Visual-Query Curriculum Learning for Rare Fetal Biometry Plane Detection in Blind-Sweep Ultrasound Video

Jong Kwon[1], Jianbo Jiao[2], Alice Self[3], Aris Papageorghiou[3], and J. Alison Noble[1]

[1] Department of Engineering Science, University of Oxford
[2] School of Computer Science, University of Birmingham
[3] Nuffield Department of Women's and Reproductive Health, University of Oxford
kwon.h.jong@gmail.com

Abstract. Blind sweep ultrasound protocols can expand access to antenatal screening in low- and middle-income countries (LMICs), but their clinical utility depends on whether a sweep contains diagnostically useful views. In routine obstetric imaging, gestational age is estimated from fetal biometry measured on standardised planes (e.g. head and abdominal circumference). In blind sweeps, these planes may appear only briefly and under large appearance variation, making manual review burdensome and automated processing unreliable without quality screening. We propose *biometry plane detection* as an application-driven quality assessment task: given a blind sweep video, identify frames suitable for femur length, head- and abdominal-circumference biometry. We formulate the problem as visual-query-guided retrieval in long ultrasound videos and introduce a spatio-temporal transformer that matches a canonical query image to candidate frames. We show that we can leverage this query with an ultrasound-specific curriculum to learn effectively under extreme label sparsity. To transfer knowledge from a large freehand dataset to sparsely labelled sweeps, we incorporate scale-robust visual encoding and domain-adversarial training. Experiments on a participant-disjoint blind-sweep test set demonstrate improved detection performance over image-based and video-based baselines, while highlighting the remaining challenge posed by ultra-low prevalence and incomplete frame-level annotation. We hope this work supports adoption of blind-sweep fetal ultrasound by enabling automatic screening and rapid re-acquisition when biometry content is absent. Project page: https://github.com/kwon-j/vqcl.

Keywords: Quality assessment · Biometry plane detection · Fetal ultrasound · Video transformers

1 Introduction

A major part of the routine obstetric ultrasound (US) examinations is biometry — where measurements are made of anatomical structures such as the abdominal circumference (AC), head circumference (HC) and femur length (FL). The

sizes of these anatomies are used to estimate the gestational age of the fetus (GA). These anatomies are measured from specific angles called *standard planes* which are highly standardised by strict visibility/absence criteria for anatomical landmarks [21,17] to promote consistent measurement. Acquiring them involves fine-tuning the probe position through real-time visual feedback so requires deep anatomical knowledge and skill. This is a major barrier to deployment in antenatal care in LMICs where access to trained sonographers may be limited [4].

Simplified acquisition protocols based on long, blind probe sweeps have therefore been proposed to reduce training requirements [5,24,1,23]. These protocols are promising for LMIC deployment, but they introduce a new failure mode: because sweeps are acquired without targeting specific views, a scan may contain insufficient information for downstream tasks. However, we cannot assume an untrained user can assess the quality of a blind sweep scan. Furthermore, US scans are cheaply retaken during the same appointment, while retaking a scan after the patient has left is a large burden to the patient.

We focus on quality assessment for biometry-based GA estimation. We define a sweep as suitable if it contains frames sufficient for approximate HC/AC/FL measurement, and refer to detecting these frames in **blind sweep videos** as **biometry plane detection**; *not* standard plane detection[4], because they are not standardised. Unlike freehand standard planes, sweep biometry planes are off-axis, brief, and often motion-blurred because they are encountered by chance rather than deliberately acquired. Despite this, approximate planes can still support clinically useful GA estimation in regions where last menstrual period remains the only, and often unreliable, method for GA estimation.

Related Work. Standard plane detection in freehand ultrasound is well studied [7,10,3], but these frame-based models degrade on blind sweeps due to domain shift and the need for temporal context.

There is recent work on direct GA estimation from standard plane images and video without biometry, for both free-hand US and blind-sweep US [14,13,23]. Promisingly, [23] found direct GA estimation of blind US gave ($\pm$3 days MAE) for two separate cohorts (the same as freehand US gold-standard biometry). Multiple previous works on manual biometry of blind sweep US videos have found good agreement ($< \pm$ 1 week MAE) of GA [5,2], which further supports the feasibility and utility of these scans for automated biometry-based GA estimation. [5] uses the presence of biometry planes as a quality metric, to validate their quality assurance process and mentions quality assessment is a scarce area of research. However automated detection of biometry planes in sweep videos remains underexplored.

