Towards Unsupervised Ultrasound Video Clinical Quality Assessment with Multi-modality Data

He Zhao[1](✉), Qingqing Zheng[2], Clare Teng[1], Robail Yasrab[1], Lior Drukker[3,4], Aris T. Papageorghiou[3], and J. Alison Noble[1]

[1] Institute of Biomedical Engineering, University of Oxford, Oxford, UK
he.zhao@eng.ox.ac.uk

[2] Guangdong Provincial Key Laboratory of Computer Vision and Virtual Reality, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China

[3] Nuffield Department of Women's and Reproductive Health, University of Oxford, Oxford, UK

[4] Department of Obsterics and Gynecology, Tel-Aviv University, Tel Aviv, Israel

Abstract. Video quality assurance is an important topic in obstetric ultrasound imaging to ensure that captured videos are suitable for biometry and fetal health assessment. Previously, one successful objective approach to automated ultrasound image quality assurance has considered it as a supervised learning task of detecting anatomical structures defined by a clinical protocol. In this paper, we propose an alternative and purely data-driven approach that makes effective use of both spatial and temporal information and the model learns from high-quality videos without any anatomy-specific annotations. This makes it attractive for potentially scalable generalisation. In the proposed model, a 3D encoder and decoder pair bi-directionally learns a spatio-temporal representation between the video space and the feature space. A zoom-in module is introduced to encourage the model to focus on the main object in a frame. A further design novelty is the introduction of two additional modalities in model training (sonographer gaze and optical flow derived from the video). Finally, our approach is applied to identify high-quality videos for fetal head circumference measurement in freehand second-trimester ultrasound scans. Extensive experiments are conducted, and the results demonstrate the effectiveness of our approach with an AUC of 0.911.

1 Introduction

Ultrasound imaging is widely used in obstetrics for fetal health assessment due to its portability, low cost, and free radiation. The high dependence on experience, and intra- and inter-observer variability is also well known. For example, it can be difficult for trainee sonographers to localize the appropriate plane for diagnosis because of fetal movement and acoustic shadowing, and even experienced sonographers can struggle to acquire good diagnostic images for subjects with

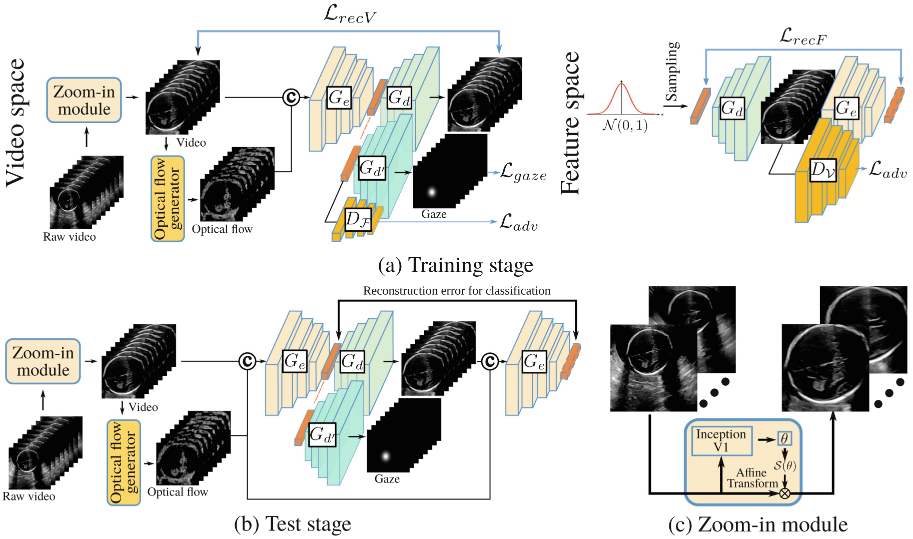
L. Wang et al. (Eds.): MICCAI 2022, LNCS 13434, pp. 228–237, 2022.
https://doi.org/10.1007/978-3-031-16440-8_22

poor acoustic windows. Assessment and audit of video quality is recommended in clinical guidelines. However, this has to be done by an experienced sonographer which is very time-consuming and labour-intensive and takes clinicians away from treating patients. Despite its importance to clinical practice, hardly any research has been reported on automated video clinical quality assessment.

