Disentangling Clinic-Specific Acquisition Variability from Motion Dynamics in Intrapartum Ultrasound

Ilhan Aytutuldu[1][0000−0003−4237−8442], Murat Yassa[2], Pinar Birol Ilter[3], Sevval Ilhan[4,5], Rumeysa Elif Erdem[6], Umur Goktug Unlu[7], and Yusuf Sinan Akgul[4]

[1] Gebze Technical University, Vocational School of Cyber Security, Kocaeli, Turkiye
iaytutuldu@gtu.edu.tr
[2] Acibadem Kartal Hospital, Istanbul, Turkiye
[3] Kutahya City Hospital, Kutahya, Turkiye
[4] Gebze Technical University, Faculty of Engineering, Department of Computer Engineering, Kocaeli, Turkiye
[5] Istanbul Okan University, Faculty of Applied Sciences, Istanbul, Turkiye
[6] Erzurum City Hospital, Erzurum, Turkiye
[7] Erbaa State Hospital, Tokat, Turkiye

Abstract. Intrapartum ultrasound enables objective monitoring of fetal head progression during labor, yet clinical practice relies predominantly on single-frame analysis, which is sensitive to noise and fails to capture continuous anatomical dynamics. Leveraging video clips offers richer information but introduces a critical challenge in multi-clinic settings: acquisition variability arising from differences in probe motion, imaging hardware, and clinical protocols distorts latent temporal representations in ways that static harmonization methods cannot address. We collected a novel intrapartum ultrasound dataset from three clinics with heterogeneous acquisition protocols and propose a self-supervised representation learning framework that explicitly disentangles clinic-specific acquisition distortions from physiologically meaningful motion dynamics. The framework integrates temporal regularization, adversarial disentanglement, and hierarchical cross-attention to effectively separate clinic-invariant and clinic-specific representations. The proposed Hierarchical Cross-Attention encoder achieves 99.70% temporal order prediction accuracy, substantially reduces clinic-induced variability, and generalizes to fetal head segmentation on the HC18 benchmark with a Dice score of 87.23%, outperforming both an in-domain autoencoder baseline and random initialization despite its encoder being pre-trained exclusively on intrapartum video. These results establish temporal disentanglement as a principled strategy for separating acquisition variability from true motion dynamics, enabling robust representation learning across heterogeneous clinical conditions.

Keywords: Intrapartum ultrasound · Multi-center harmonization · Temporal representation learning · Self-supervised learning · Disentangled representations

1 Introduction

Intrapartum ultrasound enables objective assessment of fetal head progression during labor, providing improved accuracy and reproducibility compared to digital examination [5,9,17]. Quantitative evaluation of fetal head descent, rotation, and position supports clinical decision-making, including detection of obstructed labor and delivery planning. While current practice and most automated methods rely predominantly on single-frame analysis [5,9,2], labor is a longitudinal process in which each subject undergoes continuous physiological change across multiple timepoints — making repeated temporal assessment more informative than isolated single-frame evaluations. Fetal head progression is inherently dynamic, and meaningful clinical information is encoded in motion patterns across time rather than in static appearance alone.

A key challenge in multi-clinic ultrasound analysis is acquisition variability arising from differences in hardware, probe configuration, operator technique, and imaging protocols [19,6]. Existing harmonization and domain adaptation methods, including ComBat [7], adversarial adaptation [8], and feature alignment [18], primarily address static appearance differences. Yet because clinically meaningful information resides in temporal dynamics, clinic-specific variability can distort latent temporal trajectories in ways static harmonization fundamentally cannot address (Fig. 1). This motivates explicitly modeling and suppressing acquisition-induced temporal distortions as a prerequisite for reliable longitudinal representation learning in multi-center obstetric imaging.

Self-supervised learning (SSL) enables transferable representation learning without manual annotation [3,10,1] and has shown strong performance in medical imaging and ultrasound [20,16,4]. Temporal representation learning further improves robustness by modelling motion-dependent structure [15,12], yet existing methods entangle physiological motion with acquisition variability. Disentangled representation learning offers a principled framework for separating invariant and domain-specific factors [13,14], but its integration with temporal modelling for multi-clinic ultrasound remains unexplored.

