Medical Image Analysis 90 (2023) 102977

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Memory-based unsupervised video clinical quality assessment with multi-modality data in fetal ultrasound

He Zhao [a,*], Qingqing Zheng [b], Clare Teng [a], Robail Yasrab [a], Lior Drukker [c,d], Aris T. Papageorghiou [c], J. Alison Noble [a]

[a] Institute of Biomedical Engineering, University of Oxford, United Kingdom
[b] Guangdong Provincial Key Laboratory of Computer Vision and Virtual Reality, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, China
[c] Nuffield Department of Women's and Reproductive Health, University of Oxford, United Kingdom
[d] Department of Obstetrics and Gynecology, Tel-Aviv University, Israel

ARTICLE INFO

MSC:
41A05
41A10
65D05
65D17

Keywords:
Fetal ultrasound
Clinical quality assessment
Multi-modality data

ABSTRACT

In obstetric sonography, the quality of acquisition of ultrasound scan video is crucial for accurate (manual or automated) biometric measurement and fetal health assessment. However, the nature of fetal ultrasound involves free-hand probe manipulation and this can make it challenging to capture high-quality videos for fetal biometry, especially for the less-experienced sonographer. Manually checking the quality of acquired videos would be time-consuming, subjective and requires a comprehensive understanding of fetal anatomy. Thus, it would be advantageous to develop an automatic quality assessment method to support video standardization and improve diagnostic accuracy of video-based analysis. In this paper, we propose a general and purely data-driven video-based quality assessment framework which directly learns a distinguishable feature representation from high-quality ultrasound videos alone, without anatomical annotations. Our solution effectively utilizes both spatial and temporal information of ultrasound videos. The spatio-temporal representation is learned by a bi-directional reconstruction between the video space and the feature space, enhanced by a key-query memory module proposed in the feature space. To further improve performance, two additional modalities are introduced in training which are the sonographer gaze and optical flow derived from the video. Two different clinical quality assessment tasks in fetal ultrasound are considered in our experiments, *i.e.,* measurement of the fetal head circumference and cerebellar diameter; in both of these, low-quality videos are detected by the large reconstruction error in the feature space. Extensive experimental evaluation demonstrates the merits of our approach.

1. Introduction

Ultrasound is widely used to monitor normal fetal development and well-being since it is a radiation-free imaging modality, portable and relatively low-cost (Reddy et al., 2008). During routine obstetric ultrasound scans, a sonographer is tasked with finding standard ultrasound planes to examine anatomical structures, or to measure the size of fetal structures (such as the fetal head circumference, femur length, etc.). Fetal biometry is used to estimate fetal gestational age and to monitor fetal growth (Papageorghiou et al., 2014; Self et al., 2022). Deep learning based methods are widely used in ultrasound image analysis (Fiorentino et al., 2022). However, the acquisition of good planes is highly dependent on the experience of a sonographer, fetal movement and acoustic shadowing, leading to a high intra- and inter-observer variability (Sarris et al., 2012). The importance of quality assessment for obstetric scanning has been emphasized in several studies (Dudley and Chapman, 2002; Salomon et al., 2006; Cavallaro et al., 2018). In practice, the suitability of a still image or video for biometry is ensured by an experienced sonographer manually checking whether all required anatomical structures are visible and whether the view is appropriately magnified. This is time-consuming in real time, and labor-intensive. Thus, automatic clinical quality assessment of fetal ultrasound scanning is desirable. In the literature, previous automated quality assessment algorithms (Wu et al., 2017; Lin et al., 2018, 2019; Dong et al., 2019; Yaqub et al., 2021) are typically image-based methods based on supervised learning, which require extensive annotation of fetal anatomical structures and assume pre-defined image quality criteria, normally based on anatomical appearance. Such methods aim to mimic clinical practice such as those outlined in international guidelines for ultrasound acquisition (Salomon et al.,

* Corresponding author.
E-mail address: he.zhao@eng.ox.ac.uk (H. Zhao).

https://doi.org/10.1016/j.media.2023.102977
Received 30 December 2022; Received in revised form 3 August 2023; Accepted 18 September 2023
Available online 23 September 2023

2019); and lead to an explainable representation (*i.e.*, a human user can understand algorithm performance). However, such an approach demands a lot of human effort due to the heavy annotation requirement, which also limits the generalizability of this type of approach. That is, new labels are required for every clinical quality assurance task. It is noteworthy that hardly any method (Wu et al., 2017; Lin et al., 2018) takes into consideration the temporal information, performing quality assessment only on a single frame or a series of single frames treated independently.

In this paper, we consider video-based clinical quality assessment for fetal biometry. The goal of this work is to check acquired video quality and provide recommendations for retaking scans if necessary. Compared with image-based method, our approach involves temporal information which allows the model to learn a better understanding of the fetal anatomy. It evaluates high-quality ultrasound videos, focusing on the appropriateness of a captured video clip for further clinical analysis, rather than identifying a predefined standard plane. The proposed approach follows a reconstruction-based anomaly detection pipeline with bi-directional reconstruction between the video space and the feature space. Similar to classical anomaly detection task, the intuition is that low-quality samples can be recognized by the large reconstruction error as the low-quality samples are not part of the training dataset. A 3D encoder–decoder pair is designed with decomposition to capture the spatio-temporal information in the video sequence. A key-query memory module is proposed, which stores the intrinsic information of the high-quality data and makes the model more sensitive to low-quality samples. Different from the existing methods in the literature, our approach achieves effective usage of both spatial and temporal information and requires no anatomy-specific annotations.

The contributions of our paper can be summarized as: (1) to the best of our knowledge, our approach is the first attempt to implement ultrasound clinical quality assessment by an unsupervised pipeline without the prerequisite of anatomical annotations; (2) a memory-based bi-directional reconstruction between the video and feature spaces is proposed to learn the discriminative representation for identifying high-quality data; finally, (3) multi-modality data, *i.e.,* optical flow and a gaze map, are engaged with the help of an input generator and an auxiliary prediction branch, respectively, which further improve video quality assessment performance.

This article substantially extends a conference paper (Zhao et al., 2022). Specifically, the current article, includes a new key-query memory module that is shown to enhance the spatio-temporal feature representation. This novel method is evaluated by assessing the quality of video acquisition in two exemplar tasks, *i.e.,* fetal head circumference measurement and fetal cerebellar diameter measurement, which are two essential tasks for gestational age estimation. Extensive ablation studies are well discussed to verify the significance of each component. The remainder of the paper is organized as follows. In Section 2, related work on ultrasound quality assessment and video anomaly detection is reviewed. In Section 3, we detail our approach. Experiments and results are presented and discussed in Section 4. Conclusions are presented in Section 5.

2. Related work

This section covers related technical background literature on visual quality assessment for natural images, clinical ultrasound image quality assessment, and video anomaly detection.

2.1. Visual quality assessment

Natural Image Quality Assessment (IQA) has been well studied in image processing. IQA aims to simulate human perception, which is influenced by image content sharpness, contrast, and illumination (Krasula et al., 2017). Most IQA methods have been developed for natural images, and focus on natural image clarity and noise removal (Wang et al., 2004; Heusel et al., 2017). Visual quality is normally assessed in terms of fidelity and clarity. Several prior works (Hemmsen et al., 2010; Loizou et al., 2006; Sassaroli et al., 2019) concentrate on evaluating this kind of quality for ultrasound images by using similar metrics to those proposed for natural images. In Loizou et al. (2006), statistical and texture analysis are applied to the images, and together with the quality metrics and visual perception are used to evaluate image quality.

2.2. Clinical image quality assessment for ultrasound

Clinical image quality is highly related to the specific application and context. Ultrasound images with low contrast, acoustic shadow, and speckle are categorized as low quality with respect to visual quality. However, such images may still be deemed acceptable to a clinician if sufficient diagnostic information is contained within them. Abdi et al. (2017a) propose a framework for automatic quality assessment of echo data in ultrasound by a regression convolutional neural network. A quality assessment method for cardiac ultrasound image is proposed in Liao et al. (2019) through modeling the label uncertainty in CNNs. The authors in Baum et al. (2021) propose a weighted quality score to accept or reject diagnostic-quality lung ultrasound images, which combines the classification-based quality score and the novelty detection-based quality score by Bayesian model averaging.

In obstetric examination and diagnosis, incorrect fetal biometric measurement may lead to inaccurate fetal gestational age estimation and increase the misdiagnosis risk. Therefore, ensuring high-quality fetal ultrasound acquisition is crucial and important. Several studies have been carried out following this definition of clinical quality in fetal ultrasound (Yaqub et al., 2019; Zhang et al., 2021; Yaqub et al., 2021). Standard plane detection (Chen et al., 2015; Baumgartner et al., 2017; Cai et al., 2018) can also be considered to implicitly assess image quality, typically defined in terms of the detectability of specific structures in an image. Early work that considers this type of approach is Rahmatullah et al. (2011). This uses an AdaBoost classifier with Haar-like features to score fetal ultrasound clinical quality by detecting two landmarks of the fetal abdomen. Zhang et al. (2017) propose a random forest approach to determine the quality of fetal head images with the shape and anatomical features calculated from the head region. An automatic quality assessment method for 2nd trimester fetal ultrasound is proposed in Wu et al. (2017) named FUIQA, which is realized by two deep convolutional neural network models for region extraction and anatomy detection, respectively. Finally, image quality is evaluated by assessing the goodness of depiction of the stomach bubble and the umbilical vein. Lin et al. (2018) proposes a Faster RCNN-based model to evaluate fetal head ultrasound image quality, which is further extended in Lin et al. (2019) by a more specific protocol. A detection branch aims to detect six key anatomical structures, and a classification branch is included to identify the fan-shape area. A quality score is given by a pre-defined protocol based on the key anatomies and area shape. In Abdi et al. (2017a,b), Abdi et al. propose a deep regression model and a recurrent neural network, respectively, for quality assessment of echocardiography. Dong et al. (2019) propose a multi-branch framework for fetal echocardiography quality assessment, where three anatomical structures are localized by a cascading classification and detection network. Additionally, view zoom and gain are assessed by a classification model. A semi-supervised approach is proposed in Gao et al. (2020) to select head planes in low-cost ultrasound probe video with prototype features and metric learning. Saeed et al. (2021) treat the quality assessment task as a measure of image amenability with respect to a specific task by a reinforcement learning model. Although clinical criteria are not necessary for their method, it should be noted that detailed anatomical annotations are required in the training phase. A specified pre-defined protocol and annotated locations of anatomical structures are required in most of the aforementioned methods, which limits generalization to new applications.

