Dual Agreement Consistency Learning for Semi-Supervised Fetal Ultrasound Segmentation

Fangyijie Wang[1,2,✉][0009−0003−0427−368X], Guénolé Silvestre[1,3], Ziyang Wang[4], and Kathleen M. Curran[1,2][0000−0003−0095−9337]

[1] Research Ireland Centre for Research Training in Machine Learning
[2] School of Medicine, University College Dublin, Dublin, Ireland
fangyijie.wang@ucdconnect.ie
[3] School of Computer Science, University College Dublin, Dublin, Ireland
[4] School of Computer Science and Digital Technologies, Aston University, Birmingham, UK

Abstract. Maternal-fetal Ultrasound (US) is the primary imaging modality for monitoring fetal development, yet accurate automated segmentation remains challenging due to the scarcity of pixel-level annotations. To address this issue, we propose Dual Agreement Consistency Learning (DACL), a semi-supervised framework for robust fetal US image segmentation. DACL jointly trains a deployment-oriented lightweight convolutional network (1.47 M parameters) and a Transformer-based network, leveraging labeled data for supervised learning and unlabeled data via cross pseudo supervision (CPS). To enhance prediction stability, we introduce a dual-agreement consistency loss that couples pixel-wise probabilistic divergence with entropy-guided confidence alignment. Unlike conventional CPS methods that enforce agreement only at the prediction level, DACL explicitly regularizes both distributional alignment and uncertainty, thereby suppressing unreliable pseudo-labels and enabling stable cross-architecture pseudo-label learning under extreme annotation scarcity. Furthermore, an interpolation-based consistency strategy using mixup is applied to unlabeled samples to enhance robustness. Under 5% labeled data, DACL improves Dice by up to 2.77% and reduces HD95 by up to 14.69 mm compared with the strongest recent semi-supervised methods, demonstrating significant improvements in boundary accuracy on both fetal head and abdomen datasets. These results demonstrate the effectiveness of agreement-based consistency learning for annotation-efficient fetal US segmentation. Our code is on GitHub.

Keywords: Fetal Ultrasound · Segmentation · Semi-supervised Learning · Agreement-Based Consistency Learning.

1 Introduction

Ultrasound (US) imaging is widely used for prenatal evaluation of fetal growth, fetal anatomy, gestational age estimation, and pregnancy monitoring due to its portability, low cost, and non-invasive nature [15]. Accurate measurements

enable precise evaluation of fetal biometry and effective monitoring of fetal growth [3, 12]. Therefore, precise delineation of fetal structures of interest, such as the head and abdomen, is crucial for obstetricians. However, this process is patient-specific, operator-dependent, and prone to intra-and inter-operator variability, which can introduce errors in fetal biometry assessment [3,17]. Such errors may lead to inaccurate evaluation of fetal development and health, potentially resulting in missed detection of congenital diseases [11].

Recent advancements in Deep Learning (DL) have notably improved fetal US image segmentation. Several studies have explored lightweight DL models for US segmentation [7]. Nevertheless, collecting large numbers of annotated US images for training remains labor-intensive and time-consuming, requiring medical proficiency and clinical expertise for precise pixel-level labeling [14]. Recent studies [5, 9] have also explored Semi-Supervised Learning (SSL) for fetal US image analysis; however, this line of work remains relatively underexplored. Developing lightweight models with US is particularly challenging when analyzing fetal US images under limited annotation settings.