Biometry-based GA is the current clinical gold standard, hence is more amenable for clinical adoption than direct GA estimation. Further, our work complements direct GA estimation work as it can be used as a preprocessing step to screen whether videos contain relevant and informative structures. Due to the black-box nature of deep-learning based models, a GA model is likely to still give a GA prediction for a non-informative video.

[4] Standard plane detection is a task only in regular freehand US scans

2 Data

We used 3 datasets for this work: (1) blind-sweep videos with biometry plane frames annotated, (2) freehand US videos with standard-plane-like frames annotated and (3) standard plane frames individually, i.e. not as part of videos. The amounts of each are shown respectively in Figure 1.

Biometry Plane Data. Our data [22] consisted of approximately 4 hours of real clinical blind-sweep ultrasound, comprising 334,000 frames from **115 participants'** US videos. The sweeps were acquired by a hospital sonographer after routine obstetric scanning using a GE Voluson E8, at 18–28 weeks' gestation (CALOPUS – see acknowledgements). The sonographer found only 248 biometry frames, which is **extremely** sparse ($< 0.1\%$). These were unevenly split, with much lower FL planes (63 FL vs. 84 HC and 101 AC), which reflects the difficulty of the FL task. Lower FL prevalence is consistent with other manual biometry plane detection in the literature [2,5,18]. The femur is essentially 1D, whereas the head/abdomen are approximately spherical/cylindrical, thus the chances of the imaging plane bisecting exactly along the 1D axis to get FL is much lower than getting HC/AC.

The sonographer gave each plane a subjective quality score of: Excellent, Acceptable or Poor depending on how clear/usable the plane looked. The most frequent quality was "Poor", which further obscures the already small labelled data, highlighting the difficulty of the task.

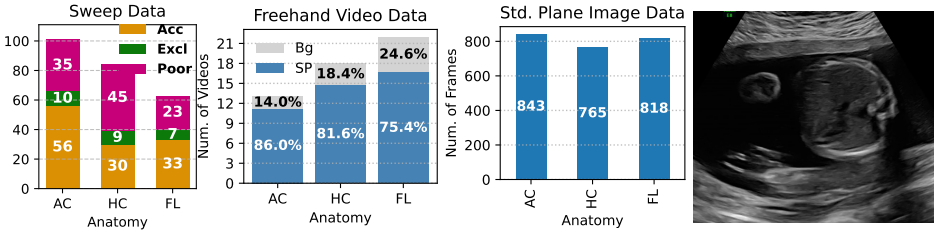

Fig. 1: The middle graph shows, on average, how much of the clip is background vs. standard plane-like (480 frames in each video). Right is an example "false positive" biometry plane.

We counted 4 frames (± 0.13s at 30fps) on either side of the expert-labelled biometry plane as a true prediction to augment the low prevalence. This was manually inspected to be a fairly consistent temporal cut-off for which frame appearance would change semantically. This could be done for this sweep data because during blind-sweep scans, the probe is moved at a constant speed in a consistent direction.

Standard Plane Data. The freehand US data was taken on the same E8 probe during real clinical appointments of 846 participants (PULSE – see acknowledge-

ments). The standard plane frames were defined by the frozen frame, where the sonographer pauses the video to take measurements.

This definition is not conducive to train a video model, as the ground truth is given as a single frame; however, because the sonographer lingers around the standard plane to fine-tune, many of the preceding frames also resemble standard planes but there is no clear temporal cut-off to determine when frames stop looking like standard planes.

Thus to obtain freehand video labelled data, we split up the freehand US videos into 360 frames before the standard plane and 120 frames after and a non-expert (lead author) manually annotated frames that resembled standard planes. This temporal window meant most videos had more positive annotations than negatives, (opposite to sweep video annotations Figure 1).