2.3. Video anomaly detection

Most of the reported video anomaly detection studies for natural scenes have considered the task of detecting abnormal frames in a surveillance video. Early anomaly detection methods were based on hand-crafted features (Basharat et al., 2008; Cong et al., 2011; Cheng et al., 2015). The most common deep learning-base approaches are based on image reconstruction and attempt to reconstruct normal frames and identify events with large reconstruction errors as anomalies. Autoencoders (Hasan et al., 2016; Zhao et al., 2017; Gong et al., 2019; Park et al., 2020) are widely adopted in this kind of method for image reconstruction. In Zhao et al. (2017), a spatio-temporal autoencoder is utilized to learn a video representation and extract features by frame reconstruction and prediction branches. A temporally-coherent sparse coding-based anomaly detection method is proposed in Luo et al. (2017). The sparse coefficients are iteratively updated via a stacked RNN to detect anomalies in videos. Wang et al. (2021) propose a multi-path convGRU-based frame prediction network to capture temporal relationships and handle informative parts of the frame. Liu et al. (2018) detect anomalies with large reconstruction error between a predicted frame and the corresponding ground truth. However, the aforementioned video anomaly detection methods treat a single frame as a target instead of a video clip. In addition, the image background of these applications is typically static which is totally different from ultrasound videos.

3. Method

Our approach is distinctive from existing ultrasound clinical quality assessment methods in two ways. Firstly, both spatial and temporal information is considered by a spatio-temporal encoder and decoder pair. Secondly, our approach is a generalizable method that does not need anatomy-specific annotation and pre-defined protocol.

We formulate the task as an anomaly detection problem, where low-quality video is regarded as the anomalous data. Denote the training dataset as $\mathcal{D} = \{x_1, \dots, x_N\}$ with N high-quality training samples only and a test dataset $\mathcal{D}_t = \{(x_{t_1}, y_1), \dots, (x_{t_M}, y_M)\}$ where $y_i = 0$ indicates low quality video clips and $y_i = 1$ indicates high quality ones. Our goal is to learn the distribution of high-quality video in the training set $\mathcal{D}$. This will allow to identify the low-quality video in a given test set $\mathcal{D}_t$ as anomalous. To achieve this, we propose a spatio-temporal encoder and decoder pair with a memory module to fully exploit the value of high-quality videos via a bi-directional reconstruction between the video space and the feature space.

As illustrated in Fig. 1, the video is first processed by an input processing module (IPM) which outputs a region of interest (ROI) for each frame in the video $x_z \in \mathbb{R}^{W\times H\times 1\times Q}$ and an optical flow map $x_o \in \mathbb{R}^{W\times H\times 1\times Q}$, where W and H denote the input video spatial size and Q refers to the number of frames in the video. The ROI-extracted video x_z and corresponding optical flow map x_o are input to model training. A spatio-temporal encoder G_e and decoder G_d pair is adopted to learn the spatio-temporal features from bi-directional reconstruction between the video and feature spaces with adversarial learning. The bi-directional information flow between the two spaces provides feedback for model training and allows the high-quality data representation to be discriminated from low-quality data. A more discriminative representation is learned due to the proposed key-query memory module that stores the intrinsic information about the high quality data. An auxiliary branch for gaze prediction is included in the video space to mimic where a sonographer looks. Our model is trained in an end-to-end manner, following a similar scheme as described in Zhu et al. (2017), where translations between two domains are performed. In the training stage, one forward process includes video reconstruction and feature reconstruction, where the raw video and the sampled features are inputs to our model and the corresponding generated videos and reconstructed features the outputs. In the inference stage shown in Fig. 1(d), the feature reconstruction error is used as an indicator to identify the low-quality data with a large reconstruction error. In the following sections, we introduce the key components of our model in more detail.

3.1. Input processing module

The proposed input processing module (IPM) is shown in Fig. 1(b) and consists of two parts for ROI extraction and anatomical structure displacement estimation.

ROI extraction. In obstetric ultrasound scans, curvilinear ultrasound transducer is commonly used, meaning the field of view is a fan-shaped area. Maternal tissues may occupy large regions of this area, which does not contribute to the clinical quality of the video focusing on the fetus. Maternal tissues can mislead the model as the background contributes to the overall reconstruction error. Inspired by Jaderberg et al. (2015), we propose a ROI extraction unit that uses a spatial transformer network to localize the region of interest. The unit shown in Fig. 1(b) contains two parts: a backbone network for affine parameters learning and a grid generator $S(\theta)$ for sampling grid generation. The inception-V1 (Szegedy et al., 2015) is utilized as the backbone network to learn parameters of the affine transformation. A sampling grid used for producing transformed outputs is generated by the grid generator $S(\theta)$ based on the learned parameters. Finally, the sampling grid is applied to each input video frame to create a spatial region of interest with high field-of-view occupancy. Different from Jaderberg et al. (2015), our ROI extraction unit is pre-trained independently as a detection task by minimizing the difference between the ROI and the approximate region surrounding a fetal structure at the pixel level. The ROI extraction unit parameters are fixed in model training and testing. Using a pretrained and fixed unit enables the model to prioritize the reconstruction and learn the representation of high-quality data rather than optimizing fetal structure localization.

Optical flow generator. We use the Farneback algorithm[1] (Farnebäck, 2003) to generate optical flow. The optical flow is derived from the video but it is considered as a additional modality as it describes movement rather than spatial patterns. This optical flow captures the displacement patterns of anatomical structures in the ultrasound video. However, the background structures in ultrasound videos can have a large diversity and the speckle can change with the movement of tissues (Prabhu et al., 2014), which can make it difficult to accurately capture these displacement patterns. To address this issue, we introduce pre-processing with a 2D median filter, which is a simple but effective technique for reducing unrelated movement. This helps improve the accuracy of the displacement patterns captured by the optical flow. This pre-processing step helps the Farneback algorithm to focus on capturing the useful displacement of anatomical structures rather than being influenced by speckle and background changes. Fig. 2 shows that obvious fetal head structures can be captured when including the pre-processing step, while the optical flow-based displacement map without pre-processing generates spurious displacements.

3.2. Adversarial bi-directional reconstruction with spatio-temporal encoder–decoder

As shown in Fig. 1, our model takes the video and optical flow as inputs and outputs reconstructed video as well as the gaze prediction. There are two directional reconstruction processes between the video space ($\mathcal{V}$) and the feature space ($\mathcal{F}$) assisted by adversarial learning, which are the video space circle $\mathcal{V} \rightarrow \mathcal{F} \rightarrow \mathcal{V}$ and the feature space circle $\mathcal{F} \rightarrow \mathcal{V} \rightarrow \mathcal{F}$, respectively. The encoder $G_e : \mathcal{V} \rightarrow \mathcal{F}$ and

[1] This algorithm is chosen for its low computational cost, while recent deep learning based method might be suitable but without finding superior results.

