Unsupervised Ultrasound Image Quality Assessment with Score Consistency and Relativity Co-learning

Juncheng Guo[1], Jianxin Lin[1,*], Guanghua Tan[1], Yuhuan Lu[1], Zhan Gao[1], Shengli Li[2], and Kenli Li[1]

[1] The College of Computer Science and Electronic Engineering, Hunan university, Changsha, China
linjianxin@hnu.edu.cn
[2] Department of Ultrasound, Affiliated Shenzhen Maternal and Child Healthcare Hospital of Nanfang Medical University, Shenzhen, China

Abstract. Selecting an optimal standard plane in prenatal ultrasound is crucial for improving the accuracy of AI-assisted diagnosis. Existing approaches, typically dependent on detecting the presence of anatomical structures as defined by clinical protocols, have been constrained by a lack of consideration for image perceptual quality. Although supervised training with manually labeled quality scores seems feasible, the subjective nature and unclear definition of these scores make such learning error-prone and manual labeling excessively time-consuming. In this paper, we present an unsupervised ultrasound image quality assessment method with score consistency and relativity co-learning (CRL-UIQA). Our approach generates pseudo-labels by calculating feature distribution distances between ultrasound images and high-quality standard planes, leveraging consistency and relativity for training regression networks in quality prediction. Extensive experiments on the dataset demonstrate the impressive performance of the proposed CRL-UIQA, showcasing excellent generalization across diverse plane images.

Keywords: Unsupervised Learning · Image Quality Assessment · Ultrasound Image.

1 Introduction

Ultrasound (US) is the primary method for assessing fetal health in prenatal screening due to its low cost and absence of harmful radiation [4]. The prenatal US examination typically includes four procedures: probe scanning, standard plane selection, growth parameter measurement and diagnosis. Among these procedures, the standard plane is crucial as it is the US image containing key anatomical structures (KASs) [21]. Any incorrect or unclear standard planes may result in limitations on the precision of measurements and diagnosis [19]. For example, measurement of abdomen and head circumference based on fetal thalamic standard plane and fetal abdomen standard plane can be used to estimate gestational age and fetal weight [1, 6, 8], and fetal four-chambered heart (4CH) standard plane is the most basic and important plane in fetal cardiac examination [11, 9].

Some researchers attempted to use artificial intelligence (AI) for the intelligent processing of prenatal US examination to reduce the burden on sonographers, which includes the automatic acquisition of standard planes [15, 5, 3, 17]. Rahmatullah et al. [18] propose to use AdaBoost to recognize the presence of stomach and umbilical vein to determine the standard plane of fetal abdomen. Wu et al. [24] utilize a convolutional neural network to classify the presence of KASs in fetal abdomen planes into four categories and combine the regions of interest (ROI) using a cumulative scoring approach for quality scoring. Lin et al. [10] propose to recognize six KASs as well as ROI in the head plane based on Faster R-CNN and use different weight scores for quality scoring based on different importance. Dong et al. [7] achieve quality scoring of fetal 4CH planes by combining image gain, scaling, and KASs.

However, these methods determine the standard plane only by detecting the presence of KASs in the plane and lack the consideration of image perceptual quality, which may lead to the use of blurred standard planes, e.g., with unclearly structures, containing image artifacts, for measurements and diagnosis, which is contrary to the strict quality requirements for measurements and diagnosis [19]. Therefore, selecting an optimal standard plane (best perceptual image quality) from a set of standard planes (US images of all KASs presented) is essential for improving the accuracy of measurement and diagnosis. However, there are two main challenges for such a task: 1) unclear definition of perceptual quality makes such dataset labeling easy to be subjective from different experts; 2) manual labeling requires clinical experts making it excessively time and labor consuming.

In this paper, we propose a score consistency and relativity co-learning framework (CRL-UIQA) for unsupervised US standard plane image quality assessment (IQA), outputting continuous perceptual quality scores for any US images. Instead of using subjective and unclearly defined score labels for supervised training, we generate pseudo-labels by calculating feature distribution distances between US images and a high-quality standard plane set. Then, to encourage the model to focus on capturing features that are robust and stable across different perspectives of the same image quality, we propose to employ weakly augmented views, like horizontal flip, vertical flip, and rotation, and compute the score consistency loss. Such consistency constraint enhances the model's ability to extract invariant features related to image quality. In addition, to strengthen the model's ability to correlate between different image quality features and quality prediction scores, we propose to use strongly augmented views, such as gaussian blur and motion blur, and compute the relativity loss. This relativity loss forces the model to learn to extract features that are closely related to image quality, thus increasing the model's sensitivity to changes in quality features. Therefore, by applying the score consistency and relativity co-learning, our method can adjust