To address this, we collect a novel intrapartum ultrasound dataset from three clinics with heterogeneous acquisition protocols and propose a self-supervised framework that explicitly disentangles clinic-specific temporal distortions from physiologically meaningful motion dynamics, integrating temporal regularization, adversarial disentanglement, and hierarchical cross-attention for robust multi-center representation learning, with the goal of establishing a methodological foundation for future clinically robust AI-driven labor monitoring systems.

2 Method

This study was approved by the Bakırköy Dr. Sadi Konuk Training and Research Hospital Clinical Research Ethics Committee (Protocol No: 2025/408). All data were collected prospectively under informed consent in accordance with the Declaration of Helsinki.

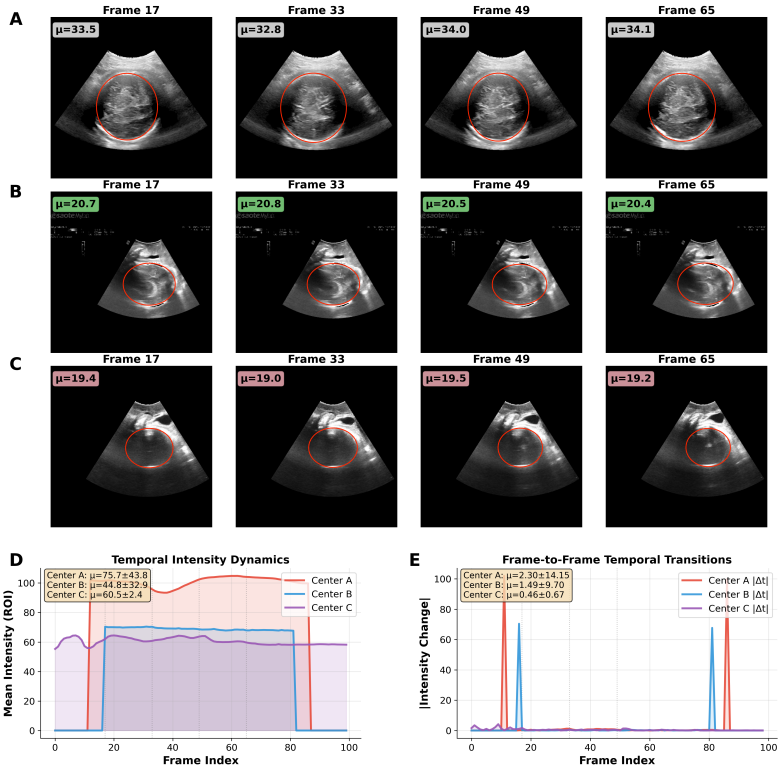

Fig. 1. (A–C) Representative frames from three different centers show pronounced differences in probe geometry, field-of-view, and brightness (μ values inset). (D) Temporal intensity dynamics reveal center-specific mean levels and variance profiles across 100 frames. (E) Frame-to-frame absolute transitions ($|\Delta_t|$) expose highly center-dependent temporal instability, with example A showing large sporadic spikes ($\mu = 2.30 \pm 14.15$) compared to near-stationary behavior in example C ($\mu = 0.46 \pm 0.67$), motivating temporal rather than purely static harmonization.

We collected 360 video clips from 30 subjects across three clinical centers with heterogeneous acquisition protocols. All clips from a given subject were assigned exclusively to a single partition to prevent data leakage, using a stratified 70/15/15 subject-level split (21 subjects for training, 4 for validation, 5 for testing), with leave-one-center-out cross-validation for generalization assessment.