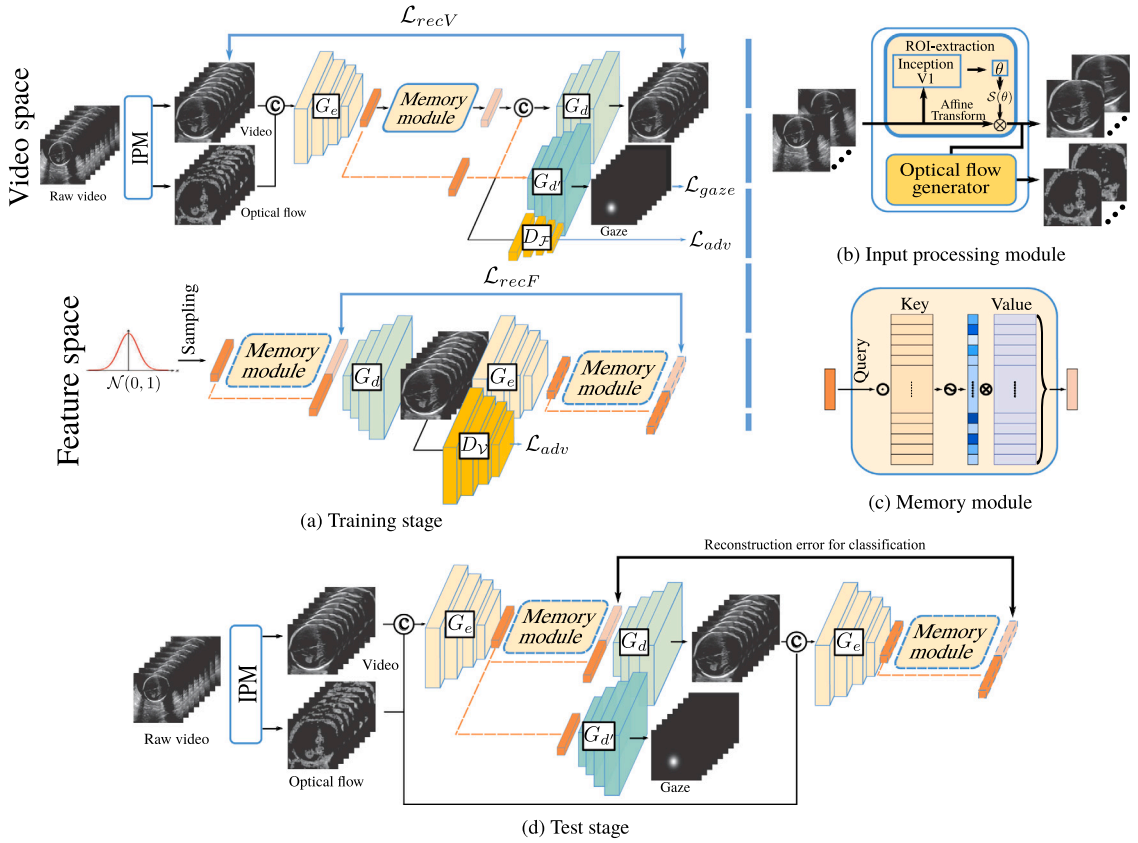


Fig. 1. Flowchart of our approach. (a) Training stage with a bi-directional reconstruction loop in video and feature spaces, where © refers to a concatenation operation. Note, the encoder G_e takes two modalities of video and optical flow as input, and two decoders G_d, G'_d output reconstructed video and predicted gaze, respectively. (b) The input processing module which includes ROI extraction and optical flow map generation. In the ROI extraction module, the sampling grid of affine transformation is generated by grid generator $S(\theta)$ with the learned parameter θ. (c) The proposed memory module captures intrinsic feature representation of high-quality data by the key-query scheme. $Value$ matrix stores prototypical features and Key matrix is the index for retrieval. (d) Test stage for ultrasound quality assessment, where the feature reconstruction error is used as a criterion to identify low-quality videos.

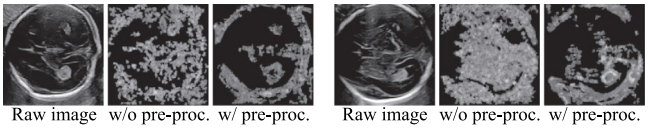


Fig. 2. Two examples of optical flow (OF) generated by the Farneback algorithm. From left to right of each panel are raw images, OF generated without pre-processing, OF generated with pre-processing.

decoder $G_d : \mathcal{F} \rightarrow \mathcal{V}$ build the bridge between two spaces. To help the model generate realistic high-quality data, two discriminators $D_\mathcal{V}$ and $D_\mathcal{F}$ are proposed. Specifically, $D_\mathcal{V}$ is employed to distinguish videos generated by G_d from the real high-quality data. $D_\mathcal{F}$ is utilized to identify features generated by G_e from the real features sampled from a multivariate Gaussian distribution. The bi-directional reconstruction makes our model gain a better understanding of the high-quality data by exploring the information from both spaces.

Encoder and decoder. The encoder G_e and decoder G_d are 3D CNN models, which fully exploit the spatial and temporal information. We prefer to decompose the temporal and spatial information, as this means the model is easier to optimize compared with a pure spatio-temporal CNN solution (*i.e.*, R3D (Tran et al., 2015)). With the spatial convolution in advance, redundant spatial information can be eliminated, leaving only the crucial spatial information to be combined with temporal information to compose the representation of high-quality data. The bottleneck feature output by G_e is concatenated with the feature retrieved from the memory module $\mathcal{M}$ and serves as input to the decoder G_d. Our encoder G_e consists of eight 3D convolutional layers. The first five layers are each only for spatial convolution, for which the kernel size is $1 \times 4 \times 4$ and the stride is $1 \times 2 \times 2$. The other three convolutional layers merge the spatial and temporal information together leading to a bottleneck feature of size C. The spatio-temporal convolutional layers have a kernel size $4 \times 4 \times 4$ and stride $2 \times 2 \times 2$. The decoder G_d has a symmetrical structure to the encoder but uses deconvolutional layers where the first three deconvolutional layers perform spatio-temporal deconvolution and the following five layers reconstruct spatial information.

Discriminators. The video space discriminator $D_\mathcal{V}$ is similar to PatchGAN (Li and Wand, 2016) but with a spatio-temporal convolutional structure. There are five layers, where the first layer is a spatial convolution with kernel size $1 \times 3 \times 3$ and stride $1 \times 2 \times 2$ and the following three layers are spatio-temporal convolutions with kernel size $3 \times 3 \times 3$ and stride $2 \times 2 \times 2$ followed by a spatial convolutional layer with kernel size $1 \times 3 \times 3$ and stride 1. The feature space discriminator $D_\mathcal{F}$ is a multi-layer perceptron network (MLP) with six fully-connected layers which have a neuron size from 512 to 1.

Gaze prediction branch. Eye-tracking data synchronized with the video was available in our study. This shows the sonographer gaze locations during scanning. We can use this for gaze prediction. Trying to predict gaze forces the model to learn salient regions of interest in high-quality video. We propose a multi-task branch for gaze prediction, where an

auxiliary decoder $G_{d'}$, sharing the same structure with G_d, is employed to learn the gaze map from the videos supervised by gaze ground truth. Compared with serving as an additional input, the gaze utilized as a prediction output has two advantages. It eliminates the requirement of gaze in the test phase and enables the model to provide a valuable signal to sonographers on where to look.

3.3. Key-query memory module

In principle, the reconstruction-based method models the high-quality data by obtaining an encoding feature representation which preserves the most important information of high-quality data. It forces the model to learn typical patterns of high-quality data in training. In order to enhance this idea, we propose to use a key-query memory module $\mathcal{M}$ to store the typical high-quality information, *e.g.,* the anatomical key structure referring to high-quality data. The intuition of using the memory is to mimic the process of quality assessment by sonographers. Like how sonographers identify high-quality samples by the critical anatomical structures in the video, the memory module remembers and focuses on the relevant features of high-quality samples during training. By integrating the proposed memory module, the retrieved features in turn help the model differentiate between high-quality and low-quality samples in the test phase. The proposed module not only improves the accuracy of our approach but also provides an interpretable analysis of high-quality data.

The proposed memory module $\mathcal{M}$ consists of two components: a memory to store typical high-quality information and an index to retrieve the most relevant prototype from this stored information. As shown in Fig. 1(c), the *Value* matrix is the memory that stores all the prototypical vectors and the *Key* matrix is the index used to retrieve prototypes from *Value* matrix. The high-quality related features extracted from $\mathcal{M}$ are concatenated with encoding feature representation of input sample and fed to the decoder G_d to reconstruct the video.

The memory (*Value*) is defined as a matrix $\boldsymbol{V} \in \mathbb{R}^{M\times C}$ that contains M prototypical feature vectors with dimension C. The index (*Key*) is defined as $\boldsymbol{K} \in \mathbb{R}^{M\times C}$, which control the information retrieval process. In our approach, the dimension C is the same as the encoding feature representation from G_e. Each element of $\boldsymbol{V}$ denotes a memory item as a row vector $\boldsymbol{v}_{i,i\in\{1,\dots,M\}}$. Given a query encoding vector from the input sample $\boldsymbol{q} \in \mathbb{R}^{1\times C}$, the retrieved representation from memory module is calculated based on a weighted summation:

$$\boldsymbol{z} = \sum_{i=1}^{M} \omega_i \boldsymbol{v_i}, \tag{1}$$

where ω_i is the weight coefficient and $\sum_1^M \omega_i = 1$. The weight is used for accessing the memory by calculating the similarity between the query vector $\boldsymbol{q}$ and the key vector $\boldsymbol{k}_i \in \mathbb{R}^{1\times C}$ that is a row vector representing each element of K. Similar to the attention score in Vaswani et al. (2017), we compute the weight coefficient ω_i with a scaled softmax operation:

$$\omega_i = \frac{\exp(\frac{1}{\sqrt{C}} sim(\boldsymbol{q}, \boldsymbol{k}_i))}{\sum_{j=1}^{M} \exp(\frac{1}{\sqrt{C}} sim(\boldsymbol{q}, \boldsymbol{k}_j))}, \tag{2}$$

where $sim(\cdot)$ indicates similarity function, which is defined as:

$$sim(q, k_i) = \frac{\boldsymbol{q} \cdot \boldsymbol{k}_i}{\|\boldsymbol{q}\| \, \|\boldsymbol{k}_i\|} \tag{3}$$

The scaling factor $\frac{1}{\sqrt{C}}$ is adopted to reduce the large magnitude by dot product due to the larger encoding representation size C, which may lead to a hard training process with an extremely small gradient.

In the training phase, the memory module is optimized by back-propagation of the video space reconstruction, which supervises the module to record the most representative features in the high-quality video. In the test phase, the most relevant features representing high-quality video are retrieved for image reconstruction. The reconstruction is carried out by using both the retrieved features from memory module $\mathcal{M}$ and the encoding feature representation from encoder G_e. Therefore, the reconstruction tends to be close to the high-quality data, leading to small reconstruction errors for high-quality samples and large errors for low-quality samples which appear similar to high-quality samples. By using the memory module, the model generates high-quality related feature representations that increase its sensitivity to detect low-quality data.

3.4. Objective functions

Training is supervised by the adversarial bi-directional reconstruction and multi-task gaze prediction. The encoder G_e and decoder G_d are alternatively optimized with discriminators $D_{\mathcal{V}}$ and $D_{\mathcal{F}}$. Specifically, the model is trained to solve the following optimization function:

$$\min_G \max_D \mathcal{L} = \omega_{adv}\mathcal{L}_{adv} + \omega_{rec}\mathcal{L}_{rec} + \omega_{gaze}\mathcal{L}_{gaze}, \tag{4}$$

where $\mathcal{L}_{adv}$, $\mathcal{L}_{rec}$, $\mathcal{L}_{gaze}$ indicate adversarial loss, bi-directional reconstruction loss, and gaze loss, respectively, and ω_* refers to the corresponding loss weights.