3 Method

Core Idea and Formulation. Whilst this can be framed as a simple 3 or 4 class (AC/ HC / FL / background) classifier problem, we use a visual query to define the anatomy to detect, then frame the task as a one-vs-all binary classification, similar to Ego4D visual query localisation challenge [9]. The visual query is simply an example image of an HC, AC, or FL plane, which the model is given along with the video and the model seeks to find only that anatomy plane.

This visual query is crucial in coaxing large, video models to learn useful representations in the ultra-sparse regime. The core idea is curriculum learning – we start with the easy task of: given this AC frame, find when in this video clip this exact same frame appears (could be done by literal pixel matching). Progresses to: here's a slightly transformed version of that target AC frame, now find when in the clip the frame occurs. To: here's a similar AC frame (but from another patient/video), find when the AC frame for this video occurs, progressively until you reach: here's a really visually different frame but semantically still AC, find me the AC frames in this clip. This progression coaxes the model to move from low-level pixel matching toward a high-level anatomical representation.

At inference, we still need to give the model an exemplar AC, HC, FL frame, but at this point, the model should have learned semantic meaning of finding all AC frames, even if the given query is poor quality. We choose the AC, HC, FL to give the model, so we can pick a high-quality standard plane (without motion blur and clear boundaries) at test time. We do not use a query bank at inference: we have a fixed, one HC, one AC and one FL frame to use as fixed queries, which adds only a trivial memory overhead relative to the model weights.

Without the query curricula, a standard 3-class temporal detector must use the same few positives to support learning appearance of 3 heterogeneous classes and background, and finding them temporally from fleeting, off-axis data – optimising too many tasks at once. But with the easiest curriculum case where the visual query is the same positive frame from the video; the model learns high query-frame similarity minimises loss and learns the mechanics of temporal

localisation without needing to learn diverse biometry appearances (just pixel match). Reducing the similarity of the query to the target frame allows learning more diverse AC/HC/FL representations after having been initialised with strong temporal priors and inductive bias.

Thus empirically, without this curriculum, video models fail to converge and only frame-level detectors converged with such a small dataset size, because they didn't need to learn temporal connection parameters. We know from sonographers that temporal context for biometry plane detection is crucial, especially to detect poor quality frames. This curriculum training paradigm cannot be done with the simple classification set up, but requires a visual query.

Another benefit of the query use is it allows using spatio-temporal pretraining from natural videos / other datasets – the spatio-temporal relationship knowledge between query and video should be more general than encoder weights - it is just temporal similarity matching.

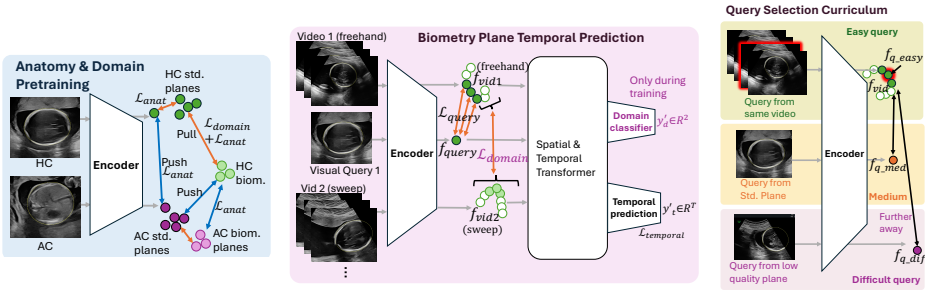

Fig. 2: Model architecture and two-stage training pipeline. Orange arrows shows loss pulls together, blue arrows shows loss will repel. Light green dots are HC planes from blind-sweep and dark green are HCs from freehand in latent space.

Model Details. We use a two-stage training process of pretraining the visual encoder on the large image dataset, then the whole network on the videos. Although the architecture could be trained end-to-end in a single stage, in our limited data regime, we find that pretraining not only boosts performance, but also drastically helps with model convergence. We use encoder ε to process each frame and visual query independently. The video features ($f_{vid} = \varepsilon(x_{vid}) \in \mathbb{R}^{T \times f_H \times f_W \times f_C}$) and query features ($f_q = \varepsilon(x_q) \in \mathbb{R}^{f_H \times f_W \times f_C}$) from the encoder are then passed through a spatial transformer and then a temporal transformer. The result is then passed through an MLP and Maxpool layer to generate a binary prediction for each frame $y' \in \mathbb{R}^T$. The temporal transformer weights were initialised from [11] pretrained on Ego4D.