We propose a self-supervised framework that suppresses center-induced variability in multi-center intrapartum ultrasound while preserving physiologically meaningful temporal dynamics (Fig. 2). Our key insight is that acquisition variability manifests as distortions in latent temporal trajectories rather than purely static appearance differences. We address this by regularizing latent temporal transitions and structuring the latent space to separate center-invariant and center-specific components across five encoder architectures of increasing capac-

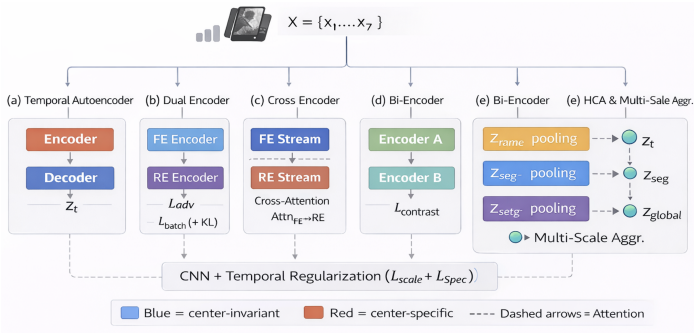

Fig. 2. Overview of the proposed temporally regularized representation learning framework. Detailed architecture and code information can be found at: `https://github.com/miccai-intrapartum/miccai_intrapartum`

ity: (1) *Temporal Autoencoder* with temporal regularization; (2) *Dual Encoder* with explicit disentanglement into a center-invariant stream (FE: fixed-effect encoder) and a center-specific stream (RE: random-effect encoder); (3) *Cross Encoder* with bidirectional cross-attention; (4) *Bi-Encoder* with contrastive temporal learning; and (5) *HCA Encoder* (Hierarchical Cross-Attention) with hierarchical multi-scale aggregation. All architectures share the temporal regularization foundation and are trained using appropriate subsets of Eq. 6.

Given a sequence $\mathbf{X} = \{x_1, \ldots, x_T\}$, encoder f_θ maps each frame to a latent embedding $\mathbf{z}_t = f_\theta(x_t)$, forming trajectory $\mathbf{Z} = \{\mathbf{z}_1, \ldots, \mathbf{z}_T\}$ with transitions $\Delta\mathbf{z}_t = \mathbf{z}_{t+1} - \mathbf{z}_t$.

Temporal *Regularization:* Two complementary losses stabilize latent trajectories. Scale regularization penalizes variance in transition magnitudes:

$$\mathcal{L}_{\text{scale}} = \frac{1}{T-1} \sum_{t=1}^{T-1} \left(\|\Delta\mathbf{z}_t\| - \mu \right)^2, \quad \mu = \frac{1}{T-1} \sum_{t=1}^{T-1} \|\Delta\mathbf{z}_t\|, \tag{1}$$

preventing erratic latent motion from acquisition noise. Spectral regularization suppresses high-frequency instability:

$$\mathcal{L}_{\text{spec}} = \sum_f w_f \left| \mathcal{F}(\Delta\mathbf{z})_f \right|^2, \tag{2}$$

where $\mathcal{F}$ is the Fourier transform and w_f increases with frequency. A decoder g_ϕ enforces semantic retention via reconstruction loss $\mathcal{L}_{\text{rec}} = \frac{1}{T} \sum_t \|x_t - g_\phi(\mathbf{z}_t)\|^2$.

Adversarial *Disentanglement:* A center-invariant encoder f^{FE} and center-specific encoder f^{RE} decompose the representation:

$$\mathbf{z}_t^{FE} = f^{FE}(x_t), \quad \mathbf{z}_t^{RE} = f^{RE}(x_t). \tag{3}$$

A center classifier $C(\cdot)$ drives functional specialization: f^{FE} is trained adversarially to suppress center information ($\mathcal{L}_{\text{adv}} = -\mathbb{E}[\log C(\mathbf{z}_t^{FE})]$), while f^{RE} explicitly captures acquisition variability ($\mathcal{L}_{\text{batch}} = \mathbb{E}[\text{CrossEntropy}(C(\mathbf{z}_t^{RE}), c_t)]$). A variational formulation on the RE stream prevents posterior collapse:

$$\mathcal{L}_{\text{KL}} = D_{\text{KL}}\left(q(\mathbf{z}^{\text{RE}} \mid x_t) \,\|\, \mathcal{N}(0, I)\right). \tag{4}$$