3.4.1. Adversarial loss

Two discriminators are proposed to distinguish the generated videos or features from the real ones. These discriminators aid the model to obtain a better reconstruction of high-quality data. A least-squares adversarial loss is used to optimize the discriminators:

$$\min_{D_{\mathcal{V}}} \mathcal{L}_{D_{\mathcal{V}}} = \frac{1}{2}|D_{\mathcal{V}}(G_d([\boldsymbol{f}, \mathcal{M}(f)]))|^2 + \frac{1}{2}|D_{\mathcal{V}}(x_z) - 1|^2, \tag{5}$$

and

$$\min_{D_{\mathcal{F}}} \mathcal{L}_{D_{\mathcal{F}}} = \frac{1}{2}|D_{\mathcal{F}}(G_e(x_z, x_o))|^2 + \frac{1}{2}|D_{\mathcal{F}}(\boldsymbol{f}) - 1|^2, \tag{6}$$

where x_z, x_o are the ROI-extracted videos and corresponding optical flow map, respectively; $\boldsymbol{f}$ is the feature vector sampled from a multivariate Gaussian distribution as in Kingma and Welling (2014); and $[\cdot,\cdot]$ refers to the concatenation operation. The encoder G_e and decoder G_d can benefit from the adversarial learning by optimizing the following formulation:

$$\mathcal{L}_{adv} = \frac{1}{2}|D_{\mathcal{F}}(G_e(x_z, x_o)) - 1|^2 + |D_{\mathcal{V}}(G_d([\boldsymbol{f}, \mathcal{M}(f)])) - 1|^2. \tag{7}$$

3.4.2. Bi-directional reconstruction loss

The bi-directional reconstruction loss allows the encoder and decoder to fully exploit the spatio-temporal representations of high-quality videos for a realistic reconstruction result by learning the information flows between the video space and feature space. For video reconstruction, we utilize the structure similarity (SSIM) (Wang et al., 2004) whereby the average SSIM score over all video frames defines the video SSIM score. Namely,

$$SSIM(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{Q}\sum_{i=1}^{Q} \frac{(2\mu_{x_i}\mu_{y_i} + c_1)(2\sigma_{x_i y_i} + c_2)}{(\mu_{x_i}^2 + \mu_{y_i}^2 + c_1)(\sigma_{x_i}^2 + \sigma_{y_i}^2 + c_2)}, \tag{8}$$

where x_i and y_i are the input and reconstruction, respectively; μ_* and σ_* refer to the mean intensity and the standard deviation of image, respectively and σ_{xy} denotes the covariance of images. The constants c_1 and c_2 are set to 0.01 and 0.03, respectively. The video reconstruction loss $\mathcal{L}_{recV}$ is defined as:

$$\mathcal{L}_{recV} = 1 - SSIM(x_z, G_d([f_z, \mathcal{M}(f_z)])), \tag{9}$$

where $f_z = G_e(x_z, x_o)$ is the feature representation generated by encoder G_e. Compared with the pixel-wise L1 loss used for image reconstruction, the SSIM loss focusing on the image content is less sensitive to pixel shifts. This attribute enables the model to learn clinical quality information related to anatomy rather than just pixel intensity during reconstruction. It also makes our approach for ultrasound clinical quality assessment less sensitive to speckle.

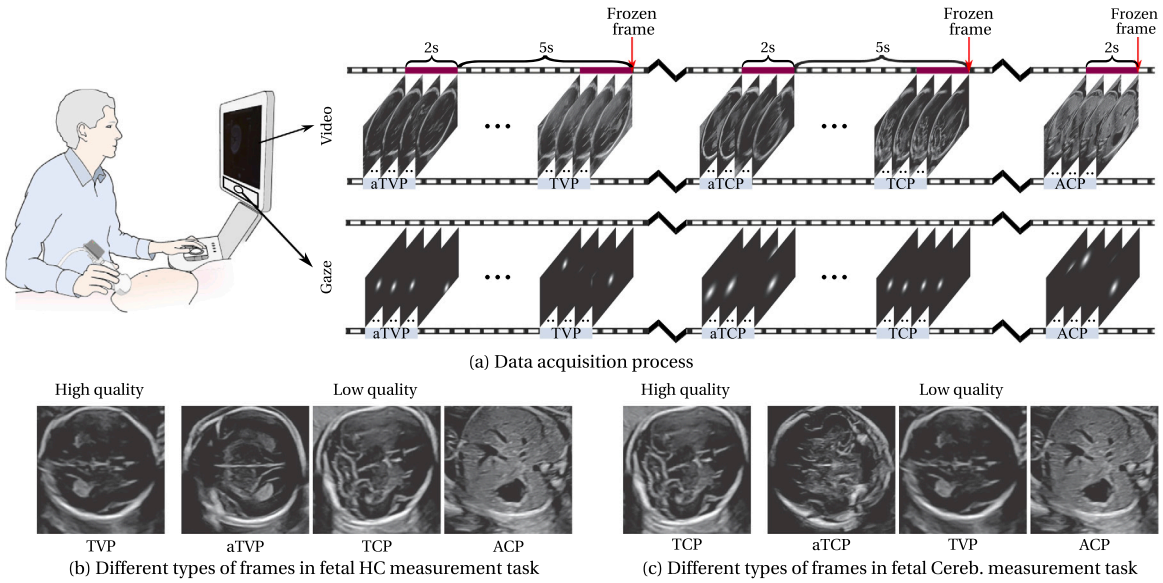


Fig. 3. Illustration of the data acquisition process and examples of high- and low-quality frames for two different tasks. (a) The process of data acquisition. Video and gaze are captured simultaneously, where two-second video clips are used as training or test samples based on frozen frames of different types. The video with aTVP/aTCP type is obtained five seconds before the frozen frame, which stands for approaching the TVP/TCP plane. (b) Exemplar frames for the fetal HC measurement task (High-quality video: TVP); (c) Exemplar frames for the fetal cerebellar measurement task (High-quality video: TCP).

For feature reconstruction, we consider the encoding feature $\boldsymbol{f}$ as well as the feature $\boldsymbol{z}$ generated from memory module together. In this case $\mathcal{L}_{recF}$ is defined as:

$$\begin{aligned}\mathcal{L}_{recF} &= \|G_e(G_d([\boldsymbol{f}, \mathcal{M}(\boldsymbol{f})]), x_o) - \boldsymbol{f}\|_1 \\ &+ \|\mathcal{M}(G_e(G_d([\boldsymbol{f}, \mathcal{M}(\boldsymbol{f})]), x_o)) - \mathcal{M}(\boldsymbol{f})\|_1,\end{aligned} \tag{10}$$

where $[\cdot,\cdot]$ indicated the concatenation operation. Finally, the bi-directional reconstruction loss $\mathcal{L}_{rec}$ is defined as:

$$\mathcal{L}_{rec} = \mathcal{L}_{recV} + \mathcal{L}_{recF}. \tag{11}$$

3.4.3. Gaze loss

An auxiliary loss is introduced by the gaze prediction task. The gaze loss function is utilized for the model to learn the gaze saliency map. This minimizes the difference between the gaze prediction map and the ground-truth by a simple L1 loss. The gaze loss $\mathcal{L}_{gaze}$ is defined as:

$$\mathcal{L}_{gaze} = \|G_{d'}(G_e(x_z, x_o)) - x_g\|_1, \tag{12}$$

where x_g is the eye gaze ground truth.

4. Experiments

In this section, we evaluate our approach on two fetal anatomy biometry tasks: head circumference measurement, and cerebellar measurement. For each task, the high-quality video is defined as the video clip containing frames which are suitable for biometry. Next we describe the dataset used in our study and the experimental configuration.

4.1. Datasets

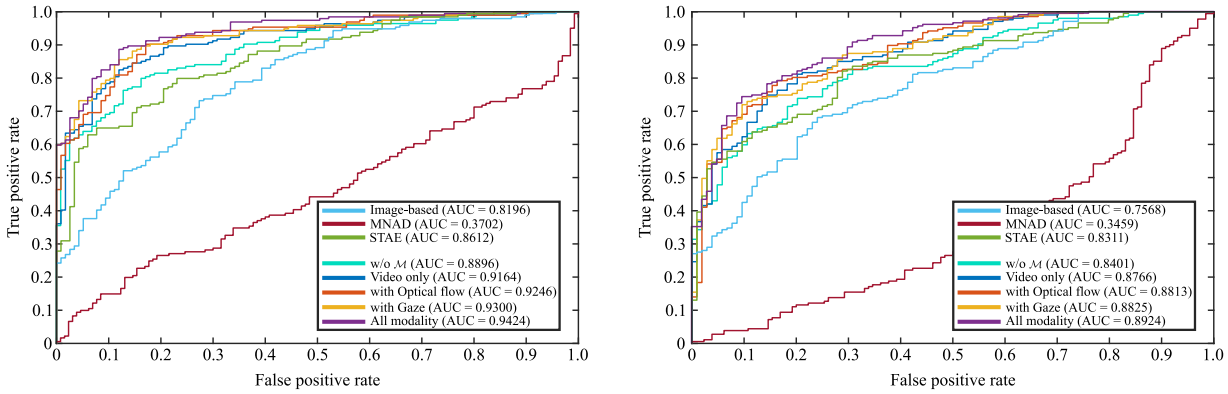
The data used in our experiments are from an existing study PULSE (Drukker et al., 2021), which is approved by the UK Research Ethics Committee (Reference 18/WS/0051). Full-length ultrasound videos were recorded by a free-hand acquisition protocol on a GE Voluson E8 scanner at 30 Hz. Simultaneously, gaze data was acquired using a Tobii Eye Tracker 4C. In total, our dataset consisted of scans from 430 subjects. Videos had a frame resolution of 1008 × 784. The data acquisition process is shown in Fig. 3(a), where an experienced sonographer moves the probe to find and freeze a biometry plane. Video clips containing the frozen frame and 2 s before are labeled with the frozen frame type, *e.g.*, transventricular plane (TVP), transcerebellar plane (TCP), abdominal circumference plane (ACP). Using a two-second video clip as our data sample achieves a balance between performance and computational cost. This clip length provides sufficient information for quality assessment and corresponds to the fine search stage of sonographer scanning that takes place shortly before the frozen frame. Clinically, TVP is used for measuring the fetal head circumference; TCP is obtained for cerebellar measurement; ACP is captured for fetal abdominal circumference measurement. An approaching transventricular/transcerebellar plane (aTVP/aTCP) is defined as the video clip collected 5–7 s before the frozen TVP/TCP frame. Based on the definitions above, we considered modeling the two tasks of fetal head circumference (HC) measurement and fetal cerebellar measurement as follows. For fetal HC measurement (*i.e.*, Fig. 3(b)), the high-quality data is TVP, while low-quality videos are TCP, aTVP, and ACP. For fetal cerebellar measurement (*i.e.*, Fig. 3(c)), the high-quality data is TCP, while low-quality data is TVP, aTCP, and ACP. Each subject scan provided one video clip for each type of input sample. For each of the two biometry tasks, 300 high-quality samples were randomly selected as the training set. The test set consists of 311 samples which are the remaining 130 high-quality clips and 181 low-quality clips randomly selected from 430 subjects. Considering the computational complexity and the content changes of ultrasound frames, we re-sampled the raw video with a sampling rate at 8 Hz and selected 8 frames to form the input samples.