Temporal Loss. Focal loss was used between the per-frame binary predictions and labels. ($\mathcal{L}_{temporal} = -\alpha(1-y'_t)^\gamma log(y'_t)$)[15] with $\gamma = 3, \alpha = 0.25$ due to class imbalance. Other loss functions were explored but were less effective (Table 2).

Visual Encoder. We pretrained the encoder to maximise the cosine similarity of feature vectors of the same anatomy class ($\mathcal{L}_{anat}$ in Figure 2). This was done using the n≈800 standard plane frame data, in a supervised contrastive manner using the SINCERE loss [6] (a supervised contrastive loss [12]).

"The anatomy taking up at least half the image" is part of the criteria for standard planes [17], so whilst blind-sweep videos are always zoomed out to increase visual coverage. To ensure learned features could transfer well from the plentiful standard planes to the smaller biometry dataset, we promote scale invariance in our encoder via (a) heavy zoom augmentations, (b) using Swin-Tiny[16] instead of ViT (which has a fixed patch size), and (c) using connections between all scales of the Swin (AvgPooling all Swin stages to the smallest H,W feature size before channel-wise concatenation to the deepest Swin layer).

Query Similarity Loss. During full network training, we add a contrastive loss (InfoNCE [20]) between f^+_{vid} (the features of the biometry frames from the video) and f_q (the visual query feature). We refer to this as $\mathcal{L}_{query}$ in Figure 2. Standard supervised contrastive losses [12] are calculated symmetrically for all samples in the batch so as not to favour a particular class for multi-class classification, however we only want the query and biometry frames to cluster together, and not the background frames to cluster to each other. Thus we use a mask so the loss is only calculated between the biometry planes and query. This neatly uses the inherent asymmetry in the InfoNCE loss to push the positives away from negatives (background), and pulls together positive classes (biometry planes) without pulling together the background classes. $\mathcal{L}_{query} = -\log \frac{\exp(\text{sim}(f^+_{vid}, f_q)/\tau)}{\sum_{k\in\mathcal{N}} \exp(\text{sim}(f^-_{vid_k}, f_q)/\tau)}$ (k iterates over all negative background frames $\mathcal{N}$).

Domain Adaptation. To encourage learning of features that are indifferent to the acquisition domain (sweep versus freehand), we add a domain classifier head and use a gradient reversal layer to maximise the loss of this classification [8]. This was done during the pretraining step with focal loss due to class imbalance, and with binary cross-entropy loss during the full network training. $\mathcal{L}_{domain\_vid} = -(y_d \log(y'_d) + (1 - y_d) \log(1 - y'_d))$.

Curriculum Learning. Our easiest curriculum is to use the query image from the same video clip (which is the frame we are trying to detect), then for the next curriculum we augment this query. Next we use non-augmented queries from high-quality standard plane images, not from the same video (manually picked by a non-expert (lead author)). Next augment, and then next we use biometry planes of progressively decreasing subjective quality as the query (see Figure 1). Here we push the model beyond the testing requirements (low quality, augmented, biometry plane query) so to learn even more robust representations.

Data Loading. Temporal augmentation was done by randomly sampling clips at intervals of 1–5 frames for sweeps and 1–15 frames for freehand videos. We enforce at least one positive per T=30 frame training clip. We also employ a

Table 1: Detection performance on blind-sweep US. All metrics were evaluated at a fixed sensitivity (TPR) of 0.75. FL is omitted from the main table as baselines failed to converge (AUROC $\approx$ 0.55), whereas our model achieved AUROC = 0.903 and AUPRC = 0.053 for FL.