Hierarchical Cross-Attention: Frame-level embeddings are pooled into segment- and sequence-level representations, with cross-attention enabling multi-scale context-aware refinement:

$$\mathbf{z}_{\text{agg}} = \text{Attention}(\mathbf{z}_{\text{frame}}, \mathbf{z}_{\text{segment}}, \mathbf{z}_{\text{global}}). \tag{5}$$

The full training objective is:

$$\mathcal{L} = \mathcal{L}_{\text{rec}} + \lambda_{\text{scale}}\mathcal{L}_{\text{scale}} + \lambda_{\text{spec}}\mathcal{L}_{\text{spec}} + \lambda_{\text{adv}}\mathcal{L}_{\text{adv}} + \lambda_{\text{batch}}\mathcal{L}_{\text{batch}} + \lambda_{\text{KL}}\mathcal{L}_{\text{KL}}. \tag{6}$$

Evaluation: Temporal order prediction (TOP) probes temporal structure by classifying ordered versus shuffled sequences from latent transition statistics $\|\Delta\mathbf{z}_t\|$, evaluated exclusively on unseen subjects to ensure generalization rather than memorization. Representations that capture meaningful motion dynamics produce structured, low-variance transitions for ordered sequences and irregular transitions for shuffled ones — making TOP accuracy a direct measure of how well the latent trajectory encodes physiologically consistent motion. Center invariance is quantified via the Average Silhouette Width (ASW) computed on the FE latent space using center identity as the grouping variable: lower ASW indicates greater overlap between center clusters, confirming successful suppression of center-specific information. Cross-domain generalization is assessed on the HC18 benchmark [11] — an antenatal fetal head segmentation dataset comprising 2D B-mode ultrasound images with pixel-level masks — used here as an external representation quality probe rather than a clinical performance claim. Both datasets share the same imaging modality and anatomical target, making HC18 a principled choice for evaluating anatomical transferability of intrapartum-trained representations. Evaluation comprises two complementary assessments: segmentation Dice score, obtained by fine-tuning a U-Net decoder on top of the frozen-then-unfrozen encoder using HC18 ground truth masks, and embedding robustness via cosine similarity under Gaussian perturbations of increasing magnitude.

3 Results

Across the five architectures, TOP accuracy rises steadily from 95.51% to 99.70% as temporal modelling capacity increases (Table 1). The Bi-Encoder is the exception, dropping to 91.62%. Unlike the other architectures, it relies solely on contrastive objectives to enforce temporal consistency without explicit regularization of latent transitions, which appears insufficient to suppress acquisition-induced temporal distortions in this multi-center setting.

Table 1. Progressive architecture comparison on temporal order prediction (TOP). All models share identical CNN backbones and training protocols, differing only in architectural mechanisms.

Architecture	TOP Acc. (%)	95% CI	Key Mechanism
Temporal AE	95.51	[93.30, 97.50]	Scale + spectral regularization
Dual Encoder	94.41	[93.85, 98.60]	FE/RE adversarial decomposition
Cross Encoder	98.04	[96.93, 99.44]	Bidirectional cross-attention
Bi-Encoder	91.62	[88.83, 94.69]	Contrastive temporal learning
HCA Encoder	**99.70**	[99.10, 100.0]	3-level hierarchical aggregation

Table 2. Leave-one-center-out (LOCO) validation evaluating generalization to unseen acquisition centers. Each center is used once as a held-out test set while training on the remaining centers.

Test Center	HCA Acc. (%)	Temporal AE Acc. (%)	Δ (%)
Center A	**98.6**	94.2	+4.4
Center B	**97.4**	92.1	+5.3
Center C	**99.1**	95.0	+4.1
Mean	**98.4**	93.8	+4.6

Under leave-one-center-out evaluation, the HCA encoder maintains strong performance on each unseen center, averaging 98.4% against 93.8% for the Temporal Autoencoder (Table 2). Gains are consistent across centers, ranging from 4.1 to 5.3 percentage points.