In this paper, we are interested in video clinical quality assessment which is task-specific for biometric measurement. High clinical quality means the video is suitable for further measurement and analysis. A novel data-driven approach is proposed by learning a model of video quality assessment directly from high-quality data. Our approach learns the spatio-temporal representation between the video and feature space bi-directionally with a reconstruction-based anomaly detection pipeline. The intuition is that a low-quality sample can be detected by its associated large reconstruction error as the sample is not present in the training data. Different from existing supervised image quality assessment methods for ultrasound [1,8,15], our approach makes effective use of both spatial and temporal information and the model learns from high-quality videos without any anatomy-specific annotations. These characteristics make our approach attractive for clinical quality assessment tasks where anatomical annotations are often rare and inaccessible. The contributions of this paper are summarized as follows: (1) To the best of our knowledge, our approach is the first video-based clinical quality assessment method that does not depend on clinical protocol definitions and anatomical annotations. (2) Bi-directional reconstruction between the video and feature spaces prompts our model to learn an informative representation of high-quality data. (3) We propose to use multi-modality data (*i.e.,* optical flow & gaze) in the training stage with the help of an input generator and an auxiliary prediction branch, respectively. This prediction branch further enables our model to highlight informative structures by the predicted gaze.

2 Related Work

Image quality assessment has been studied extensively in image processing with various assessment metrics proposed such as PSNR, SSIM [14], and FID [5]. These image quality metrics focus on image clarity and noise removal. The definition of quality assessment in ultrasound is different in that it needs to factor in clinical context; it is task-specific and aims to ensure that a frame is useful for diagnosis. Prior work has mainly aimed to automate the clinical criteria checklist specified in clinical scanning protocol guideline standards. Early work is reported in [11] and [16]. Wu *et al.* [15] propose two convolutional networks to locate the ROI and detect two anatomies of the fetal abdomen in the 2nd trimester, where a quality score is based on the appearance of the ROI and anatomies. A multi-task Fast R-CNN based quality assessment network for scoring head images is described in [8,9]. In [1], a three-step framework is proposed to give a quality score for the fetal cardiac plane. Firstly, the cardiac four-chamber planes are detected and then a detection network locates the anatomical structures. The authors also propose a classification network that considers two other indices (*i.e.,* view zoom and gain), which is not used in previous studies. A semi-supervised approach using metric learning is proposed in [4] for selecting head planes in low-cost ultrasound probe video. In [12], the authors propose a reinforcement learning method to select images which are amenable to a target task. Although it is not based on clinical criteria, detailed anatomical annotations are still required in training. A recent evaluation of a real-time Artificial Intelligence (AI) based system that automatically keeps track of acquired images and checks images conform to imaging protocol standards is reported in [17] where five experienced sonographers are used as the reference. A specified pre-defined protocol and annotated locations of anatomical structures are required in the aforementioned methods, which limits transferability to new applications.



(a) Training stage

(b) Test stage

(c) Zoom-in module

Fig. 1. Flowchart of our approach. (a) Training stage with bi-directional reconstruction loop in video and feature spaces. (b) Test stage with feature reconstruction error for classification. (c) Details of the zoom-in module.

3 Method

Our approach assesses clinical quality of ultrasound videos using only qualified scans without anatomical annotation. We formulate the video quality assessment task as an anomaly detection problem, where low-quality video is regarded as anomalous data. Denote the training dataset as $\mathcal{D}$ with N high-quality training samples only, *i.e.,* $\mathcal{D} = \{x_i, ..., x_N\}$, and a test set $\mathcal{D}_t$, *i.e.,* $\mathcal{D}_t = \{(x_{t_1}, y_1), ..., (x_{t_M}, y_{t_M})\}$ where $y \in \{0, 1\}$ indicates a video label (0 for high quality and 1 for low quality). Our goal is to train a model to learn the distribution of high-quality videos from the training dataset $\mathcal{D}$ and to identify the low-quality video in the test dataset $\mathcal{D}_t$ as anomalous. A three-dimensional encoder G_e and decoder G_d pair is proposed to learn the spatio-temporal representation. The bi-directional information flow between video space and feature space provides feedback for the model during training. This information allows the high-quality data feature representation to be informative and discriminative from that of the low-quality data.