Model	**AUROC ↑**		**TNR ↑**		**Sweep TNR ↑**		**AUPRC ↑**	
	HC	AC	HC	AC	HC	AC	HC	AC
VGG16 (Pretrained)	0.855	0.923	0.93	0.89	0.85	0.83	0.054	0.115
VQLoC (Vanilla)	0.855	0.834	0.79	0.75	0.75	0.72	0.055	0.013
VQLoC (Curric.)	0.912	0.914	0.87	0.84	0.84	0.82	0.046	0.022
TimeSFormer	0.759	0.753	0.71	0.60	0.61	0.64	0.005	0.011
Ours (Full)	**0.973**	**0.979**	**0.95**	**0.93**	**0.89**	**0.84**	**0.114**	**0.135**
Inter-annotator Agreement*	—	—	0.999	0.999	0.95	0.93	—	—

* Inter-annotator agreement is shown only to indicate annotation variability, not as a model-performance upper bound. The second reader identified substantially fewer planes than the primary annotator, with frame-level TPR of AC: 0.078, HC: 0.118 and FL: 0.127. Sweep-level TPR was AC: 0.34, HC: 0.42, FL: 0.10

balanced batching strategy, mixing high-prevalence freehand with sparse sweep videos to stabilise training and prevent majority-class bias.

4 Experiments and Results

All reported blind-sweep test results use a participant-disjoint split. The test set contains 23 participants, 57 unique annotated biometry planes.

The VGG-16 baseline was initialised with weights from [7], which pretrained using a SIM-CLR scheme and finetuned for standard plane detection; [7] reported highly competitive performance on this task. We finetuned on the biometry plane detection task. TimeSFormer did not converge without curriculum learning, whereas VQLoC did; so we report VQLoC both with and without our curriculum. We found the best performance when they were initialised using pretrained weights on SomethingSomethingV2 and the Ego4D Visual Query Challenge, respectively.

[19] is conceptually related to our query-frame attention formulation, but its query-selection module is not applicable to blind-sweep biometry: unlike spatially constrained fetal-heart videos, these sweeps provide no reliable cue for selecting a similar AC/HC/FL query during inference. Its temporal uncertainty-aware localization loss is related to the start/end KL loss ablated in Table 2.

We also compare these to the inter-annotator agreement. A second sonographer found less than a quarter of the annotations of the first sonographer, which highlights the task difficulty. For a fairer comparison, instead of using all videos, the inter-annotator agreement was from the subset of participants for whom the second sonographer found planes.

Metrics. We report frame-level AUROC and AUPRC one-vs-all for each anatomy, e.g. AC versus all non-AC frames and HC versus all non-HC frames. For acquisition-

time screening, we also report TNR (TNR = $\frac{TN}{TN+FP}$) at a fixed sensitivity of 0.75 (TPR=0.75) as a quality screening tool, we are interested in detecting scans that don't contain any biometry planes i.e. maximise TNR. This operating point prioritises rejecting scans likely to lack usable biometry content, for which re-acquisition is feasible. However, we recognise prioritising the TPR is a much more difficult task due to the low prevalence, thus we provide full AUPRC and AUROC values.

However, because positives constitute $< 0.1\%$ of frames, AUPRC is strongly affected by prevalence and missing annotations. We therefore report them for transparency, but interpret them as frame-level ranking metrics for model performance comparison, rather than expected clinical false-alarm rates. For real-world quality screening, the clinically relevant unit is the sweep rather than an individual frame (you retake the whole sweep not the frame). Thus we report sweep level true negative rate too.

Results. Our method outperforms the others we tested against in all fields except inter-annotator true negative rate (unsurprising as the second sonographer found much less frames). Surprisingly, the VGG16 performs better than most video models in terms of AUPRC. The difficulty of the task is reflected in the PR curves – the extremely low prevalence makes it very hard to get any precision. At such low scores for AUPRC, it is strongly influenced by a few well-classified images. If the model is very confident for just a few images, the top left of the PR curve provides a lot of area under the curve (when it only counts for very few images).

Definitionally, the second sonographer's "false positives" are not actually false positives; they were picked by a sonographer specifically because they are feasible biometry planes, and from these planes, all led to estimated GA within 14 days of true GA [22]. Thus, the dataset contains missingness – the sonographers picked one or two planes per patient instead of finding all possible biometry planes. This not only gives mixed training signals, but also underestimates the model's performance. Fig. 1 (right) shows an example "false positive" that appears biometry-plausible.

