FUSQA: Fetal Ultrasound Segmentation Quality Assessment

Sevim Cengiz, Ibrahim Almakky, Mohammad Yaqub

Computer Vision Department, Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, United Arab Emirates

Background

- Inaccurate crown-rump length (CRL) measurement can lead to incorrect gestation age (GA) calculation, resulting in a wrong assessment of fetal growth.
- CRL measurements can be either obtained manually or through segmentation-based algorithms.
- The quality of the segmentation-based algorithms needs to be checked for CRL measurement performance.

Aim

- To verify segmentation quality automatically with a deep learning-based method called FUSQA.
- To demonstrate the importance of an automated fetal ultrasound image quality assessment approach on a clinical essential downstream task (accurate CRL measurement and GA estimation).

Materials & Methods

- D_A: 696 fetal ultrasound images (UK-acquired)
- D_B: 226 fetal ultrasound images (UAE-acquired)
- A single CNN model used to estimate the probability of being good segmentation (Fig1).
- A set of altered masks from ground truth masks were generated (Fig2).

Proposed Model

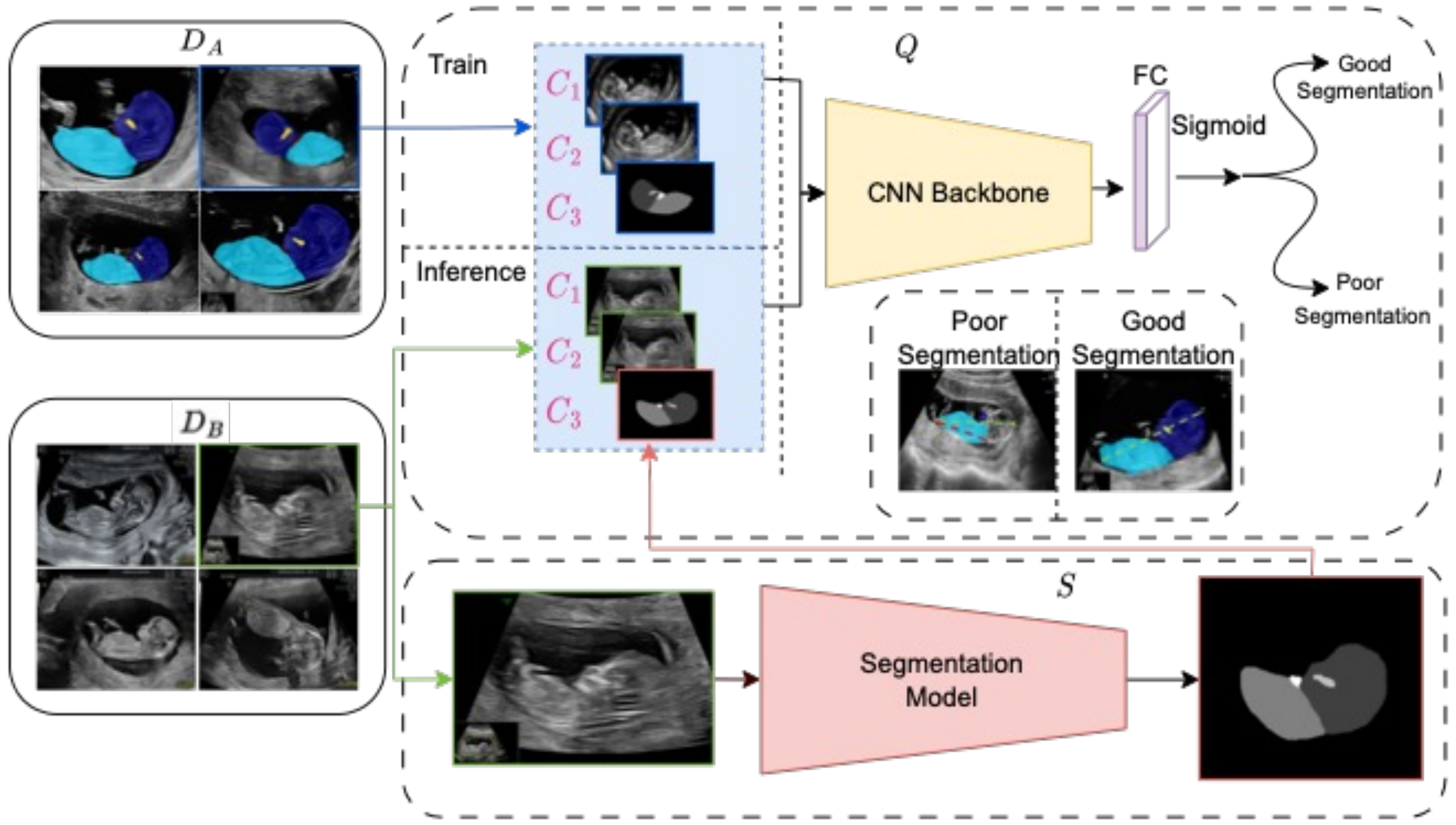

Fig1: The train and inference pipelines for FUSQA method using two datasets D_A and D_B. The bottom-right shows samples of good and bad segmentation masks and their impact on the CRL estimation.