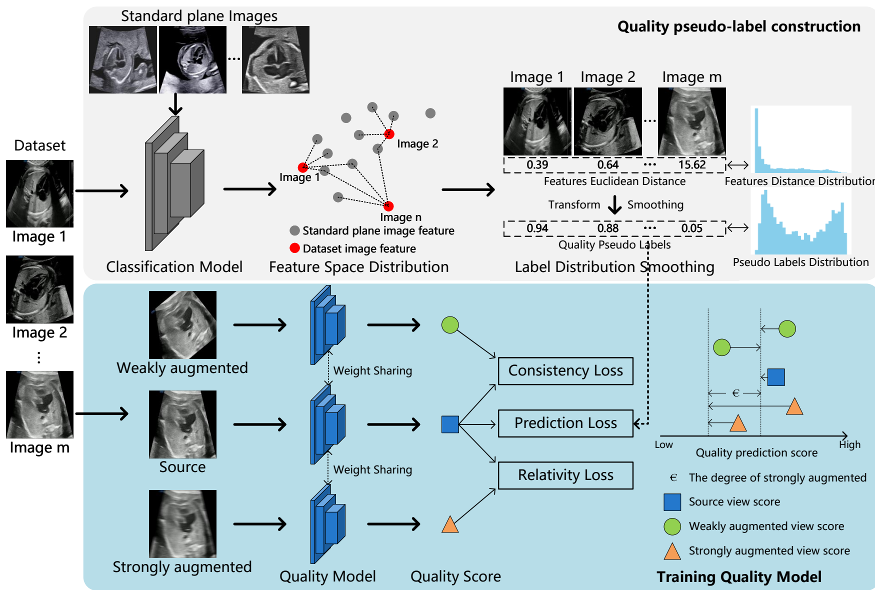

Fig. 1. The CRL-UIQA framework. Step 1: Construct quality pseudo-labels for the image dataset using pre-trained classification models and smooth the label distribution. Step 2: Train quality prediction models using co-learning of score consistency and relativity under the constraint of pseudo-labels.

the pseudo score to a more proper quality-aware score by forcing the model to align with human perceptual quality priors, as shown in the bottom right corner of Figure 1. Our method achieves an accuracy of 0.910 and 0.867 on the 4CH and abdomen datasets, significantly outperforms existing IQA methods, and shows a high correlation between quality prediction scores and US images quality.

2 Method

An overview of the proposed method is shown in Fig. 1. First, we compute the feature distances between the image dataset and the standard plane image set in a pre-trained US image classification model, and construct quality pseudo-labels via the Yeo-Johnson transform to improve the normality of the data distribution. Then, we let the weight-sharing model learn the representations of three views, which include the source, weakly augmented, and strongly augmented views, and train a quality prediction model for fetal US images by regressing the quality pseudo-labels on the source view and by co-learning the consistency of the weakly augmented view and the relativity of the strongly augmented view.

2.1 Quality pseudo-label construction

For images sharing the same content yet having different levels of quality, within the intraclass distribution of features, high-quality images exhibit aggregated feature embeddings, whereas low-quality images exhibit dispersion around the boundary [12]. Inspired by this, we adopt the Euclidean distance between the image dataset and the high-quality standard plane in the feature space distribution as the basis for constructing the quality pseudo-label. Specifically, we define the image dataset $X = \{x_i | i = 1, 2, ..., m\}$, the set of high-quality standard planes $G = \{g_j | j = 1, 2, ..., n\}$, where m and n denote the number of images in the dataset and high-quality standard planes, and the embedding function $P(\cdot)$ of the pre-trained classification model. Hence, for each image in the dataset, we compute the distance in feature space with n high-quality standard planes:

$$D_{x_i} = \{d_j | < P(x_i), P(g_j) >, j = 1, 2, ..., n\}, \tag{1}$$

where $< \cdot >$ denotes the Euclidean distance between feature pairs. To avoid bias caused by chance factors such as outliers in the feature space distribution and to ensure positive correlation between image quality and quality labels, we further choose the top-k smallest distance in the sequence D_{x_i}, and compute the quality pseudo-label as $s_i = -mean(top(D_{x_i}, k))$.