3.1 Model Structure

The pipeline of our approach is shown in Fig. 1. For each given ultrasound video, the main object of interest (*e.g.*, fetal head) is first extracted by the pre-trained zoom-in module. An optical flow generator is followed to estimate an optical flow field describing displacement from the zoomed-in video, which serves as the second modality input in our model. An encoder G_e and decoder G_d pair with 3D convolutional layers is adopted to learn spatio-temporal features from both video-based modalities. Two adversarial reconstruction processes are performed bi-directionally between the video and feature spaces with different alternative combinations of G_e and G_d. Besides video and optical flow, a third modality, gaze, is used by an auxiliary branch to predict where a sonographer looks. Feature reconstruction error is used as the indicator to recognize low-quality data as this will have a large reconstruction error.

Spatial Zoom-In Module and Optical Flow Generator. The goal of the zoom-in module is to extract the spatial region of interest in a video frame. As shown in Fig. 1(c), the original ultrasound video may contain fetal structures (*e.g.*, head) with a low field-of-view occupancy. This may mislead the model as the background has a major influence on overall reconstruction error. Inspired by [6], a zoom-in module is introduced to locate and transform the image to center the region of interest around the fetal structure. Inside the zoom-in module, we use InceptionV1 [13] to learn its affine transformation parameters. This plug-in module is pre-trained with approximate bounding boxes around the fetal structures and is fixed in the following stage. The optical flow generator is developed to capture displacement patterns that characterize the appearance of anatomical structures in videos. We employ the Farneback algorithm [3] with a window size of 3×3 to generate a dense optical flow field. A median filter with a kernel size of 21×21 is applied as pre-processing to reduce the effect of speckle on optical flow field estimation.

Bi-directional Reconstruction Between Two Spaces. As shown in Fig. 1 there are two directional reconstruction processes assisted by adversarial learning. One is video reconstruction following *video* → *feature* → *video* by G_e–G_d; the second is feature reconstruction going along with *feature* → *video* → *feature* by G_d–G_e. The encoder G_e consists of eight 3D convolutional layers. The first five layers are with kernel size $1 \times 4 \times 4$ and stride $1 \times 2 \times 2$ performing spatial convolution, while the last three layers are with kernel size $4 \times 4 \times 4$ and stride $2 \times 2 \times 2$ performing spatio-temporal convolution leading to a bottleneck feature with size of 1024. The decoder G_d is with symmetrical structure but uses deconvolutional layers instead. The bi-directional information flow helps the model gain better understanding of high-quality videos. Two discriminators (*i.e.,* $D_{\mathcal{V}}$ and $D_{\mathcal{F}}$) are also proposed in the video space and feature space, respectively, for generating realistic high-quality data. The discriminator $D_{\mathcal{V}}$ has the similar structure of encoder and $D_{\mathcal{F}}$ consists of a stack of fully connected layers with neurons from 64 to 1.

Auxiliary Gaze Branch. Eye-tracking data records sonographer gaze locations during scanning. Trying to predict gaze forces the model to learn the salient regions of interest of high-quality video. To take full advantage of this prior knowledge, we introduce an auxiliary decoder $G_{d'}$ with the same structure as G_d, to learn gaze map. Compared with using the eye-tracking data as additional input, the training scheme as prediction eliminates the requirement for gaze in the test phase. It also enables the model to provide guidance to novice sonographers on where to look and which spatial parts are essential.