To isolate the contribution of each component, we train incrementally from reconstruction only to the full model (Table 3). Reconstruction alone yields near-chance TOP accuracy (52.8%), confirming that standard autoencoders do not inherently preserve temporal structure. Scale regularization provides the largest single gain (52.8% to 85.9%), followed by spectral regularization (91.4%) and adversarial training (ASW 0.32 to 0.28). The full HCA model reaches 99.3% TOP and ASW of 0.24, demonstrating that each component addresses a complementary aspect of center-induced temporal variability. Against external baselines, both proposed variants outperform ComBat and DANN on TOP accuracy and center invariance, with the Temporal Autoencoder already competitive at lower computational cost (Table 4).

Table 3. The contribution of each component to temporal order prediction accuracy and center invariance on the test set.

Configuration	TOP Acc. (%)	ASW (FE)
Reconstruction only (L_{rec})	52.8	0.41
+ Scale regularization (L_{scale})	85.9	0.36
+ Spectral regularization (L_{spec})	91.4	0.32
+ Adversarial loss (L_{adv})	96.1	0.28
+ Hierarchical aggregation (full HCA)	**99.3**	**0.24**

Table 4. Comparison with existing harmonization and domain adaptation methods on the held-out test set.

Method	TOP Acc. (%)	ASW (FE)	Training
Raw (no training)	47.3 ± 0.0	0.40	–
ComBat + Autoencoder	88.6 ± 2.0	0.30	Fast
DANN	92.3 ± 1.6	0.31	Slow
Temporal AE (ours)	95.5 ± 1.1	0.25	Fast
HCA (ours)	**99.3 ± 0.4**	**0.24**	Moderate

On HC18, HCA embeddings degrade substantially less under Gaussian noise than the in-domain autoencoder at every noise level tested (Table 5). At $\sigma = 0.20$, cosine similarity remains at 0.907 compared to 0.713, with the gap widening consistently as noise intensity increases ($p < 10^{-22}$ throughout).

Table 5. Noise robustness evaluation on HC18 dataset using cosine similarity under Gaussian perturbations.

Noise σ	PlainAE	HCA	t-statistic	p-value
0.02	0.9972 ± 0.0017	$\mathbf{0.9995 \pm 0.0002}$	12.84	8.50×10^{-23}
0.05	0.9775 ± 0.0154	$\mathbf{0.9973 \pm 0.0009}$	12.52	3.99×10^{-22}
0.10	0.9120 ± 0.0569	$\mathbf{0.9874 \pm 0.0039}$	12.94	5.28×10^{-23}
0.15	0.8241 ± 0.0984	$\mathbf{0.9605 \pm 0.0110}$	13.50	3.53×10^{-24}
0.20	0.7126 ± 0.1368	$\mathbf{0.9068 \pm 0.0226}$	14.20	1.31×10^{-25}

To probe whether the learned representations encode anatomically transferable structure, we use HC18 as an external evaluation benchmark. A fully supervised in-domain model achieves approximately 92–93% Dice, establishing the performance ceiling. Despite being trained exclusively on intrapartum video, the HCA encoder (87.23%) outperforms both an in-domain autoencoder baseline (83.12%) and random initialization (80.64%), suggesting that temporal regularization encodes anatomically meaningful structure beyond the intrapartum domain. Complementing this, HCA embeddings show significantly greater robustness under Gaussian perturbation at every noise level tested (Table 5), further supporting the transferability of the learned representations.

4 Discussion

This work demonstrates that center-induced variability in multi-center intrapartum ultrasound manifests primarily as dynamic distortions in latent temporal trajectories, and that explicitly regularizing these trajectories is more effective than conventional static harmonization. The progressive architecture comparison (Table 1) shows consistent TOP gains from 95.51% to 99.70% with increasing temporal modelling capacity, confirming that sequential dependencies encode diagnostically meaningful information beyond static appearance [15,12].