Then, in order to avoid the long-tailed distribution of pseudo-labels that causes the model to tend to optimize for high-frequency labels and ignore the learning of low-frequency labels, we smooth the label distributions using the Yeo-Johnson transform [25], which is not subject to any data constraints and improves the normality of the data distribution. Finally, the pseudo-labels are regularized in the range of (0, 1), where 1 and 0 denote the pseudo-labels of the highest and lowest quality images, respectively. Thus, the new quality pseudo-label q_{x_i} for each image in the dataset is defined as:

$$\hat{S} = \text{Yeo-Johnson}(S), \tag{2}$$

$$q_{x_i} = \frac{\hat{s}_i - \min(\hat{S})}{\max(\hat{S}) - \min(\hat{S})}, \tag{3}$$

where $S = \{s_i | i = 1, 2, ..., m\}$ and $\hat{S} = \{\hat{s}_i | i = 1, 2, ..., m\}$.

2.2 Score consistency and relativity co-learning

Using the constructed quality pseudo-labels, we can train a model for continuous quality assessment of fetal US images that is capable of outputting image quality scores end-to-end and does not rely on the computation of feature distances between images and high-quality standard planes. Intuitively, we first construct the following loss function to force the model to learn the quality pseudo-label:

$$L_d = \frac{1}{m} \sum_{i=1}^{m} (q_{x_i} - p_{x_i})^2, \tag{4}$$

where p_{x_i} is the image quality model prediction result.

However, considering that the pseudo-labels are not accurate and lack the consideration of image perceptual quality, the forced fitting of incorrect pseudo-labels is not sufficient for IQA. Therefore, we further extend each image data into three views: source view, weakly augmented view and strongly augmented view. In model training, we require the quality prediction of the source view and the weakly augmented view to be consistent and ensure that the quality prediction of the source view is better than that of the strongly augmented view, thus improving the robustness of the model on perceptual quality comprehending.

Consistency of Weakly Augmented Views To encourage the model to focus on capturing features that are robust and stable across different perspectives of the same image quality, we generate weakly augmented views using horizontal flip, vertical flip, and rotation of the source view that do not change the image quality, and compute the consistency loss as:

$$L_c = \frac{1}{m}\sum_{i=1}^{m}(\hat{p}_{x_i} - p_{x_i})^2, \tag{5}$$

where $\hat{p}_{x_i}$ denotes the quality prediction score for the weakly augmented views.

Relativity of Strongly Augmented Views In order to reduce the error caused by the deviation of pseudo-labels from the real image quality, we use different degrees of corruption such as gaussian blur and motion blur for each source view x_i to generate strongly augmented views $x_i^{'}$. In this way, we can construct accurate samples with relativity $B = (x_i, x_i^{'}, \epsilon_i)$, where ϵ_i represents the degree of augmentation, and compute the relativity loss, defined as follows:

$$L_r = max((p_{x_i^{'}} - p_{x_i}) + \epsilon_i, 0), \tag{6}$$

where p_{x_i} and $p_{x_i^{'}}$ denote the quality prediction results for the source and strongly augmented views, respectively.
Our model is trained in an end-to-end way with total loss defined as follows:

$$L = \alpha * L_d + \theta * L_c + \beta * L_r, \tag{7}$$

where α, θ and β are trade-off parameters.

Through the co-learning of consistency and relativity, the proposed CRL-UIQA framework is able to learn the fetal US image quality features better and improve the model robustness.

3 Experiments and Results

3.1 Dataset

We obtained ethics committee approval to collect 262 videos of 2D fetal US scans in the middle and late stages of pregnancy in the same type of US equipment, where each 2D video was from a different individual.

Table 1. Ablation study results of our proposed various strategies on the 4CH and abdomen datasets. Results are reported using quality prediction accuracy mentioned in Section 3.2.

Pseudo-label	Distribution Smoothing	L_c	L_r	Dataset	
				4CH	Abdomen
✓				0.857	0.825
✓	✓			0.876	0.836
✓	✓	✓		0.901	0.846
✓	✓		✓	0.836	0.819
✓	✓	✓	✓	**0.910**	**0.867**

Specifically, we randomly collect 8696 4CH images and 7619 abdomen images from 217 videos. We randomly select 6956 images and 6425 images (m) from the two datasets as the training set, and 1740 images and 1194 images as the validation set to observe the training status of the model. Subsequently, three experienced sonographers select 45 (n) high-quality 4CH standard planes and abdomen standard planes in 25 videos. 477 pairs of 4CH images for the 4CH set and 488 pairs of abdomen images for the abdomen test set in another 20 videos, in which each pair of images is clearly defined as to which image has the higher quality.