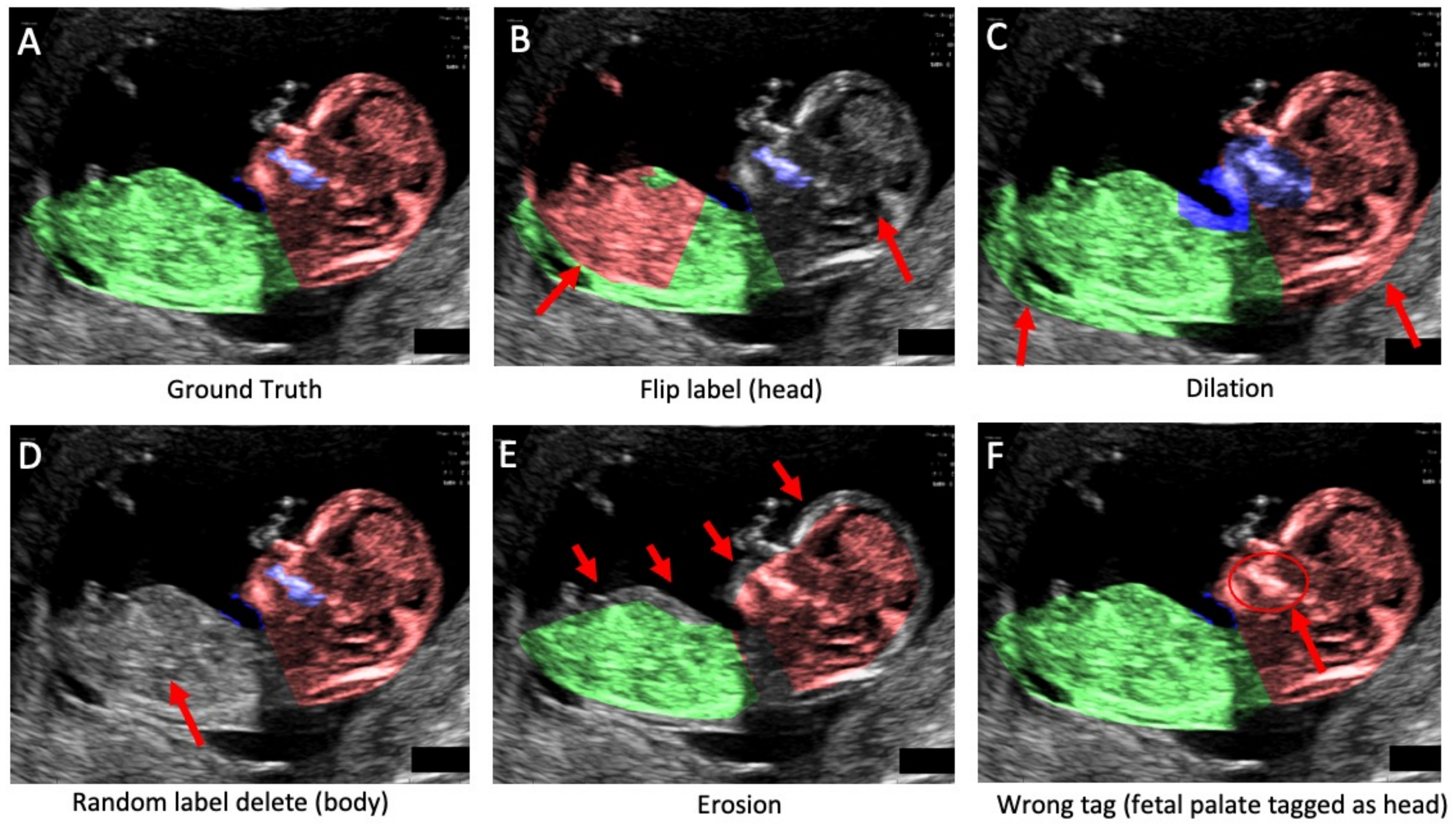

Fig2: An example ground truth segmentation mask (A) and altered poor segmentation masks from it with the following variants: flipping of the head (B), segmentation over-estimation (C), randomly selected label deletion (D), segmentation under-estimation (E), fetal palate tagged as the head (F).

Results

Table 1: The comparison on unseen D_B between the Siamese, Synergic, and Single CNN models with different backbones.

Model	Network	Precision	Recall	Acc.	F1 Score
Siamese	ResNet 18	0.673	0.97	0.75	0.795
	ResNet 50	0.5	1.0	0.5	0.66
Synergic	ResNet 18	0.525	0.983	0.547	0.685
	ResNet 50	0.746	0.87	0.787	0.803
Proposed	ResNet 18	0.795	**1.0**	0.871	**0.886**
	ResNet 50	0.804	0.982	**0.902**	0.87
Anomaly Detection [1]	CAE	**1.0**	0.72	0.70	0.83

Table 2: The mean CRL and GA estimation errors in predicted good and poor-quality segmentation masks on unseen images from (D_B) (Top). A breakdown of the compliance of predicted good and poor segmentation masks with FASP criteria on the dataset D_B (Bottom).

		Poor Seg.	Good Seg.
Downstream task on proposed model	CRL diff (mm) ↓	13.63	2.64
	GA diff (days) ↓	7.73	1.45
Downstream task on CAE	CRL diff (mm) ↓	17.71	5.01
	GA diff (days) ↓	10.06	2.79
Image Guidance Criteria	Neutral position ↑	67.6%	79.7%
	Fetal palate ↑	64.9%	93.6%
	Magnification ↑	75%	99.1%
	Fetal face direction ↑	76%	99.1%
	Horizontal orientation ↑	67.6%	100%
	Left caliper definition ↑	45.4%	85.2%
	Right caliper definition ↑	63.9%	95.4%
	Acceptance of CRL ↑	64.9%	96.3%

Conclusions

- We formulate the segmentation quality assessment process as an automated classification task to distinguish between good and poor quality segmentation masks for more accurate gestational age estimation.
- We validate the performance of our proposed approach on two datasets we collect from two hospitals using different ultrasound machines.
- Our best-performing architecture achieved over 90% classification accuracy on an unseen dataset.

[1] Francesco Galati and Maria A. Zuluaga. Functional Imaging and Modeling of the Heart, Lecture Notes in Computer Science, pages 101–111, Cham, 2021. doi: 10.1007/978-3-030-78710-3 11.