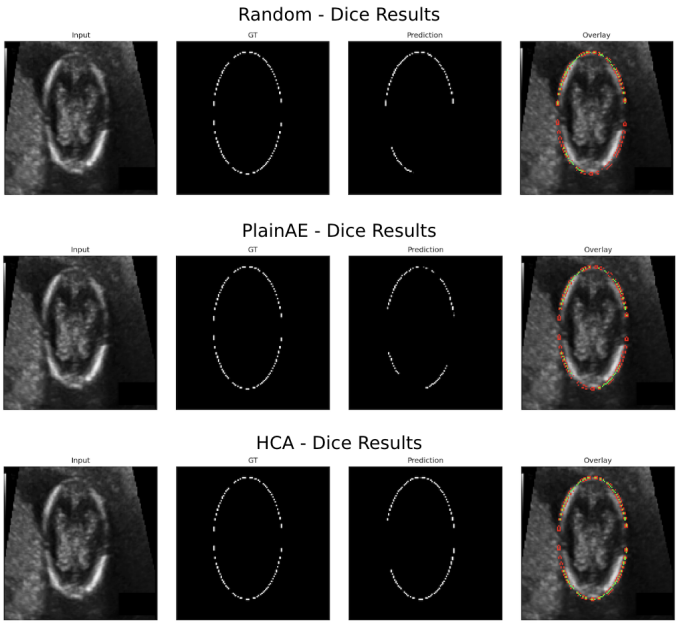

Fig. 3. Qualitative fetal head segmentation results on HC18. Each row shows input, ground truth, prediction, and overlay. The HCA encoder (bottom row) produces substantially more complete cranial boundary delineation compared to random initialization (top) and PlainAE (middle), with the most notable improvement in the lower cranial region where both baselines show partial or missing contours.

The ablation analysis (Table 3) provides mechanistic validation of each component. Scale regularization alone improves TOP from 52.8% to 85.9%, supporting the hypothesis that center-dependent acquisition differences disrupt latent temporal consistency. The dual-stream design achieves clear functional separation, with RE and FE streams reaching ASW scores of 0.725 and 0.24, respectively, indicating effective isolation of center-specific variability while preserving invariant representations [13,7]. In contrast, ComBat and DANN operate on global statistics without temporal awareness and cannot achieve comparable structured decomposition (Table 4).

Cross-domain evaluation on HC18 further demonstrates the generality of the learned representations. Despite being trained only on intrapartum video, the encoder improves fetal head segmentation performance over an in-domain autoencoder (87.23% vs 83.12% Dice) and maintains significantly greater robustness under Gaussian perturbations (Table 5, $p < 10^{-22}$). This suggests that temporal regularization actively organizes the latent space into anatomically transferable representations rather than merely suppressing nuisance variability [1,20]. Qualitative results (Fig. 3) visually confirm improved structural completeness.

Conventional harmonization and domain adaptation methods address static distribution shifts [7,8,18], but the baseline comparison (Table 4) highlights their limitations in temporal settings. The LOCO evaluation (Table 2) shows consistent improvements of 4.1–5.3 percentage points on unseen centers, supporting robustness across heterogeneous acquisition environments [9,5].

Limitations include the modest dataset size and three-center scope. Future work should validate on larger cohorts, explore transformer-based temporal architectures to capture longer-range dependencies [3,10], investigate deployment-oriented preprocessing strategies such as adaptive frame trimming and more comprehensive loss-component ablations, and benchmark against contrastive self-supervised methods such as CPC and MoCo [15,10] in multi-center temporal settings.

5 Conclusion

We presented a self-supervised framework for disentangling center-induced temporal distortions from physiologically meaningful motion dynamics in multi-center intrapartum ultrasound. Through progressive architectural refinement, the framework achieves 99.70% temporal order prediction accuracy, effective functional separation of center-invariant and center-specific representations, and cross-domain transfer to fetal head segmentation (Dice 87.23%) without in-domain training. These results establish temporal regularization as a principled strategy for robust multi-center ultrasound representation learning.

Acknowledgments. This work received no funding.