4.2. Experimental settings

The spatial resolution of videos is resized by bi-cubic interpolation to 256 × 256 pixels. The pixel values are standardized to $[-1, 1]$ in each video frame. We employ data augmentation on the spatial dimension including image flipping and contrast adjustment. The parameters of the Farneback algorithm are set to the defaults except the filter size is determined as 3 according to the image size. The median filter used for pre-processing has a large kernel size of 21 to reduce the unrelated pixel displacement. Our approach is implemented in PyTorch v1.10.0. Our model is trained with an Adam optimizer and a decayed learning rate. Following the CycleGAN training scheme (Zhu et al., 2017), the model is trained for 200 epochs with a starting learning rate set to 0.0002,

Table 1
Performance of different methods based on the ROI-extracted videos with the evaluation metric of AUC, F1 (%), ACC (%), SEN (%) and SPE (%) on fetal HC measurement task and fetal cerebellar measurement task, respectively.

			Fetal HC measurement					Fetal Cereb. measurement				
			AUC	F1	ACC	SEN	SPE	AUC	F1	ACC	SEN	SPE
Single modality	Image-based		0.8196	83.61	77.09	93.69	49.57	0.7568	82.83	72.67	**99.03**	20.19
	MNAD (Park et al., 2020)		0.3702	73.40	58.52	**98.34**	3.08	0.3459	73.36	58.20	98.90	1.54
	STAE (Zhao et al., 2017)		0.8612	83.13	77.81	87.63	64.57	0.8311	83.32	75.56	91.43	43.91
Multiple modalities	Our approach	Video w/o $\mathcal{M}$ (Zhao et al., 2022)	0.8896	84.88	80.06	89.69	64.10	0.8401	83.87	75.88	94.20	44.74
		Video only	0.9164	88.27	85.21	89.18	78.63	0.8766	85.65	79.10	93.72	50.00
		with Optical flow	0.9246	89.69	87.14	89.69	82.91	0.8813	86.61	80.71	93.72	54.81
		with Gaze	0.9300	89.64	87.14	89.18	83.76	0.8825	86.20	79.10	98.07	41.37
		All modality (ours)	**0.9424**	**90.34**	**88.10**	89.18	**86.32**	**0.8924**	**88.38**	**83.60**	93.72	**63.46**



(a) Fetal HC measurement (b) Fetal Cereb. measurement

Fig. 4. Receiver operating characteristic (ROC) curves of the comparison methods on two different fetal measurement tasks.

which linearly decays to 0 in the last 100 epochs. The loss weights ω_{adv}, ω_{rec} and ω_{gaze} are empirically set to 1, 10 and 10, respectively.

We choose the area under the receiver operating characteristic curve (AUC), F1-score, accuracy, sensitivity and specificity as the evaluation metrics. For reference, the low-quality video is defined as positive sample and the high-quality video is defined as negative sample. The overall performance of quality assessment is judged by AUC and F1-score, while sensitivity and specificity reflect the ability to correctly detect positive and negative samples. The threshold to calculate the scores is determined based on the best value of F1-score.

4.3. Experimental results

We benchmark our approach with other methods in the literature for the two biometry tasks. As there is no existing clinical quality assessment method can be trained without anatomical annotations, we compare our approach with the most related video anomaly detection methods. Specially, we compare with two technically similar approach, (1) a Spatio-Temporal Auto-Encoder (STAE) (Zhao et al., 2017) and (2) MNAD (Park et al., 2020); (3) an image-based version of our approach which only takes the last frozen frame of video as input; and (4) our variant without memory module (Zhao et al., 2022) (Video w/o $\mathcal{M}$). STAE (Zhao et al., 2017) is a video space-only method to detect anomalous cases by learning the reconstruction of videos alone. MNAD (Park et al., 2020) is a video anomaly detection method which detects anomalous frames in a video. In contrast to our work, it does not consider the video as a whole sample. Given the different setting of the current work, we suppose that their model would not achieve satisfactory results with respect to other compared methods.

Fetal HC measurement task. We first evaluate our approach on the fetal HC measurement task where the TVP video is regarded as high-quality data. The results shown in Table 1 indicate that the proposed approach with multi-modality data has the best performance in terms of AUC and F1-score. Note that here is a significant performance gap between the image-based and video-based methods. These results suggest that temporal information is useful to identify the clinical quality of a specific task. Intuitively this makes sense, as the last frozen frame is not always the best diagnostic frame for biometry. Sometimes, the sonographer will "roll back" the video in the buffer to manually find the best measurement plane.

Among the video-based methods, our approach achieves the best performance with an AUC of 0.9164, which is about 5% higher than STAE. Compared with our earlier paper (Zhao et al., 2022), the proposed memory-based model achieves improvements of 3% and 4.3% in terms of AUC score for the HC and Cereb. measurement tasks, respectively. In addition, the increasing specificity demonstrates that the memory module based model avoids misclassifying high-quality data as low quality. This is because the memory module learns prototypical representations of high-quality samples during training and generates features close to high-quality data regardless of the input. This characteristic results in a model with lower reconstruction error for high-quality data and higher error for low-quality data. The improvement observed for the single modality case clearly indicates the benefits of our bi-directional reconstruction scheme and the proposed memory module. With the aid of the optical flow map and gaze, the AUC further increases from 0.9164 to 0.9424. Our approach also achieves the best specificity of 86.32% and the third highest sensitivity of 89.18%. This demonstrates our approach is able to detect most of the low-quality data while keeping a high true negative detection rate. The ROC curves for the models are shown in Fig. 4. Our approach outperforms the other models. Among all of the variants of our approach, the model with all modalities is the best, especially in the true positive rate axis range of $[0.8, 0.9]$. This corresponds to the most useful range for quality assessment as it achieves a high detection rate of low-quality data.

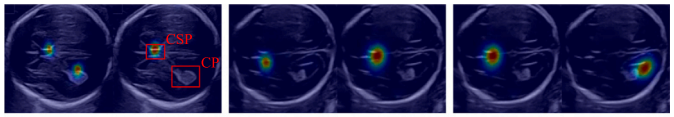
Fetal cerebellar measurement task. We next test our approach for modeling the fetal cerebellar measurement task and treat TCP as high-quality data. The results are shown in the right panel of Table 1, and the

Table 2
Model performance with and without the ROI extraction module for the fetal HC measurement task.

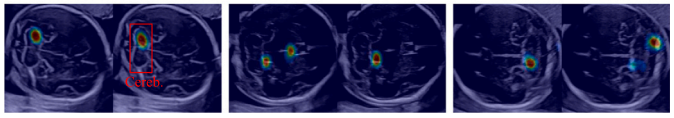
ROI module	AUC	F1	ACC	SEN	SPE
No	0.7048	77.60	68.81	82.82	35.38
Yes	0.9246	89.69	87.14	89.69	82.91

Table 3
Performance of our model with different gaze loss weightings (ω_{gaze}) in Eq. (4).

ω_{gaze}	AUC	F1	ACC	SEN	SPE
0	0.9246	89.69	87.14	89.69	82.91
1	0.9316	89.98	87.46	**90.21**	82.91
10	**0.9424**	**90.34**	**88.10**	89.18	86.32
15	0.9330	89.49	87.46	85.57	**90.60**



(a) Gaze predictions on HC measurement task



(b) Gaze predictions on Cerebellar measurement task

Fig. 5. Three examples of gaze prediction between two consecutive frames for the two fetal measurement tasks. Exemplar anatomical key structures for each tasks are shown with red box, such as CSP and CP for HC measurement task, and Cereb. for cerebellar measurement task.

corresponding ROC curve is displayed in Fig. 4(b). A similar trend to HC measurement is observed with our memory-based bi-directional reconstruction approach achieving the best performance. Compared to HC measurement task, this task's overall performance has decreased, especially the image-based methods with a relatively low specificity indicating that spatial information alone is inadequate for fetal ultrasound quality assessment. The reason for performance degradation is that it is difficult for the model to reconstruct the fetal cerebellar as the structure in the TCP is more complicated than the TVP. This can be observed from Fig. 3. However, our approach still obtains the highest performance with AUC of 0.8924 and F1-score of 88.38%.