Note that each of the above sets does not require manual labeling of image quality.

3.2 Implementation Details

We implement fetal ultrasound IQA by introducing an additional classification branch on the backbone of YOLOv5 [22] pretrained for KASs detection, which utilizes a projection layer to reduce the number of feature map channels and feeds pooled features into a fully connected layer to output quality scores. Our approach is implemented on an NVIDIA RTX 3090 24GB GPU using Pytorch. We set the loss trade-off parameters α, θ and β to 0.05, 0.1, and 0.1, respectively, train 35 epochs using an SGD optimizer with an initial learning rate of 0.001, and set the batch size to 16, and use YOLOv5 default parameters for the rest.

For intuitive analysis, we use accuracy as the metric to evaluate the performance of the model. For each pair of US images in the test set, we use the model to predict the quality scores of both and denote the number of correctly differentiated high and low qualities as T, the number of errors as F, and the accuracy is calculated as $\frac{T}{T+F}$.

3.3 Ablation Study

We experimentally investigate the effects of distribution smoothing, consistency loss L_c, and relativity loss L_r on model performance, as shown in Table 1. The accuracy decreases when only L_r is added to the regression training for pseudo-labels. It is because the introduced strong augmented view increases the sensitivity of the model to feature changes, which reduces the ability of the model

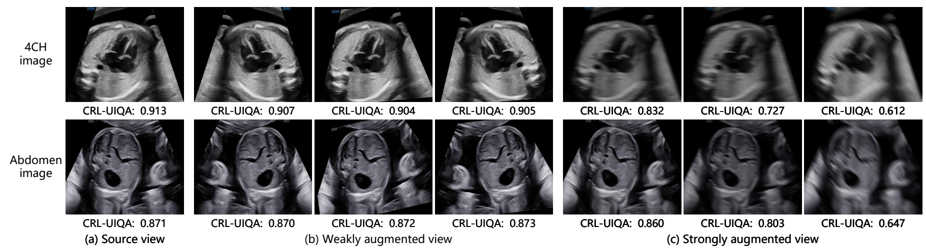

Fig. 2. Quality prediction results of our method for different views, where (a) is the source view, (b) are the weakly augmented views with randomized Flip and Rotation degrees, and (c) are the strongly augmented views with increasing distortion degree from left to right.

to generalize to different features of the same quality level, but this is improved by adding L_c. Overall, the addition of the three strategies enables our model to achieve optimal performance. We show the quality prediction scores of the model for different views of the US image in Fig. 2, illustrating the effectiveness of score consistency and relativity co-learning, where our model successfully gives similar quality scores to the source and weakly augmented views, and lower the quality score to the strongly augmented view.

3.4 Comparison with Other Algorithms

We compare the proposed CRL-UIQA with the traditional IQA methods NIQE [14], BRISQUE [13], and MSSIM [23], as well as the learning-based IQA methods SER-FIQ [20], SDD-FIQA [16], and Weighted score. Weighted score is a scoring system that we soften the existing standard plane cumulative scoring method [10, 24] by using structure confidence as weights [2]. SDD-FIQA* is trained in our framework based on pseudo-labels constructed from SDD-FIQA. Table 2 shows the accuracy of existing IQA methods on US images. BRISQUE, NIQE and MSSIM use natural scene images to construct statistical features to quantify losses such as image distortion, which is not applicable to US images, thus resulting in relatively low accuracy. Our method achieves the best results of 0.910 and 0.867 in both the 4CH and abdominal test sets, respectively, illustrating the excellent performance and outstanding generalization ability of CRL-UIQA.

Fig. 3 shows the quality prediction results of learning-based IQA methods for different quality US images. Relative to other methods, our CRL-UIQA shows higher accuracy and displays a high correlation between quality prediction scores and US image quality. For example, for the second and third images in the first row of (a), the third image has a higher quality score due to the clear display of the left atrium and left ventricle as well as less noise, and for the second and third images in the first row of (b), the second image has a higher quality score due to the perfect display of the stomach and umbilical vein as well as a more completed abdominal wall contour.