To address this challenge, we propose Dual Agreement Consistency Learning (DACL), a semi-supervised framework that enforces prediction agreement between heterogeneous models at the pixel level. DACL adopts a cross-teaching strategy involving a lightweight Convolutional Neural Network (CNN), a Transformer network, and an Exponential Moving Average (EMA) Transformer teacher. The CNN and Transformer students are independently supervised using Ground Truth (GT) annotations on labeled data, while unlabeled data are leveraged through CPS. To further improve training stability, we introduce a Dual-Agreement Consistency (DAC) loss that jointly enforces (i) probabilistic agreement via Kullback–Leibler divergence (KL) divergence and (ii) uncertainty-aware confidence agreement via entropy regularization. Unlike conventional CPS methods that primarily rely on hard pseudo-label agreement, DAC regularizes both prediction distributions and prediction uncertainty across heterogeneous CNN-Transformer architectures. In addition, we apply an interpolation-based consistency regularization strategy (mixup) to unlabeled data, enabling the teacher model to guide the Transformer student with consistent targets. Together, these components enhance the robustness of the lightweight CNN under limited annotation settings.

The main contributions of this paper are as follows: (1) We propose Dual Agreement Consistency Learning (DACL), a semi-supervised SSL framework that stabilizes cross pseudo supervision through a novel pixel-wise dual-agreement consistency strategy, jointly enforcing probabilistic alignment and confidence-aware regularization between heterogeneous models. (2) By explicitly suppressing unreliable pseudo-labels, DACL effectively leverages unlabeled data to improve fetal US segmentation, particularly enhancing boundary accuracy under limited annotation settings. (3) Extensive experiments on two fetal head and abdomen datasets demonstrate consistent improvements over recent semi-supervised methods across overlap and boundary metrics.

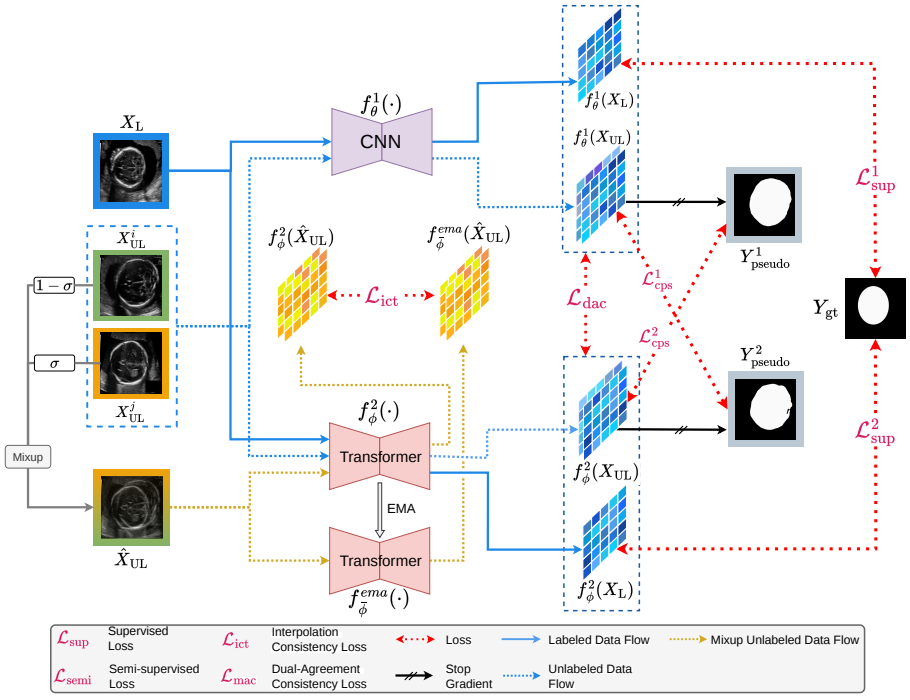

Fig. 1. Overview of the proposed Dual Agreement Consistency Learning (DACL) framework. A lightweight CNN and a Transformer are jointly trained using supervised learning on labeled data and cross pseudo supervision on unlabeled data. A dual-agreement consistency loss enforces pixel-wise distribution alignment and uncertainty-aware agreement between their predictions, while mixup-based interpolation consistency with an EMA teacher further stabilizes training.

2 Methodology

This section reviews Swin-Unet [1] and UNeXt [20]. Subsequently, we explain our semi-supervised framework DACL in the following sections 2.1, 2.2, 2.3, and 2.4. The overview of our method is illustrated in Fig.1.