4.4. Ablation study

We conducted an ablation study to understand the effect of each component in our model, the parameter settings, and the reconstruction schemes. All the ablation experiments are conducted on the fetal HC measurement task.

4.4.1. ROI extraction

Table 2 reports the performance metrics for our model with and without the ROI extraction module. Observe that there is a significant model performance improvement when the ROI-extraction module is used. With the ROI-extraction module, the AUC increases from 0.7048 to 0.9246 which is a statistically significant improvement (p-value = $0.0003 < 0.05$). The explanation for this is that the affine transformation network enables the model to ignore background tissue and to focus on fetal structures as being critical for image reconstruction. This operation simulates the sonographer assessing the quality, where the crucial part will be zoomed in to make a large field of view followed by diagnosis.

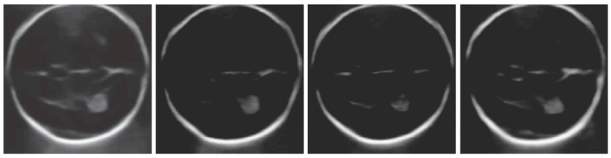


Fig. 6. Reconstructed images from different memory feature vectors stored in the memory bank $\boldsymbol{V}$.

Table 4
Effect of memory module size (Msize) on model performance for the fetal HC measurement task.

Msize	AUC	F1	ACC	SEN	SPE
5	0.9231	88.61	85.53	**90.21**	77.78
15	**0.9246**	**89.69**	**87.14**	89.69	**82.91**
30	0.9239	89.17	86.17	81.24	77.78
60	0.9236	88.32	85.21	89.69	77.78

4.4.2. Gaze prediction

To investigate the influence of the gaze loss we conducted the following experiments with the full model (model with the IPM and memory module) where we varied ω_{gaze} in Eq. (4) to be $0, 1, 10, 15$. Results are summarized in Table 3. When $\omega_{gaze} = 0$, it means no gaze data is used in model training. It is observed that as the gaze loss weight increases, model performance metrics increase until $\omega_{gaze} = 10$. At this value the AUC is 0.9424, which is approximately 2% higher than the model without gaze. Based on the these results, we select $\omega_{gaze} = 10$ as our gaze weight for our experiments.

Gaze predictions with $\omega_{gaze} = 10$ on consecutive frames are shown in Fig. 5. The top row displays results for the fetal HC measurement task, while the bottom shows gaze prediction for the fetal cerebellar measurement task. For the HC measurement task, the sonographer aims to find several anatomical key structures, such as the cavum septi pellucidi (CSP) and choroid plexus (CP). In our predictions, the gaze saliency maps show a high intensity (red spot) on those structures. For the cerebellar measurement task, gaze predictions mainly focus on the edges of cerebellar as well as the CSP.

4.4.3. Memory module

The memory module is based on a key-query mechanism and stores the prototypical representations in a value matrix. To validate this module, we first quantitatively analyze the influence of memory size and qualitatively investigate images generated by different memory feature vectors. Model performance using a memory module with different memory sizes is summarized in Table 4, where the model is trained with optical flow and video modalities. The results show that memory size has little impact on model performance. Indeed, a small memory size can achieve a good result. We attribute that the variance within the high-quality data is not large, allowing a small memory bank to effectively handle the prototypical representations. As for using a large memory bank, such as 30 and 60, we find that it does not improve the model performance, but rather slightly decreases it. we hypothesize that a large bank size (*e.g.*, 30 or 60) with limited prototypical representations may not provide any additional information, thereby failing to improve the performance. Instead, it increases the number of parameters that need to be optimized, leading to a slightly worse result. Based on the experiment results, we choose the memory size as 15 for our memory module. Reconstructed images based on random selected memory feature vectors are shown in Fig. 6. Compared with the real high-quality data (*e.g.*, Fig. 2), the reconstructed images only from the memory feature vectors are smooth with less tissue structures. This means the representations only keep the essential information of the high-quality data and discard the variant information of each input sample, such as the unrelated anatomical structures for high-quality decision. Among images generated by different prototypical

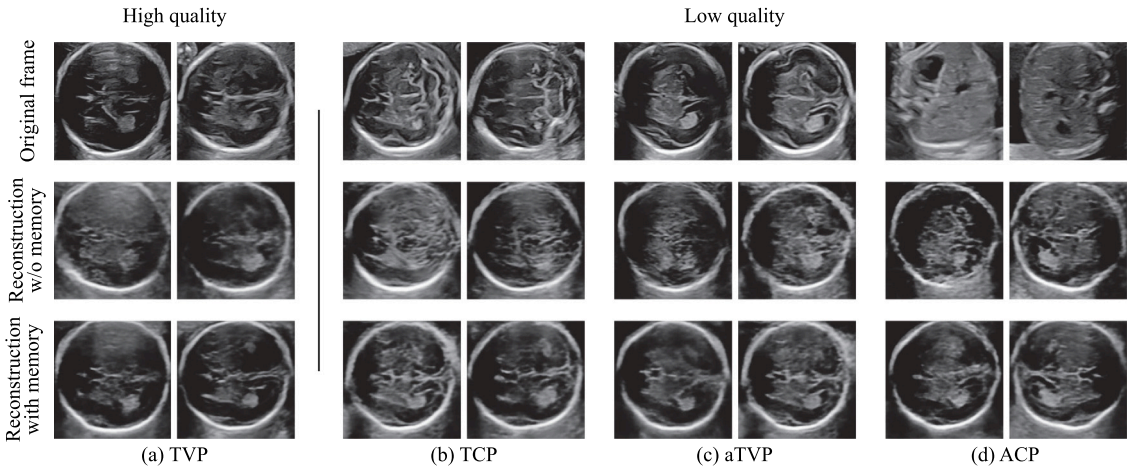


Fig. 7. Image-space reconstruction results of our model for the fetal HC measurement task with respect to the existence of our memory module. Different panels display different types of videos, namely (a) TVP, (b) TCP, (c) aTVP, and (d) ACP, respectively. Two different exemplar frames and their corresponding reconstructions regarding memory modules are shown for each type.

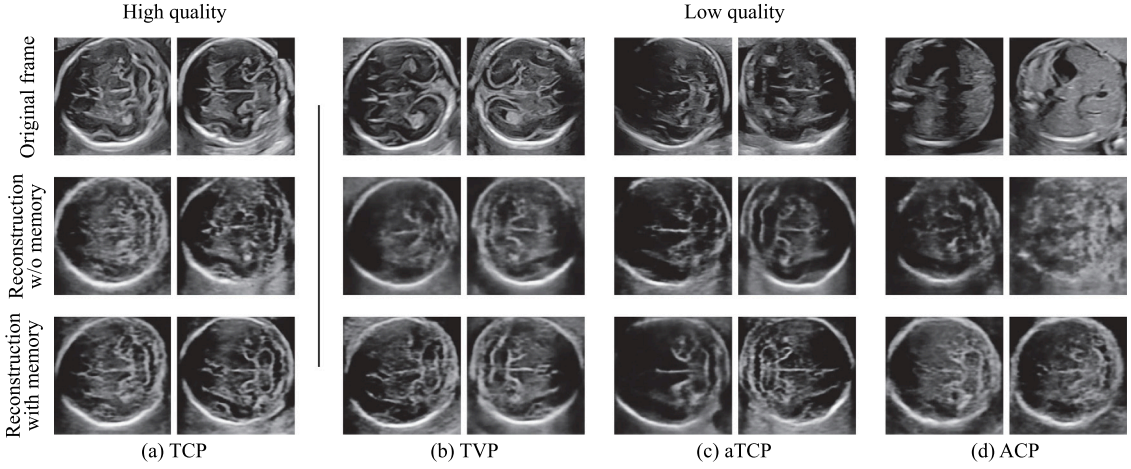


Fig. 8. Image-space reconstruction results of our model for the fetal Cereb. measurement task with respect to the existence of our memory module. Different panels display different types of videos, namely (a) TCP, (b) TVP, (c) aTCP, and (d) ACP, respectively. Two different exemplar frames and their corresponding reconstructions regarding memory modules are shown for each type.

representations, all the reconstructions contain the essential structures that allow for TVP identification. These features include the boundary of head skull, the middle line, and the choroid plexus (CP).

We also compare video reconstruction generated by the model with and without the memory module for the fetal HC measurement task and fetal cerebellar measurement task. Fig. 7 shows the exemplar original frames of high-quality (TVP) and low-quality (TCP, aTVP, ACP) video as well as their reconstructions. Two different examples are shown in each column of each case, while from top to bottom rows are the original frame, the frame reconstructed using the model without the memory module, and the reconstructed frame generated by the model with the memory module. We analyze the superior performance of the proposed memory module in two scenarios: high-quality data reconstruction and low-quality data reconstruction. For the high-quality data (*i.e.*, (a) TVP in our case), the real video frames exhibit clear structures of cavum septi pellucidi (CSP) and a well-defined shape of CP, which are two essential anatomical features for quality assessment by clinicians. The model without memory can generate a blurry head skull similar to the original frame, but the anatomies inside the head are ambiguous, resulting in a large difference not only for human perception but also for feature reconstruction. On the contrary, our memory-based model produces reconstructed images with well-defined anatomies that visually resemble the real high-quality data more closely. As to the low-quality data (*i.e.*, (b) TCP, (c) aTVP, and (d) ACP), our model is able to generate realistic anatomical structures and content information based on the original input frames. Despite the significant difference between the ACP and TVP, our model is able to generate outputs that closely match the key anatomies of a TVP, which indicates the ability of memory module to remember the high-quality related structures. However, the non-memory model produces outputs with poorly defined anatomical structures which can hardly be recognized as high-quality frames.