Ablations Table 2. We believe that without the curriculum learning, the domain discrimination and query-sim loss create competing losses as regularisation during training, causing instability, which is why the model failed to converge, whilst it converged for VQLoC and other smaller models. We also find that using the start and end time KL loss (instead of per frame binary classification), although it shows promise in literature [25], the fleeting and sometimes non-contiguous appearances of biometry-planes is not conducive for this loss.

5 Conclusion

We introduced biometry plane detection in blind sweep obstetric ultrasound as an application-driven quality assessment task, motivated by the need to ensure

Table 2: Ablation study (mean of AC and HC) on blind-sweep videos.

Model Variation	AUROC	AUPRC
Encoder Architectures		
Swin-FPN (Ours)	**0.976**	**0.111**
ViT-B	0.817	0.012
Swin (no FPN)	0.949	0.053
Swin (no pretrain, no FPN)	0.921	0.055
Component Analysis		
No Query-Sim Loss	0.956	0.081
No Domain-Discrimination	0.959	0.068
No Curriculum Learning†	0.550	—
KL Loss (vs. BCE Temporal)†	0.560	—

†*Indicates training failure to converge.*

sweeps contain clinically meaningful content for biometry-based gestational age estimation. We proposed a visual-query-guided spatio-temporal model trained with an ultrasound-specific curriculum, together with scale-robust visual encoding and domain-adversarial learning to transfer knowledge from freehand data to sparsely labelled sweeps. Experiments show consistent improvements over image and video baselines, while highlighting the remaining challenge posed by extreme class imbalance and annotation variability. We expect this direction to support practical sweep screening and rapid re-acquisition when biometric information is absent, but larger multi-centre validation and more exhaustive annotation are needed before deployment.

Acknowledgments. This research was supported by the Oxford EPSRC Centre for Doctoral Training in Health Data Science (EP/S02428X/1). The data was supported by the CALOPUS project (Global Challenge Research Fund EP/R013853/01), and the ERC Project PULSE (ERC-ADG-2015 69458). J. Jiao is supported by an Amazon Research Award.

Disclosure of Interests. The authors have no competing interests to declare.

References

1. Abuhamad, A., Zhao, Y., Abuhamad, S., Sinkovskaya, E., Rao, R., Kanaan, C., Platt, L.: Standardized six-step approach to the performance of the focused basic obstetric ultrasound examination. American Journal of Perinatology **33**, 90–98 (8 2015). https://doi.org/10.1055/s-0035-1558828
2. Arroyo, J., Marini, T.J., Saavedra, A.C., Toscano, M., Baran, T.M., Drennan, K., Dozier, A., Zhao, Y.T., Egoavil, M., Tamayo, L., Ramos, B., Castaneda, B.: No sonographer, no radiologist: New system for automatic prenatal detection of fetal biometry, fetal presentation, and placental location. PLoS ONE **17** (2 2022). https://doi.org/10.1371/journal.pone.0262107
3. Baumgartner, C.F., Kamnitsas, K., Matthew, J., Fletcher, T.P., Smith, S., Koch, L.M., Kainz, B., Rueckert, D.: Sononet: Real-time detection and localisation of fetal standard scan planes in freehand ultrasound. IEEE Transactions on Medical Imaging **36**(11), 2204–2215 (2017). https://doi.org/10.1109/TMI.2017.2712367

