of sonographers, obstetricians, and technologists, both with more than seven years of professional experience.

Manual measurements were performed by three sonographers with experience in ultrasound imaging.

### Training and test case characteristics

a) State what is meant by one case in this challenge. A case encompasses all data that is processed to produce one result that is compared to the corresponding reference result (i.e. the desired algorithm output).

Examples:

- Training and test cases both represent a CT image of a human brain. Training cases have a weak annotation (tumor present or not and tumor volume (if any)) while the test cases are annotated with the tumor contour (if any).
- A case refers to all information that is available for one particular patient in a specific study. This information always includes the image information as specified in data source(s) (see above) and may include context information (see above). Both training and test cases are annotated with survival (binary) 5 years after (first) image was taken.

**A case (training or test case) consists of one ultrasound image.**

b) State the total number of training, validation and test cases.

Training cases: 31421 (300)

Validation cases: 100

Test cases: 501

c) Explain why a total number of cases and the specific proportion of training, validation and test cases was chosen.

The number of training cases is a trade-off between the effort of manual measurement and the necessity for sufficient training data. To our knowledge, this is the largest publicly available labeled intrapartum ultrasound dataset.

d) Mention further important characteristics of the training, validation and test cases (e.g. class distribution in classification tasks chosen according to real-world distribution vs. equal class distribution) and justify the choice.

Manual landmark annotations were made by using in-house developed software. The annotations were performed by two experienced doctors. Both doctors were given training on the annotation software by the software developers.

e) Challenge organizers are encouraged to (partly) use unseen, unpublished data for their challenges. Describe if new data will be used for the challenge and state the number of cases along with the proportion of new data.

30% of the data is unseen and unpublished.

### Annotation characteristics

a) Describe the method for determining the reference annotation, i.e. the desired algorithm output. Provide the information separately for the training, validation and test cases if necessary. Possible methods include manual image annotation, in silico ground truth generation and annotation by automatic methods.

If human annotation was involved, state the number of annotators.

The three annotators are in accordance with the guidelines outlined in ".ISUOG Practice Guidelines: intrapartum ultrasound".

b) Provide the instructions given to the annotators (if any) prior to the annotation. This may include description of a training phase with the software. Provide the information separately for the training, validation and test cases if necessary. Preferably, provide a link to the annotation protocol.

The following annotation protocol was defined:

Step 1: Identification of intrapartum ultrasound standard planes by selecting high-quality images containing intact targets (i.e., maternal pubic symphysis (PS) and fetal head (FH)) from an ultrasound video;

Step 2: Manual measurement of biometric parameters as defined by clinical guidelines: AOP is the angle between the long axis of the pubic bone and a line from the lowest edge of the pubis drawn tangentially to the deepest bony part of the fetal skull.

c) Provide details on the subject(s)/algorithm(s) that annotated the cases (e.g. information on level of expertise such as number of years of professional experience, medically-trained or not). Provide the information separately for the training, validation and test cases if necessary.

Three experts annotated training and test data:

At the First Affiliated hospital of Jinan University, a obstetrician with 7 years of experience in intrapartum ultrasound examinations;

At the Zhujiang Hospital of Southern Medical University, a sonographer with 10 years of experience in ultrasound imaging and experience in machine learning research annotated all data;

At the Third Affiliated Hospital of Sun Yat-sen University, a sonographer with 8 years of experience in intrapartum ultrasound examinations.

Regarding annotation quality, the process is overseen by the challenge's clinical chair, Dr. Gaowen Chen, a senior

radiologist with over 10 years of experience in intrapartum ultrasound imaging.

d) Describe the method(s) used to merge multiple annotations for one case (if any). Provide the information separately for the training, validation and test cases if necessary.

N/A

### Data pre-processing method(s)

Describe the method(s) used for pre-processing the raw training data before it is provided to the participating teams. Provide the information separately for the training, validation and test cases if necessary.

The US images were anonymized by first removing any patient-related information on each image. Then all images were renamed and converted to the same format.

### Sources of error

a) Describe the most relevant possible error sources related to the image annotation. If possible, estimate the magnitude (range) of these errors, using inter-and intra-annotator variability, for example. Provide the information separately for the training, validation and test cases, if necessary.