Table 2. Comparison with existing quality assessment methods.

	BRISQUE [13]	NIQE [14]	MSSIM [23]	SER-FIQ [20]	SDD-FIQA* [16]	Weighted Score	CRL-UIQA
4CH	0.421	0.497	0.585	0.562	0.702	0.798	**0.910**
Abdomen	0.498	0.355	0.434	0.537	0.607	0.764	**0.867**

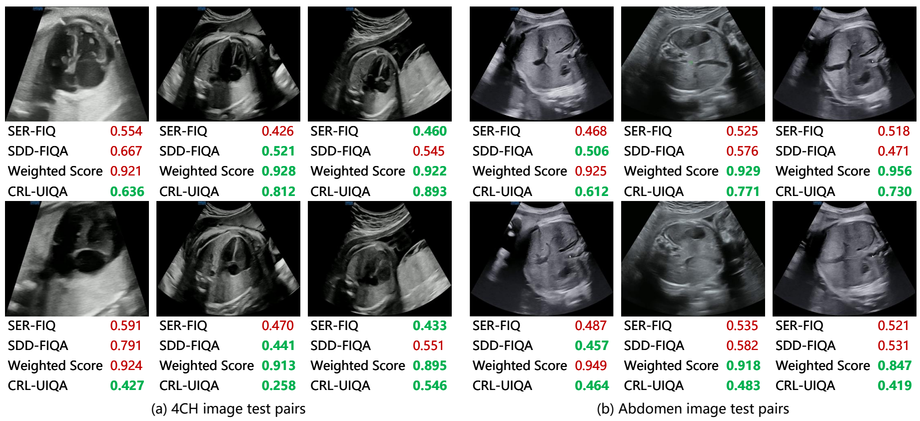

Fig. 3. Quality prediction results of learning-based methods for different quality US images. The two images in the same column are higher quality and lower quality images respectively. We label the scores that can successfully differentiate between high and low quality images by predicting the results as green. Conversely, we mark them as red.

4 Conclusion

In this paper, we propose an unsupervised ultrasound image quality assessment method, CRL-UIQA. Our method constructs pseudo-labels by calculating feature distribution distances between an image and high-quality standard plane images without relying on any annotated information about image quality. Different from the direct regression pseudo-labels approach, we use consistency loss to encourage the model to focus on capturing robust and stable features in different views of the same image quality, and relevancy loss to emphasize the correlation between different image qualities and quality prediction scores. Experimental results on fetal 4CH and abdomen US images demonstrate the effectiveness of the proposed method and demonstrate excellent generalization capabilities over different plane images.

Acknowledgments. This work was supported by the National Key R&D Program of China (grant number 2022YFF0606302), the National Natural Science Foundation of China (grant number 62272159), and the Postgraduate Scientific Research Innovation Project of Hunan Province (grant number QL20230100).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Altman, D., Chitty, L.: New charts for ultrasound dating of pregnancy. Ultrasound in Obstetrics and Gynecology: The Official Journal of the International Society of Ultrasound in Obstetrics and Gynecology **10**(3), 174–191 (1997)
2. Baum, Z.M., Bonmati, E., Cristoni, L., Walden, A., Prados, F., Kanber, B., Barratt, D.C., Hawkes, D.J., Parker, G.J., Wheeler-Kingshott, C.A.G., et al.: Image quality assessment for closed-loop computer-assisted lung ultrasound. In: Medical Imaging 2021: Image-Guided Procedures, Robotic Interventions, and Modeling. vol. 11598, pp. 183–189. SPIE (2021)
3. Baumgartner, C.F., Kamnitsas, K., Matthew, J., Smith, S., Kainz, B., Rueckert, D.: Real-time standard scan plane detection and localisation in fetal ultrasound using fully convolutional neural networks. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19. pp. 203–211. Springer (2016)
4. Bucher, H.C., Schmidt, J.G.: Does routine ultrasound scanning improve outcome in pregnancy? meta-analysis of various outcome measures. British Medical Journal **307**(6895), 13–17 (1993)
5. Chen, H., Dou, Q., Ni, D., Cheng, J.Z., Qin, J., Li, S., Heng, P.A.: Automatic fetal ultrasound standard plane detection using knowledge transferred recurrent neural networks. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part I 18. pp. 507–514. Springer (2015)
6. Degani, S.: Fetal biometry: clinical, pathological, and technical considerations. Obstetrical & gynecological survey **56**(3), 159–167 (2001)
7. Dong, J., Liu, S., Liao, Y., Wen, H., Lei, B., Li, S., Wang, T.: A generic quality control framework for fetal ultrasound cardiac four-chamber planes. IEEE journal of biomedical and health informatics **24**(4), 931–942 (2019)
8. Dudley, N.: A systematic review of the ultrasound estimation of fetal weight. Ultrasound in Obstetrics and Gynecology: The Official Journal of the International Society of Ultrasound in Obstetrics and Gynecology **25**(1), 80–89 (2005)
9. Jeanty, P., Chaoui, R., Tihonenko, I., Grochal, F.: A review of findings in fetal cardiac section drawings: Part 1: The 4-chamber view. Journal of Ultrasound in Medicine **26**(11), 1601–1610 (2007)
10. Lin, Z., Li, S., Ni, D., Liao, Y., Wen, H., Du, J., Chen, S., Wang, T., Lei, B.: Multitask learning for quality assessment of fetal head ultrasound images. Medical image analysis **58**, 101548 (2019)
11. McGahan, J.P.: Sonography of the fetal heart: findings on the four-chamber view. AJR. American journal of roentgenology **156**(3), 547–553 (1991)
12. Meng, Q., Zhao, S., Huang, Z., Zhou, F.: Magface: A universal representation for face recognition and quality assessment. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14225–14234 (2021)
13. Mittal, A., Moorthy, A.K., Bovik, A.C.: No-reference image quality assessment in the spatial domain. IEEE Transactions on image processing **21**(12), 4695–4708 (2012)