4. Darmstadt, G.L., Lee, A.C., Cousens, S., Sibley, L., Bhutta, Z.A., Donnay, F., Osrin, D., Bang, A., Kumar, V., Wall, S.N., et al.: 60 million non-facility births: who can deliver in community settings to reduce intrapartum-related deaths? International Journal of Gynecology & Obstetrics **107**, S89–S112 (2009)
5. Dougherty, A., Kasten, M., DeSarno, M., Badger, G., Streeter, M., Jones, D.C., Sussman, B., DeStigter, K.: Validation of a telemedicine quality assurance method for point-of-care obstetric ultrasound used in low-resource settings. Journal of Ultrasound in Medicine **40**(3), 529–540 (2021)
6. Feeney, P., Hughes, M.C.: Sincere: Supervised information noise-contrastive estimation revisited. arXiv preprint arXiv:2309.14277 (2023)
7. Fu, Z., Jiao, J., Yasrab, R., Drukker, L., Papageorghiou, A.T., Noble, J.A.: Anatomy-aware contrastive representation learning for fetal ultrasound. In: European Conference on Computer Vision. pp. 422–436. Springer (2022)
8. Ganin, Y., Lempitsky, V.: Unsupervised domain adaptation by backpropagation. In: International conference on machine learning. pp. 1180–1189. PMLR (2015)
9. Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X., et al.: Ego4d: Around the world in 3,000 hours of egocentric video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18995–19012 (2022)
10. Guo, J., Tan, G., Wu, F., Wen, H., Li, K.: Fetal ultrasound standard plane detection with coarse-to-fine multi-task learning. IEEE Journal of Biomedical and Health Informatics **27**(10), 5023–5031 (2022)
11. Jiang, H., Ramakrishnan, S.K., Grauman, K.: Single-stage visual query localization in egocentric videos. Advances in Neural Information Processing Systems **36** (2024)
12. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. Advances in neural information processing systems **33**, 18661–18673 (2020)
13. Lee, C., Willis, A., Chen, C., Sieniek, M., Watters, A., Stetson, B., Uddin, A., Wong, J., Pilgrim, R., Chou, K., et al.: Development of a machine learning model for sonographic assessment of gestational age. JAMA network open **6**(1), e2248685–e2248685 (2023)
14. Lee, L.H., Bradburn, E., Craik, R., Yaqub, M., Norris, S.A., Ismail, L.C., Ohuma, E.O., Barros, F.C., Lambert, A., Carvalho, M., et al.: Machine learning for accurate estimation of fetal gestational age based on ultrasound images. NPJ digital medicine **6**(1), 36 (2023)
15. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 2980–2988 (2017)
16. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows (2021), https://arxiv.org/abs/2103.14030
17. Loughna, P., Chitty, L., Evans, T., Chudleigh, T.: Fetal size and dating: Charts recommended for clinical obstetric practice. Ultrasound **17**(3), 160–166 (2009). https://doi.org/10.1179/174313409X448543, https://doi.org/10.1179/174313409X448543
18. Marini, T.J., Oppenheimer, D.C., Baran, T.M., Rubens, D.J., Toscano, M., Drennan, K., Garra, B., Miele, F.R., Garra, G., Noone, S.J., et al.: New ultrasound telediagnostic system for low-resource areas: Pilot results from peru. Journal of Ultrasound in Medicine **40**(3), 583–595 (2021)

19. Mishra, D., Saha, P., Zhao, H., Patey, O., Papageorghiou, A.T., Noble, J.A.: STANLOC: Visual query-based video clip localization for fetal ultrasound sweep videos. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. pp. 742–752. Springer Nature Switzerland (2024)
20. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
21. 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.: Isuog practice guidelines: ultrasound assessment of fetal biometry and growth. Ultrasound in obstetrics & gynecology **53**(6), 715–723 (2019)
22. Self, A., Marsano, B., Papageorghiou, A.: Ep02. 01: Machine learning enabled point-of-care scanning protocols: a validation study of the calopus ultrasound protocol. Ultrasound in Obstetrics & Gynecology **60** (2022)
23. Stringer, J.S., Pokaprakarn, T., Prieto, J.C., Vwalika, B., Chari, S.V., Sindano, N., Freeman, B.L., Sikapande, B., Davis, N.M., Sebastião, Y.V., et al.: Diagnostic accuracy of an integrated ai tool to estimate gestational age from blind ultrasound sweeps. Jama **332**(8), 649–657 (2024)
24. Toscano, M., Marini, T.J., Drennan, K., Baran, T.M., Kan, J., Garra, B., Dozier, A.M., Ortega, R.L., Quinn, R.A., Zhao, Y.T., et al.: Testing telediagnostic obstetric ultrasound in peru: a new horizon in expanding access to prenatal ultrasound. BMC pregnancy and childbirth **21**, 1–13 (2021)
25. Yang, A., Miech, A., Sivic, J., Laptev, I., Schmid, C.: Tubedetr: Spatio-temporal video grounding with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16442–16453 (2022)