Due to the relatively low brightness of tissue regions, some landmarks are illegible, which would likely result in inconsistent annotation that affects the inference performance of models and the selection of models. The majority of the landmarks are clear enough to allow accurate annotation. For landmarks on soft tissue, the two senior doctors manually adjust the image contrast using in-house developed software. This process enhances the discernibility of soft tissue structures and ensures the accurate annotation of corresponding landmarks. In addition, we have calculated the inter-observer variability of the three senior doctors to assess human performance for each landmark. Current findings suggest that Interclass Correlation Coefficients (ICCs) are excellent for all landmarks (ICC over 0.90 for each landmark). Moreover, our double-check and the quality control approach will largely reduce this source of error.

b) In an analogous manner, describe and quantify other relevant sources of error.

The delineation of structures in ultrasound images is a challenging task, as some of the boundaries are less well-defined

## ASSESSMENT METHODS

### Metric(s)

a) Define the metric(s) to assess a property of an algorithm. These metrics should reflect the desired algorithm properties described in assessment aim(s) (see above). State which metric(s) were used to compute the ranking(s) (if any).

- Example 1: Dice Similarity Coefficient (DSC)
- Example 2: Area under curve (AUC)

1) Mean Radial Error (MRE);

2) The difference between predicted and manually measured ultrasound parameters.

All metrics will be used to compute the ranking.

b) Justify why the metric(s) was/were chosen, preferably with reference to the biomedical application.

The two metrics are complementary.

(1) Specifically, the Mean Radial Error (MRE) measures the difference between two landmarks. The error between the ground truth implant and the automatic results from participants' algorithm can be effectively measured using this metric. MRE serves as a distance error metric, with a lower value indicating better performance, formulated as follows:  $MRE = \frac{\sum(\sqrt{\Delta x_i^2 + \Delta y_i^2})}{N}$ ,  $i$  from 1 to  $N$ , where  $\Delta x_i$  is the absolute distance in x-direction between the predicted and the reference landmark,  $\Delta y_i$  is the absolute distance in y-direction between the predicted and the reference landmark, and  $N$  represents the number of detection landmark.

(2) The absolute value of the difference between predicted and manually measured ultrasound parameters indicates whether the prediction is consistent with the label.

These evaluation metrics will guide the algorithm to strike a balance between performance and efficiency.

### Ranking method(s)

a) Describe the method used to compute a performance rank for all submitted algorithms based on the generated metric results on the test cases. Typically the text will describe how results obtained per case and metric are aggregated to arrive at a final score/ranking.

Step 1: Separate rankings will be computed based on each metric;

Step 2: From the three ranking tables, the ranking of each participant will be computed as  $0.5 \cdot (MRE) + 0.5 \cdot (\text{Absolute Parameter Difference})$ ;

Step 3: In case of equal ranking, the achieved biometrical parameter will be used as a tiebreak of the single rankings;

b) Describe the method(s) used to manage submissions with missing results on test cases.

Submissions with missing results on test cases will not be considered for the leaderboard.

c) Justify why the described ranking scheme(s) was/were used.

The ranking scheme was selected based on the robustness analysis capabilities provided by the challengeR package. This decision was driven by the need to ensure the reliability and fairness of the competition rankings. The challengeR package utilizes Kendall's tau as its core metric for evaluating rank correlation, which effectively measures the degree of similarity between different sets of rankings. A significant aspect of employing Kendall's tau is its ability to identify consistent ranking patterns, thereby affirming the stability and consistency of the results amidst potential variations. The implementation of this measure is crucial in preserving the integrity of the competition, as it guarantees that the rankings reflect a fair and accurate assessment of participant performance, thus maintaining the trust and confidence of all involved parties.

### Statistical analyses

a) Provide details for the statistical methods used in the scope of the challenge analysis. This may include

- description of the missing data handling,
- details about the assessment of variability of rankings,
- description of any method used to assess whether the data met the assumptions, required for the particular statistical approach, or
- indication of any software product that was used for all data analysis methods.

We will exclude the participants who fail to report on the whole testing set. Besides the statistical values such as mean, standard deviation of the four evaluation metrics, we use the p-value in Wilcoxon test to assess whether the top performing/ranking algorithms are significantly better than the rest of algorithms. To measure the variability, ranking variability will be characterized using the bootstrap.

b) Justify why the described statistical method(s) was/were used.

The Wilcoxon test is chosen because it is non-parametric and allows us to perform the analysis with minimal hypotheses, and the Bootstrap is a simple nonparametric method that relies on minimal assumptions

### Further analyses

Present further analyses to be performed (if applicable), e.g. related to

- combining algorithms via ensembling,
- inter-algorithm variability,
- common problems/biases of the submitted methods, or
- ranking variability.

The further analyses will be discussed in a further publication after the challenge.