14. Mittal, A., Soundararajan, R., Bovik, A.C.: Making a "completely blind" image quality analyzer. IEEE Signal processing letters **20**(3), 209–212 (2012)
15. Ni, D., Li, T., Yang, X., Qin, J., Li, S., Chin, C.T., Ouyang, S., Wang, T., Chen, S.: Selective search and sequential detection for standard plane localization in ultrasound. In: Abdominal Imaging. Computation and Clinical Applications: 5th International Workshop, Held in Conjunction with MICCAI 2013, Nagoya, Japan, September 22, 2013. Proceedings 5. pp. 203–211. Springer (2013)
16. Ou, F.Z., Chen, X., Zhang, R., Huang, Y., Li, S., Li, J., Li, Y., Cao, L., Wang, Y.G.: Sdd-fiqa: unsupervised face image quality assessment with similarity distribution distance. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7670–7679 (2021)
17. Pu, B., Li, K., Li, S., Zhu, N.: Automatic fetal ultrasound standard plane recognition based on deep learning and iiot. IEEE Transactions on Industrial Informatics **17**(11), 7771–7780 (2021)
18. Rahmatullah, B., Sarris, I., Papageorghiou, A., Noble, J.A.: Quality control of fetal ultrasound images: Detection of abdomen anatomical landmarks using adaboost. In: 2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro. pp. 6–9. IEEE (2011)
19. Salomon, L., Bernard, J., Duyme, M., Doris, B., Mas, N., Ville, Y.: Feasibility and reproducibility of an image-scoring method for quality control of fetal biometry in the second trimester. Ultrasound in obstetrics & gynecology **27**(1), 34–40 (2006)
20. Terhorst, P., Kolf, J.N., Damer, N., Kirchbuchner, F., Kuijper, A.: Ser-fiq: Unsupervised estimation of face image quality based on stochastic embedding robustness. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5651–5660 (2020)
21. Timor-Tritsch, I., Monteagudo, A.: Transvaginal fetal neurosonography: standardization of the planes and sections by anatomic landmarks. Ultrasound in Obstetrics and Gynecology: The Official Journal of the International Society of Ultrasound in Obstetrics and Gynecology **8**(1), 42–47 (1996)
22. Ultralytics: YOLOv5: A state-of-the-art real-time object detection system. https://docs.ultralytics.com (2021)
23. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)
24. Wu, L., Cheng, J.Z., Li, S., Lei, B., Wang, T., Ni, D.: Fuiqa: fetal ultrasound image quality assessment with deep convolutional networks. IEEE transactions on cybernetics **47**(5), 1336–1349 (2017)
25. Yeo, I.K., Johnson, R.A.: A new family of power transformations to improve normality or symmetry. Biometrika **87**(4), 954–959 (2000)