3.2 Objective Function

Training is supervised by the bi-directional reconstruction and gaze groundtruth. The encoder $G_e(x, o) : \mathcal{V} \rightarrow \mathcal{F}$ takes the video and optical flow as input and transforms them into the feature space. The decoder $G_d(f) : \mathcal{F} \rightarrow \mathcal{V}$ converts the feature representation back into the video space. Zero-sum games are played between G_e, G_d and the two discriminators. Our model is trained to solve the following optimization function:

$$\min_{G_e, G_d} \max_{D_{\mathcal{F}}, D_{\mathcal{V}}} \mathcal{L} = \omega_{adv}\mathcal{L}_{adv} + \omega_{rec}\mathcal{L}_{rec} + \omega_{gaze}\mathcal{L}_{gaze}, \tag{1}$$

where $\mathcal{L}_{rec}$, $\mathcal{L}_{gaze}$ are the bi-directional reconstruction loss and gaze loss, respectively. The adversarial loss function $\mathcal{L}_{adv}$ is defined by the least-squares adversarial loss:

$$\mathcal{L}_{adv} = |D_{\mathcal{F}}(f) - 1|^2 + |D_{\mathcal{F}}(G_e(x, o))|^2 + |D_{\mathcal{V}}(x) - 1|^2 + |D_{\mathcal{V}}(G_d(f))|^2, \tag{2}$$

where x, o are the video and the optical flow, respectively, and f is the feature vector sampled from a standard multivariate Gaussian distribution similarly as in [7]. The adversarial loss aims to learn more realistic reconstructions in both video and feature space by $D_{\mathcal{V}}$ and $D_{\mathcal{F}}$, respectively.

Reconstruction Loss. The reconstruction loss allows the encoder-decoder or decoder-encoder models to learn spatio-temporal representations of high-quality videos. Instead of the widely used pixel-wise L1 loss, the structure similarity (SSIM) [14] loss is applied for a perceptual spatial constraint. The bi-directional reconstruction loss $\mathcal{L}_{rec}$ in the video space and feature space is defined as:

$$\mathcal{L}_{rec} = \mathcal{L}_{recV} + \mathcal{L}_{recF}, \tag{3}$$

where $\mathcal{L}_{recV}$ and $\mathcal{L}_{recF}$ are defined as: $\mathcal{L}_{recV} = 1 - SSIM(x, G_d(G_e(x, o))$ and $\mathcal{L}_{recF} = |G_e(G_d(f), o) - f|$, respectively.

Gaze Loss. We introduce a new loss function for the model to learn the gaze saliency map. The gaze loss aims to minimize the difference between the gaze prediction map and the ground truth and is defined as:

$$\mathcal{L}_{gaze} = |G_{d'}(G_e(x, o)) - g|, \qquad (4)$$

where g is the eye gaze ground truth.