## ADDITIONAL POINTS

### References

Please include any reference important for the challenge design, for example publications on the data, the annotation process or the chosen metrics as well as DOIs referring to data or code.

Ghi T, Eggebø T, Lees C, et al. ISUOG Practice Guidelines: intrapartum ultrasound[J]. Ultrasound in Obstetrics & Gynecology, 2018, 52(1): 128-139.

Rizzo G, Ghi T, Henrich W, et al. Ultrasound in labor: clinical practice guideline and recommendation by the WAPM-World Association of Perinatal Medicine and the PMF-Perinatal Medicine Foundation[J]. Journal of Perinatal medicine, 2022, 50(8): 1007-1029.

Bai J, Zhou Z, Ou Z, et al. PSFHS challenge report: Pubic symphysis and fetal head segmentation from intrapartum ultrasound images[J]. Medical Image Analysis, 2025, 99: 103353.

Zhou Z, Lu Y, Bai J, et al. Segment Anything Model for fetal head-pubic symphysis segmentation in intrapartum ultrasound image analysis[J]. Expert Systems with Applications, 2024: 125699.

Jiang J, Wang H, Bai J, et al. Intrapartum Ultrasound Image Segmentation of Pubic Symphysis and Fetal Head Using Dual Student-Teacher Framework with CNN-ViT Collaborative Learning[C]//International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer Nature Switzerland, 2024: 448-458.

Chen Z, Ou Z, Lu Y, et al. Direction-guided and multi-scale feature screening for fetal head-pubic symphysis segmentation and angle of progression calculation[J]. Expert Systems with Applications, 2024, 245: 123096.

Ou Z, Bai J, Chen Z, et al. RTSeg-Net: a lightweight network for real-time segmentation of fetal head and pubic symphysis from intrapartum ultrasound images[J]. Computers in biology and medicine, 2024, 175: 108501.

Qiu R, Zhou M, Bai J, et al. PSFHSP-Net: an efficient lightweight network for identifying pubic symphysis-fetal head standard plane from intrapartum ultrasound images[J]. Medical & Biological Engineering & Computing, 2024: 1-12.

Chen G, Bai J, Ou Z, et al. PSFHS: intrapartum ultrasound image dataset for AI-based segmentation of pubic symphysis and fetal head[J]. Scientific Data, 2024, 11(1): 436.



- Chen Z, Lu Y, Long S, et al. Fetal head and pubic symphysis segmentation in intrapartum ultrasound image using a dual-path boundary-guided residual network[J]. IEEE Journal of Biomedical and Health Informatics, 2024.
- Lu Y, Zhou M, Zhi D, et al. The JNU-IFM dataset for segmenting pubic symphysis-fetal head[J]. Data in brief, 2022, 41: 107904.
- Bai J, Sun Z, Yu S, et al. A framework for computing angle of progression from transperineal ultrasound images for evaluating fetal head descent using a novel double branch network[J]. Frontiers in physiology, 2022, 13: 940150.
- Lu Y, Zhi D, Zhou M, et al. Multitask deep neural network for the fully automatic measurement of the angle of progression[J]. Computational and mathematical methods in medicine, 2022, 2022(1): 5192338.
- Zhou M, Yuan C, Chen Z, et al. Automatic angle of progress measurement of intrapartum transperineal ultrasound image with deep learning[C]//Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part VI 23. Springer International Publishing, 2020: 406-414.
- Bai, J., Lekadir, K., Ni, D., Slimani, S., Campello, V. M., Ohene-Botwe, B., Lu, Y., Chen, G., Hou, H., Qiu, D., & Zhou, Z. (2024). Intrapartum Ultrasound Grand Challenge 2024. 27th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI 2024). Zenodo.  
<https://doi.org/10.5281/zenodo.10979813>
- Jieyun Bai, Zhanhong Ou, Yaosheng Lu, Dong Ni, Gaowen Chen, Gaowen Chen, Zhanhong Ou, Zhanhong Ou, Gaowen Chen, & Yaosheng Lu. (2023). Pubic Symphysis-Fetal Head Segmentation from Transperineal Ultrasound Images. International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI) 2023 (MICCAI 2023). Zenodo. <https://doi.org/10.5281/zenodo.7861699>  
<https://ps-fh-aop-2023.grand-challenge.org/>  
<https://codalab.lisn.upsaclay.fr/competitions/18413>

### Further comments

Further comments from the organizers.

The IUGC2025 is part of ASMUS 2025: The 6th International Workshop on Advances in Simplifying Medical UltraSound. This event is co-organized by Dong Ni, Karim Lekadir and Jieyun Bai.