We follow the original UNeXt [20] design to build our lightweight model $f_\theta^1(\cdot)$. The number of channels in the encoder path is set to $\{32, 64, 128, 160, 256\}$ to ensure computational efficiency. The $f_\phi^2(\cdot)$ model is implemented using the tiny Swin-Unet architecture introduced in the original work [1]. To match the resolution of our input images, we set the shifted window size in Swin-Unet to 7. In the encoder path of Swin-Unet, the feature dimensions of each block are $[96, 192, 384]$, while the bottleneck feature dimension is 768. In terms of computational efficiency, Swin-Unet uses 27.15M parameters and 71.17 GFLOPs, representing a high-capacity model with strong representation ability. In con-

trast, UNeXt serves as our target lightweight model, requiring only 1.47M parameters and 7.03 GFLOPs. This substantial gap in model complexity motivates our semi-supervised framework, which aims to transfer knowledge from the powerful Transformer-based model to the lightweight model, enabling accurate yet computationally efficient fetal US image analysis.

2.1 Cross-supervision between CNN and Transformer

Inspired by existing cross-supervision methods, including Deep Co-Training [13] and CPS [2], DACL adopts a similar cross-teaching strategy that couples a lightweight CNN with a Transformer-based network. Given the unlabeled training dataset $\boldsymbol{X}_{\mathrm{UL}}$, the proposed DACL produces two predictions: $f_\theta^1(\boldsymbol{X}_{\mathrm{UL}})$ and $f_\phi^2(\boldsymbol{X}_{\mathrm{UL}})$. Based on these predictions, pseudo labels for CPS are generated as $\tilde{Y}^1 = f_{\mathrm{OH}}(f_\theta^1(\boldsymbol{X}_{\mathrm{UL}}))$ and $\tilde{Y}^2 = f_{\mathrm{OH}}(f_\phi^2(\boldsymbol{X}_{\mathrm{UL}}))$. No mini-batch gradient back-propagation is performed between $f_\theta^1(\boldsymbol{X}_{\mathrm{UL}})$ and $\tilde{Y}^1$, or between $f_\phi^2(\boldsymbol{X}_{\mathrm{UL}})$ and $\tilde{Y}^2$. Here, $f_{\mathrm{OH}}(\boldsymbol{x} = \mathrm{c}) = \mathbb{1}_{[\boldsymbol{x}=\mathrm{c}]}$ denotes the one-hot encoding function used to generate pseudo labels. The semi-supervised loss, $\mathcal{L}_{\mathrm{cps}}$, is the summation term $\mathcal{L}_{\mathrm{cps}}^1 + \mathcal{L}_{\mathrm{cps}}^2$ where

$$\begin{aligned} \mathcal{L}_{\mathrm{cps}}^1 &= \mathcal{H}\Big(f_\theta^1\left(\boldsymbol{X}_{\mathrm{UL}}\right), \tilde{Y}^2\Big) + \mathcal{D}\Big(f_\theta^1\left(\boldsymbol{X}_{\mathrm{UL}}\right), \tilde{Y}^2\Big), \\ \mathcal{L}_{\mathrm{cps}}^2 &= \mathcal{H}\Big(f_\phi^2\left(\boldsymbol{X}_{\mathrm{UL}}\right), \tilde{Y}^1\Big) + \mathcal{D}\Big(f_\phi^2\left(\boldsymbol{X}_{\mathrm{UL}}\right), \tilde{Y}^1\Big). \end{aligned} \tag{1}$$

$\mathcal{H}(\cdot)$ and $\mathcal{D}(\cdot)$ denote the Cross-Entropy (CE) loss and the standard Dice based coefficient (Dice) loss, respectively. Within our framework, the Transformer-based network f_ϕ^2 is used solely during training and is not used for inference.