In Fig. 8, we present the visual reconstruction results of the fetal cerebellar measurement task. Similar to the results shown in Fig. 7, the reconstruction achieved with our memory module exhibits the best performance. For the high-quality data, the reconstruction by our approach successfully maintains clearer and more well-defined cerebellar structures resembling the original frames than the results by non-memory model. It is obvious that the reconstructed results without the memory exhibit less distinguishable structures compared to Fig. 7. This is because the anatomies within TCP are more complex and challenging to be learned using a simple encoder–decoder reconstruction approach. This issue is effectively alleviated by our proposed memory

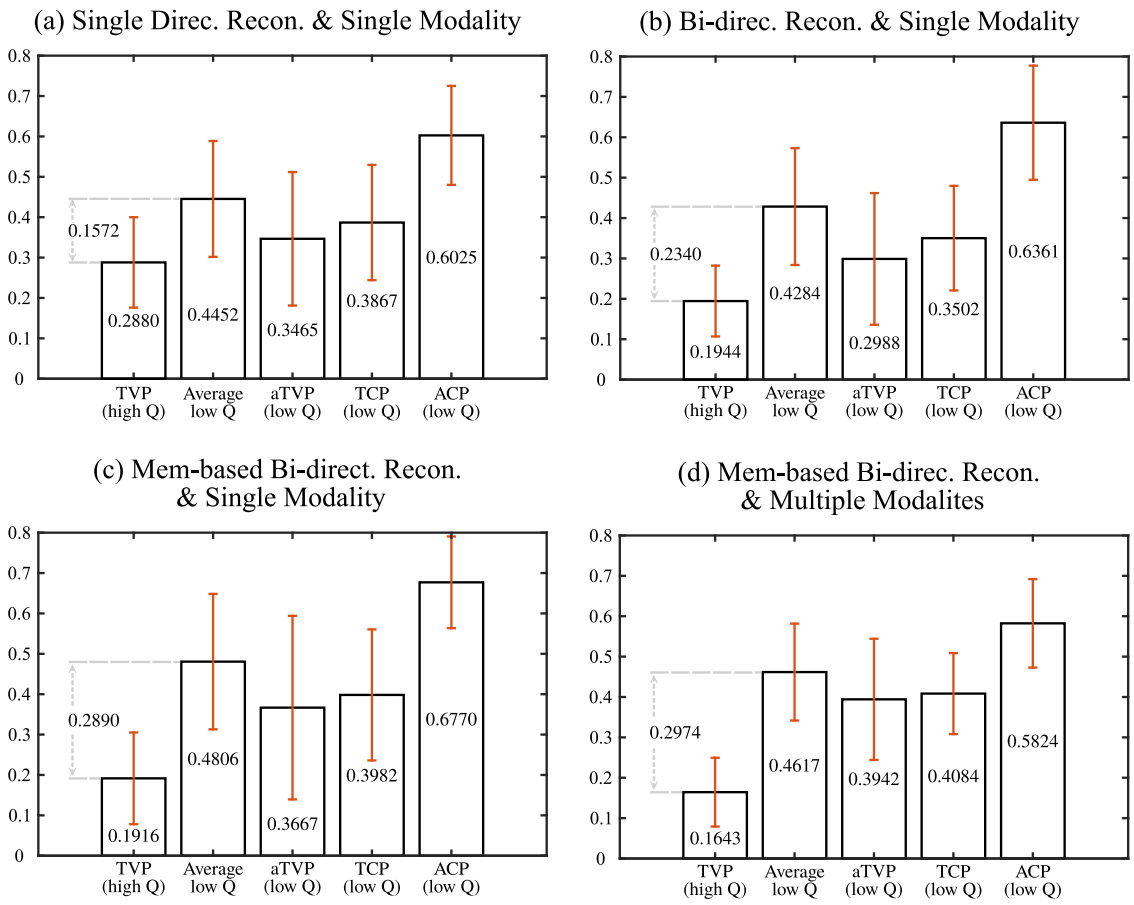


Fig. 9. Reconstruction error in feature space with different settings of reconstruction method, modality and model structure (*i.e.,* memory module). (a), (b), and (c) are with video modality only; (b), (c), and (d) are based on bi-directional reconstruction; (c) and (d) are with memory module.

module, which preserves the clarity of cerebellar structures even in low-quality data. The visual reconstruction results shown in Figs. 7 and 8 demonstrates the effectiveness of our memory module to capture features of different anatomies.

4.4.4. Reconstruction in the feature space

In the test phase, we utilize the reconstruction error in the feature space as the metric to discriminate low-quality videos from high-quality ones. We quantitatively analyze the reconstruction error for the high-quality and low-quality data under different conditions, including the use of memory module, as well as different modalities and reconstruction strategies. Fig. 9 reports the mean and standard deviation of the reconstruction error in the feature space.

Referring to Fig. 9(a) and (b), the high-quality videos (*i.e.*, TVP) report a higher average reconstruction error when the model is trained with only one-directional reconstruction flow. Besides, the reconstruction error difference between TVP (high Q) and Average low Q in (a) is smaller than that in (b), which makes it easier for the two-way reconstruction model to discriminate low-quality samples from high-quality ones. It demonstrates that more information is able to be learned from both video and feature spaces, which leads to superior performance. Furthermore, the memory-based model reports a smaller reconstruction error on high-quality data but a larger error on low-quality data (shown in (b) and (c)). This shows the benefits of the memory module, which enlarges the distance between high and low quality leading to increase the discriminative power of the model. The margin between high- and low-quality data is larger for the multi-modality case comparing Fig. 9(c) and (d), especially for aTVP that is the closest low-quality class to the high-quality class. Besides, the standard deviation is smaller for the multi-modality based model compared with the single modality based model. This suggests that multi-modality data is able to help the model learn a better representation and provide a more stable performance to identify low-quality videos.

5. Conclusion

In this paper, we propose a framework to assess video clinical quality in fetal ultrasound with an anomaly detection pipeline based on bi-directional reconstruction. The new approach avoids the drawbacks of traditional quality assessment that is dependent on anatomical annotations and a pre-defined protocol. Instead of an image-based assessment, we utilize a spatio-temporal encoder and decoder pair to exploit both the spatial and temporal information in the video. A memory module with key-query information retrieval mechanism is proposed to learn the typical information of high-quality data in the training phase. The additional modalities of optical flow and gaze are found to improve model performance by providing additional information on clinically relevant regions. Our approach provides a new idea about how to evaluate clinical ultrasound video quality in a data-driven fashion without relying on manual data annotations. Experiments on two obstetric examination applications demonstrate the effectiveness of our approach. Our approach may be readily generalized to different task-specific clinical ultrasound and non-ultrasound video quality assessment tasks.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data that has been used is confidential.

Acknowledgments

Funding for this study was granted by the European Research Council (ERC-ADG-2015 694581, project PULSE). ATP is supported by the National Institute for Health Research (NIHR) Oxford Biomedical Research Centre (BRC). The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.

References

Abdi, A.H., Luong, C., Tsang, T., Allan, G., Nouranian, S., Jue, J., Hawley, D., Fleming, S., Gin, K., Swift, J., et al., 2017a. Automatic quality assessment of echocardiograms using convolutional neural networks: feasibility on the apical four-chamber view. IEEE Trans. Med. Imaging 36 (6), 1221–1230.

Abdi, A.H., Luong, C., Tsang, T., Jue, J., Gin, K., Yeung, D., Hawley, D., Rohling, R., Abolmaesumi, P., 2017b. Quality assessment of echocardiographic cine using recurrent neural networks: Feasibility on five standard view planes. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 302–310.

Basharat, A., Gritai, A., Shah, M., 2008. Learning object motion patterns for anomaly detection and improved object detection. In: IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 1–8.

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., 2021. Image quality assessment for closed-loop computer-assisted lung ultrasound. In: Medical Imaging 2021: Image-Guided Procedures, Robotic Interventions, and Modeling, Vol. 11598. SPIE, pp. 183–189.

Baumgartner, C.F., Kamnitsas, K., Matthew, J., Fletcher, T.P., Smith, S., Koch, L.M., Kainz, B., Rueckert, D., 2017. SonoNet: real-time detection and localisation of fetal standard scan planes in freehand ultrasound. IEEE Trans. Med. Imaging 36 (11), 2204–2215.

Cai, Y., Sharma, H., Chatelain, P., Noble, J.A., 2018. Multi-task sonoeyenet: detection of fetal standardized planes assisted by generated sonographer attention maps. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part I. Springer, pp. 871–879.

Cavallaro, A., Ash, S.T., Napolitano, R., Wanyonyi, S., Ohuma, E.O., Molloholli, M., Sande, J., Sarris, I., Ioannou, C., Norris, T., et al., 2018. Quality control of ultrasound for fetal biometry: results from the intergrowth-21st project. Ultrasound Obstet. Gynecol. 52 (3), 332–339.

Chen, H., Ni, D., Qin, J., Li, S., Yang, X., Wang, T., Heng, P.A., 2015. Standard plane localization in fetal ultrasound via domain transferred deep neural networks. IEEE J. Biomed. Health Inform. 19 (5), 1627–1636.

Cheng, K.-W., Chen, Y.-T., Fang, W.-H., 2015. Video anomaly detection and localization using hierarchical feature representation and Gaussian process regression. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 2909–2917.

Cong, Y., Yuan, J., Liu, J., 2011. Sparse reconstruction cost for abnormal event detection. In: IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 3449–3456.

Dong, J., Liu, S., Liao, Y., Wen, H., Lei, B., Li, S., Wang, T., 2019. A generic quality control framework for fetal ultrasound cardiac four-chamber planes. IEEE J. Biomed. Health Inf. 24 (4), 931–942.