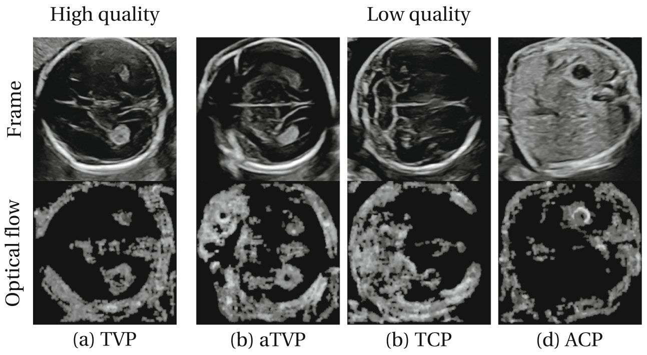
4 Experiment and Results

As part of the PULSE study [2], a dataset of 430 subjects with a resolution of 1008 × 784, including video and gaze data, is used in our experiments. During a scan, an experienced sonographer finds and freezes a biometry plane. The video clip consists of the frozen frame and 2s before freezing and is labeled by the frozen frame type, *e.g.*, transventricular plane (TVP), transcerebellar plane (TCP), abdominal circumference plane (ACP). An approaching the transventricular plane (aTVP) video clip is collected 5–7s before the frozen TVP frame. We collect 430 high-quality TVP video clips (one clip per subject) and 181 low-quality clips. For training, 300 high-quality video clips (TVP) are randomly selected, and the remaining 130 high-quality and 181 low-quality clips are used for test. Each input sample to the model consists of 8 frames sampled from 2 s video clips at an 8-frame interval and is further resized to 256 × 256. Our approach is implemented in PyTorch with a 12 GB TitanX GPU. [1] The model was trained for 200 epochs with an Adam optimizer and the learning rate is set to 0.0002, which is linearly decays to 0 in the last 100 epochs. The loss weights ω_{adv}, ω_{rec} were empirically set to 1 and 10, respectively, to make the value of each loss stay at the same numerical level. The gaze loss weight ω_{gaze} was set to 0.1 based on a parameter study reported in the following section.

Figure 2 presents exemplar frames of high- and low-quality videos together with their dense optical flow field estimated by the optical flow generator. Observe that the different planes have different displacement patterns. For example, for the TVP, the choroid plexus (CP) and brain midline region change the most during scanning; for the TCP, the displacement pattern is high in the cerebellum region. These patterns provide useful additional information for the model to learn the feature representation of high-quality data.

Quantitative Results. We compare our approach with three single-modality methods: a SpatioTemporal Auto-Encoder (STAE) [18], MNAD [10] and an image-based approach which only takes the last frozen frame of the video clip as input. MNAD is a video anomaly detection method which detects anomalous frames in a video. It is obviously unsuitable for our task, thus leading to a rather low performance. Table 1 compares these reference methods with variants of our architecture in terms of the area under the ROC curve (AUC), F1-score, accuracy, sensitivity, and specificity. For all performance metrics, there is a large

[1] Code is available at https://github.com/IBMEOX/UltrasoundVQA.



(a) TVP (b) aTVP (b) TCP (d) ACP

Fig. 2. Exemplar frames and corresponding optical flow fields of high- and low-quality videos from the output of the zoom-in module.

Table 1. Performance of different methods based on the zoomed-in videos with the evaluation metric of AUC, F1 (%), ACC (%), SEN (%) and SPE (%).

Methods			AUC	F1	ACC	SEN	SPE
Single modality	Image-based		0.790 ± 0.006	72.29	71.06	80.11	62.05
	MNAD [10]		0.308 ± 0.009	73.32	57.88	99.45	1.54
	STAE [18]		0.824 ± 0.009	80.46	76.07	84.61	64.18
Multiple modalities	Our approach	Video only	0.863 ± 0.005	82.66	78.78	86.90	67.47
		with Optical flow	0.889 ± 0.006	85.40	82.54	87.69	75.39
		with Gaze	0.886 ± 0.004	84.88	81.67	88.40	72.31
		All modalities	**0.911 ± 0.003**	**86.99**	**84.56**	**88.62**	**78.92**

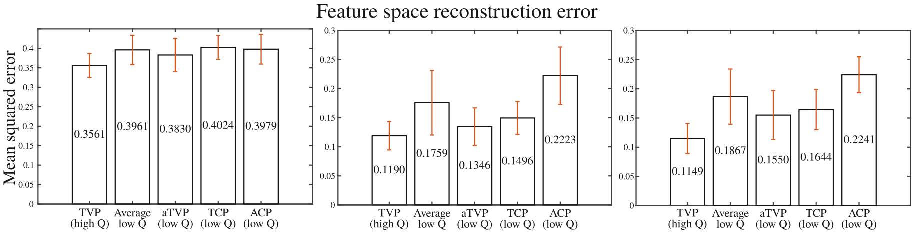
gap between the image-based and video-based methods, supporting a hypothesis that temporal information is useful to assess clinical quality for clinical tasks. This result is also explainable clinically, the last frozen frame is not always the best diagnostic frame for biometry. The conclusion from this experiment is that including temporal information is helpful to distinguish between task-specific low-quality and high-quality videos. Among the video-based methods, our bi-directional reconstruction approach performs better than single-modality video reconstruction with an improvement of AUC by 4.8%. With the addition of other data modalities, *i.e.,* optical flow, and gaze, the AUC further increases from 0.863 to 0.911, respectively. Moreover, simple perturbations (*e.g.*, flipping, adding Gaussian noise) are applied on test images leading to the AUC of 0.906, which indicates the robustness of our approach. The paired t-test between our approach and STAE [18] is performed with p-value of 8×10^{-5}, which demonstrates the statistically significant benefit of our approach.