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

References

1. Azizi, S., Mustafa, B., Ryan, F., Beaver, Z., Freyberg, J., Deaton, J., Loh, A., Karthikesalingam, A., Kornblith, S., Chen, T., et al.: Big self-supervised models advance medical image classification. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3478–3488 (2021)
2. Chen, S., Wang, H., Long, S., Bai, J., Jiang, J.: Ultrasound video segmentation of pubic symphysis and fetal head for angle of progression measurement. In: Proceedings of the 6th ACM International Conference on Multimedia in Asia. pp. 1–8 (2024)
3. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PmLR (2020)
4. Ciga, O., Xu, T., Martel, A.L.: Self supervised contrastive learning for digital histopathology. Machine learning with applications **7**, 100198 (2022)

5. Eggebø, T., Heien, C., Økland, I., Gjessing, L., Romundstad, P., Salvesen, K.: Ultrasound assessment of fetal head–perineum distance before induction of labor. Ultrasound in Obstetrics and Gynecology: The Official Journal of the International Society of Ultrasound in Obstetrics and Gynecology **32**(2), 199–204 (2008)
6. Ferraioli, G., Raimondi, A., De Silvestri, A., Filice, C., Barr, R.G.: Toward acquisition protocol standardization for estimating liver fat content using ultrasound attenuation coefficient imaging. Ultrasonography **42**(3), 446–456 (2023)
7. Fortin, J.P., Cullen, N., Sheline, Y.I., Taylor, W.D., Aselcioglu, I., Cook, P.A., Adams, P., Cooper, C., Fava, M., McGrath, P.J., et al.: Harmonization of cortical thickness measurements across scanners and sites. Neuroimage **167**, 104–120 (2018)
8. Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., March, M., Lempitsky, V.: Domain-adversarial training of neural networks. Journal of machine learning research **17**(59), 1–35 (2016)
9. Ghi, T., Eggebø, T., Lees, C., Kalache, K., Rozenberg, P., Youssef, A., Salomon, L., Tutschek, B.: Isuog practice guidelines: intrapartum ultrasound. Ultrasound in Obstetrics & Gynecology **52**(1), 128–139 (2018)
10. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9729–9738 (2020)
11. van den Heuvel, T.L., de Bruijn, D., de Korte, C.L., Ginneken, B.v.: Automated measurement of fetal head circumference using 2d ultrasound images. PloS one **13**(8), e0200412 (2018)
12. Hyvarinen, A., Morioka, H.: Unsupervised feature extraction by time-contrastive learning and nonlinear ica. Advances in neural information processing systems **29** (2016)
13. Locatello, F., Bauer, S., Lucic, M., Raetsch, G., Gelly, S., Schölkopf, B., Bachem, O.: Challenging common assumptions in the unsupervised learning of disentangled representations. In: international conference on machine learning. pp. 4114–4124. PMLR (2019)
14. Mathieu, M.F., Zhao, J.J., Zhao, J., Ramesh, A., Sprechmann, P., LeCun, Y.: Disentangling factors of variation in deep representation using adversarial training. Advances in neural information processing systems **29** (2016)
15. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
16. Taleb, A., Loetzsch, W., Danz, N., Severin, J., Gaertner, T., Bergner, B., Lippert, C.: 3d self-supervised methods for medical imaging. Advances in neural information processing systems **33**, 18158–18172 (2020)
17. Tutschek, B., Braun, T., Chantraine, F., Henrich, W.: A study of progress of labour using intrapartum translabial ultrasound, assessing head station, direction, and angle of descent. BJOG: An International Journal of Obstetrics & Gynaecology **118**(1), 62–69 (2011)
18. Tzeng, E., Hoffman, J., Saenko, K., Darrell, T.: Adversarial discriminative domain adaptation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7167–7176 (2017)
19. Venkataraman, V., Browning, T., Pedrosa, I., Abbara, S., Fetzer, D., Toomay, S., Peshock, R.M.: Implementing shared, standardized imaging protocols to improve cross-enterprise workflow and quality. Journal of Digital Imaging **32**(5), 880–887 (2019)
20. Zhou, Z., Sodha, V., Pang, J., Gotway, M.B., Liang, J.: Models genesis. Medical image analysis **67**, 101840 (2021)