Drukker, L., Sharma, H., Droste, R., Alsharid, M., Chatelain, P., Noble, J.A., Papageorghiou, A.T., 2021. Transforming obstetric ultrasound into data science using eye tracking, voice recording, transducer motion and ultrasound video. Sci. Rep. 11 (1), 1–12.

Dudley, N., Chapman, E., 2002. The importance of quality management in fetal measurement. Ultrasound Obstet. Gynecol. 19 (2), 190–196.

Farnebäck, G., 2003. Two-frame motion estimation based on polynomial expansion. In: Scandinavian Conference on Image Analysis. Springer, pp. 363–370.

Fiorentino, M.C., Villani, F.P., Di Cosmo, M., Frontoni, E., Moccia, S., 2022. A review on deep-learning algorithms for fetal ultrasound-image analysis. Med. Image Anal. 102629.

Gao, Y., Beriwal, S., Craik, R., Papageorghiou, A.T., Noble, J.A., 2020. Label efficient localization of fetal brain biometry planes in ultrasound through metric learning. In: Medical Ultrasound, and Preterm, Perinatal and Paediatric Image Analysis. Springer, pp. 126–135.

Gong, D., Liu, L., Le, V., Saha, B., Mansour, M.R., Venkatesh, S., Hengel, A.v.d., 2019. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In: IEEE International Conference on Computer Vision. pp. 1705–1714.

Hasan, M., Choi, J., Neumann, J., Roy-Chowdhury, A.K., Davis, L.S., 2016. Learning temporal regularity in video sequences. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 733–742.

Hemmsen, M.C., Petersen, M.M.l., Nikolov, S.I., Nielsen, M.B., Jensen, J.r.A., 2010. Ultrasound image quality assessment: A framework for evaluation of clinical image quality. In: Medical Imaging 2010: Ultrasonic Imaging, Tomography, and Therapy, Vol. 7629. SPIE, pp. 105–116.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S., 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. Adv. Neural Inf. Process. Syst. 30.

Jaderberg, M., Simonyan, K., Zisserman, A., et al., 2015. Spatial transformer networks. Adv. Neural Inf. Process. Syst. 28, 2017–2025.

Kingma, D.P., Welling, M., 2014. Auto-encoding variational bayes. In: International Conference on Learning Representations. pp. 1–14.

Krasula, L., Le Callet, P., Fliegel, K., Klíma, M., 2017. Quality assessment of sharpened images: challenges, methodology, and objective metrics. IEEE Trans. Image Process. 26 (3), 1496–1508.

Li, C., Wand, M., 2016. Precomputed real-time texture synthesis with markovian generative adversarial networks. In: European Conference on Computer Vision. Springer, pp. 702–716.

Liao, Z., Girgis, H., Abdi, A., Vaseli, H., Hetherington, J., Rohling, R., Gin, K., Tsang, T., Abolmaesumi, P., 2019. On modelling label uncertainty in deep neural networks: Automatic estimation of intra-observer variability in 2d echocardiography quality assessment. IEEE Trans. Med. Imaging 39 (6), 1868–1883.

Lin, Z., Le, M.H., Ni, D., Chen, S., Li, S., Wang, T., Lei, B., 2018. Quality assessment of fetal head ultrasound images based on faster R-CNN. In: Simulation, Image Processing, and Ultrasound Systems for Assisted Diagnosis and Navigation. Springer, pp. 38–46.

Lin, Z., Li, S., Ni, D., Liao, Y., Wen, H., Du, J., Chen, S., Wang, T., Lei, B., 2019. Multi-task learning for quality assessment of fetal head ultrasound images. Med. Image Anal. 58, 101548.

Liu, W., Luo, W., Lian, D., Gao, S., 2018. Future frame prediction for anomaly detection–a new baseline. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 6536–6545.

Loizou, C.P., Pattichis, C.S., Pantziaris, M., Tyllis, T., Nicolaides, A., 2006. Quality evaluation of ultrasound imaging in the carotid artery based on normalization and speckle reduction filtering. Med. Biol. Eng. Comput. 44 (5), 414–426.

Luo, W., Liu, W., Gao, S., 2017. A revisit of sparse coding based anomaly detection in stacked rnn framework. In: IEEE International Conference on Computer Vision. pp. 341–349.

Papageorghiou, A.T., Ohuma, E.O., Altman, D.G., Todros, T., Ismail, L.C., Lambert, A., Jaffer, Y.A., Bertino, E., Gravett, M.G., Purwar, M., et al., 2014. International standards for fetal growth based on serial ultrasound measurements: the fetal growth longitudinal study of the intergrowth-21st project. Lancet 384 (9946), 869–879.

Park, H., Noh, J., Ham, B., 2020. Learning memory-guided normality for anomaly detection. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 14372–14381.

Prabhu, S.J., Kanal, K., Bhargava, P., Vaidya, S., Dighe, M.K., 2014. Ultrasound artifacts: classification, applied physics with illustrations, and imaging appearances. Ultrasound Q. 30 (2), 145–157.

Rahmatullah, B., Sarris, I., Papageorghiou, A., Noble, J.A., 2011. Quality control of fetal ultrasound images: Detection of abdomen anatomical landmarks using AdaBoost. In: IEEE International Symposium on Biomedical Imaging. IEEE, pp. 6–9.

Reddy, U.M., Filly, R.A., Copel, J.A., 2008. Prenatal imaging: ultrasonography and magnetic resonance imaging. Obstet. Gynecol. 112 (1), 145.

Saeed, S.U., Fu, Y., Baum, Z., Yang, Q., Rusu, M., Fan, R.E., Sonn, G.A., Barratt, D.C., Hu, Y., 2021. Learning image quality assessment by reinforcing task amenable data selection. In: International Conference on Information Processing in Medical Imaging. Springer, pp. 755–766.

Salomon, L., Alfirevic, Z., Da Silva Costa, F., Deter, R., Figueras, F., Ghi, T.a., Glanc, P., Khalil, A., Lee, W., Napolitano, R., et al., 2019. ISUOG practice guidelines: ultrasound assessment of fetal biometry and growth. Ultrasound Obstet. Gynecol. 53 (6), 715–723.

Salomon, L., Bernard, J., Duyme, M., Doris, B., Mas, N., Ville, Y., 2006. Feasibility and reproducibility of an image-scoring method for quality control of fetal biometry in the second trimester. Ultrasound Obstet. Gynecol. 27 (1), 34–40.

Sarris, I., Ioannou, C., Chamberlain, P., Ohuma, E., Roseman, F., Hoch, L., Altman, D., Papageorghiou, A., 2012. Intra-and interobserver variability in fetal ultrasound measurements. Ultrasound Obstet. Gynecol. 39 (3), 266–273.

Sassaroli, E., Crake, C., Scorza, A., Kim, D.-S., Park, M.-A., 2019. Image quality evaluation of ultrasound imaging systems: advanced B-modes. J. Appl. Clin. Med. Phys. 20 (3), 115–124.

Self, A., Daher, L., Schlussel, M., Roberts, N., Ioannou, C., Papageorghiou, A.T., 2022. Second and third trimester estimation of gestational age using ultrasound or maternal symphysis-fundal height measurements: A systematic review. BJOG: Int. J. Obstet. Gynaecol. 129 (9), 1447–1458.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2015. Going deeper with convolutions. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 1–9.

Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M., 2015. Learning spatiotemporal features with 3d convolutional networks. In: IEEE International Conference on Computer Vision. pp. 4489–4497.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Adv. Neural Inf. Process. Syst. 30.

Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P., 2004. Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. 13 (4), 600–612.

Wang, X., Che, Z., Jiang, B., Xiao, N., Yang, K., Tang, J., Ye, J., Wang, J., Qi, Q., 2021. Robust unsupervised video anomaly detection by multipath frame prediction. IEEE Trans. Neural Netw. Learn. Syst..

Wu, L., Cheng, J.-Z., Li, S., Lei, B., Wang, T., Ni, D., 2017. FUIQA: Fetal ultrasound image quality assessment with deep convolutional networks. IEEE Trans. Cybern. 47 (5), 1336–1349.

Yaqub, M., Kelly, B., Stobart, H., Napolitano, R., Noble, J., Papageorghiou, A., 2019. Quality-improvement program for ultrasound-based fetal anatomy screening using large-scale clinical audit. Ultrasound Obstet. Gynecol. 54 (2), 239–245.

Yaqub, M., Sleep, N., Syme, S., Chen, Z., Ryou, H., Walton, S., Noble, J.A., Papageorghiou, A.T., 2021. 491 ScanNav® audit: an AI-powered screening assistant for fetal anatomical ultrasound. Amer. J. Obstet. Gynecol. 224 (2), S312.

Zhang, L., Dudley, N.J., Lambrou, T., Allinson, N., Ye, X., 2017. Automatic image quality assessment and measurement of fetal head in two-dimensional ultrasound image. J. Med. Imaging 4 (2), 024001.

Zhang, B., Liu, H., Luo, H., Li, K., 2021. Automatic quality assessment for 2D fetal sonographic standard plane based on multitask learning. Medicine 100 (4).

Zhao, Y., Deng, B., Shen, C., Liu, Y., Lu, H., Hua, X.-S., 2017. Spatio-temporal autoencoder for video anomaly detection. In: Proceedings of the 25th ACM International Conference on Multimedia. pp. 1933–1941.

Zhao, H., Zheng, Q., Teng, C., Yasrab, R., Drukker, L., Papageorghiou, A.T., Noble, J.A., 2022. Towards unsupervised ultrasound video clinical quality assessment with multi-modality data. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 228–237.

Zhu, J.-Y., Park, T., Isola, P., Efros, A.A., 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: IEEE International Conference on Computer Vision. pp. 2223–2232.