Ablation Study. Experiments were performed to study the effect of model components and parameter settings. The top panel of Table 2 demonstrates the effectiveness of the zoom-in module. Observe that a significant improvement is achieved by inclusion of the zoom-in module, with an AUC increase from 0.744 to

0.889. The explanation for this improvement is that the zoom-in module forces the encoder and decoder to concentrate on the essential region of the video instead of reconstructing background pixels which are not of interest. The bottom panel of Table 2 reports model performance for different ω_{gaze}. This additional training guidance further improves the AUC performance of our model from 0.889 to 0.911.

Table 2. Ablation study performance summary of the zoom-in module and different settings of the gaze loss weight. Note models are trained with inputs of video and optical flow.

		AUC	F1	ACC	SEN	SPEC
w/o zoom-in module		0.744	75.39	70.85	86.25	54.39
Zoom-in module		0.889	85.40	82.54	87.69	75.39
Gaze loss	$\omega_{gaze} = 0$	0.889	85.40	82.54	87.69	75.39
	$\omega_{gaze} = 0.1$	**0.911**	**86.99**	**84.56**	88.62	**78.92**
	$\omega_{gaze} = 0.5$	0.899	85.87	82.96	**88.95**	74.62
	$\omega_{gaze} = 1$	0.888	85.25	82.64	86.19	77.69



(a) Single directional reconstruction & single-modality

(b) Bi-directional reconstruction & single-modality

(c) Bi-directional reconstruction & multi-modality

Fig. 3. Reconstruction error in feature space with respect to reconstruction method and modality.

Figure 3 (a)–(c) report the mean and standard deviation of the feature space reconstruction error for the high-quality data (*i.e.*, TVP) and low-quality data (*i.e.*, aTVP, TCP, ACP). The strength of bi-directional reconstruction is demonstrated in Fig. 3(a) and (b). The difference in reconstruction error using a single directional model is very small. Therefore it is not as easy to distinguish between high- and low-quality videos. Conversely, the bi-directional model shows a larger error. The results demonstrate that more information can be learned by the bi-directional reconstruction, thus leading to better performance. The effect of using multi-modality in our model is reported in Fig. 3(b) and (c). The difference in reconstruction error between low quality and high quality is small for the single modality model, especially for aTVP which is the closest video clip to high-quality data. The margin between high- and low-quality data is greater for the multi-modality case. We conclude that the model trained with multi-modal data is able to better distinguish low-quality videos from high-quality videos, compared with just modelling from video alone.

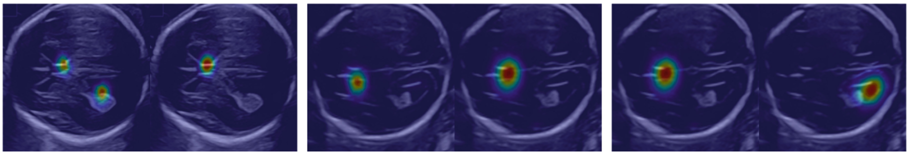


Fig. 4. Three examples of gaze prediction between two consecutive frames.

Gaze Prediction. Our model architecture uses eye gaze in an auxiliary branch instead of an input which allows the model to filter low-quality videos and also performs gaze prediction. Figure 4 shows three example gaze predictions on consecutive test frames. Observe that the gaze predictions mainly focus on the cavum septi pellucidi (CSP) and choroid plexus (CP), which are two anatomical structures that a sonographer pays attention to during scanning. The accuracy of gaze prediction is approximate 89%, where most of the eye gaze falls on CP, CSP, middle line, and the skull boundary.

5 Conclusion

In conclusion, we propose a data-driven method to assess ultrasound video clinical quality. Our approach directly learns a model from high-quality data without any anatomical annotations or protocol. The bi-directional reconstruction between video space and feature space aids the model in learning a meaningful representation of high-quality video. The addition of gaze and optical flow to video improved model performance by providing additional information about clinically important regions. Our approach provides a new idea to evaluate ultrasound video quality in a data-driven fashion without relying on data annotations. It may be readily applied to different task-specific clinical video quality assessment problems.

Acknowledgement. This paper is funded by the ERC (ERC-ADG-2015 694581 project PULSE), the EPSRC (EP/MO13774/1. EP/R013853/1), and the NIHR Biomedical Research Centre funding scheme.

References

1. Dong, J., et al.: A generic quality control framework for fetal ultrasound cardiac four-chamber planes. IEEE J. Biomed. Health Inform. **24**(4), 931–942 (2019)

2. Drukker, L., et al.: Transforming obstetric ultrasound into data science using eye tracking, voice recording, transducer motion and ultrasound video. Sci. Rep. **11**(1), 1–12 (2021)
3. Farnebäck, G.: Two-frame motion estimation based on polynomial expansion. In: Bigun, J., Gustavsson, T. (eds.) SCIA 2003. LNCS, vol. 2749, pp. 363–370. Springer, Heidelberg (2003). https://doi.org/10.1007/3-540-45103-X_50
4. Gao, Y., Beriwal, S., Craik, R., Papageorghiou, A.T., Noble, J.A.: Label efficient localization of fetal brain biometry planes in ultrasound through metric learning. In: Hu, Y., et al. (eds.) ASMUS/PIPPI -2020. LNCS, vol. 12437, pp. 126–135. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-60334-2_13
5. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
6. Jaderberg, M., Simonyan, K., Zisserman, A., et al.: Spatial transformer networks. In: Advances in Neural Information Processing Systems, vol. 28, pp. 2017–2025 (2015)
7. Kingma, D.P., Welling, M.: Auto-encoding variational Bayes. In: International Conference on Learning Representations, pp. 1–14 (2014)
8. Lin, Z., et al.: Quality assessment of fetal head ultrasound images based on faster R-CNN. In: Stoyanov, D., et al. (eds.) POCUS/BIVPCS/CuRIOUS/CPM -2018. LNCS, vol. 11042, pp. 38–46. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01045-4_5
9. Lin, Z., et al.: Multi-task learning for quality assessment of fetal head ultrasound images. Med. Image Anal. **58**, 101548 (2019)
10. Park, H., Noh, J., Ham, B.: Learning memory-guided normality for anomaly detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14372–14381 (2020)
11. 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)
12. Saeed, S.U., et al.: Learning image quality assessment by reinforcing task amenable data selection. In: Feragen, A., Sommer, S., Schnabel, J., Nielsen, M. (eds.) IPMI 2021. LNCS, vol. 12729, pp. 755–766. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-78191-0_58
13. Szegedy, C., et al.: Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–9 (2015)
14. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004)
15. Wu, L., Cheng, J.Z., Li, S., Lei, B., Wang, T., Ni, D.: FUIQA: fetal ultrasound image quality assessment with deep convolutional networks. IEEE Trans. Cybern. **47**(5), 1336–1349 (2017)
16. Yaqub, M., Kelly, B., Papageorghiou, A.T., Noble, J.A.: A deep learning solution for automatic fetal neurosonographic diagnostic plane verification using clinical standard constraints. Ultrasound Med. Biol. **43**(12), 2925–2933 (2017)
17. Yaqub, M., et al.: 491 scannav® audit: an AI-powered screening assistant for fetal anatomical ultrasound. Am. J. Obstet. Gynecol. **224**(2), S312 (2021)
18. Zhao, Y., Deng, B., Shen, C., Liu, Y., Lu, H., Hua, X.S.: Spatio-temporal autoencoder for video anomaly detection. In: Proceedings of the 25th ACM International Conference on Multimedia, pp. 1933–1941 (2017)