2.2 Interpolation Consistency Learning

To further improve the utilization of unlabeled data and the generalization of the Transformer model $f_\phi^2(\cdot)$, we encourage the predictions for interpolated unlabeled pixels to be consistent with the interpolation of the corresponding predictions [21]. We adopt a mean-teacher model $f_{\bar{\phi}}^{\mathrm{ema}}(\cdot)$, where the parameters $\bar{\phi}$ are an EMA of the student parameters ϕ. During training, ϕ is updated to enforce consistency between $f_\phi^2(\hat{\boldsymbol{X}}_{\mathrm{UL}})$ and $\mathrm{Mix}(f_{\bar{\phi}}^{\mathrm{ema}}(\boldsymbol{X}_{\mathrm{UL}}^i), f_{\bar{\phi}}^{\mathrm{ema}}(\boldsymbol{X}_{\mathrm{UL}}^j))$, where $\hat{\boldsymbol{X}}_{\mathrm{UL}} = \mathrm{Mix}(\boldsymbol{X}_{\mathrm{UL}}^i, \boldsymbol{X}_{\mathrm{UL}}^j)$. Here, Mix denotes the mixup operation with combination ratio $\sigma = 0.5$, and $i+j$ equals the unlabeled batch size B_{UL}. The consistency loss $\mathcal{L}_{\mathrm{ict}}$ is defined as $\mathcal{L}_{\mathrm{ict}} = \sum(f_\phi^2(\hat{\boldsymbol{X}}_{\mathrm{UL}}) - \mathrm{Mix}(f_{\bar{\phi}}^{\mathrm{ema}}(\boldsymbol{X}_{\mathrm{UL}}^i), f_{\bar{\phi}}^{\mathrm{ema}}(\boldsymbol{X}_{\mathrm{UL}}^j)))^2$.

2.3 Dual-Agreement Consistency

To improve robustness in semi-supervised segmentation, we introduce the DAC loss $\mathcal{L}_{\mathrm{dac}}$. This loss enforces prediction agreement between two collaborative models: a lightweight CNN f_θ^1 and a Transformer f_ϕ^2. The proposed $\mathcal{L}_{\mathrm{dac}}$ consists

of two complementary components: a pixel-wise KL loss [6], $\mathcal{L}_{\mathrm{KL}}$, and an entropy-based agreement regularization term, $\mathcal{L}_{\mathrm{ent}}$. Specifically, $\mathcal{L}_{\mathrm{KL}}$ aligns the predicted class distributions, while $\mathcal{L}_{\mathrm{ent}}$ encourages consistent and confident predictions between the two models. The proposed $\mathcal{L}_{\mathrm{dac}}$ is defined as: $\mathcal{L}_{\mathrm{dac}} = \mathcal{L}_{\mathrm{KL}} + \mathcal{L}_{\mathrm{ent}}$. $\mathcal{L}_{\mathrm{KL}}$ and $\mathcal{L}_{\mathrm{ent}}$ are defined as:

$$\mathcal{L}_{\mathrm{KL}} = \mathbb{E}\left[\frac{1}{HW}\sum_{u}\mathrm{KL}(p(u)\|q(u))\right]$$
$$\mathcal{L}_{\mathrm{ent}} = \mathbb{E}\left[\frac{1}{HW}\sum_{u}\left(H(p(u)) + H(q(u)) - H\left(\frac{p(u)+q(u)}{2}\right)\right)\right]$$

where u is a spatial pixel index (h, w), and $p(\cdot)$ and $q(\cdot)$ denote the predicted class distributions at pixel u from f_θ^1 and f_ϕ^2, respectively. Here, $H(\cdot)$ denotes the Shannon entropy of a categorical distribution [18]. The entropy-based agreement term penalizes inconsistent high-entropy predictions while favoring mutually confident low-entropy agreement between the CNN and Transformer. Therefore, unreliable pseudo-labels are suppressed without requiring explicit confidence thresholding.

2.4 The Overall Objective Function

All training losses are indicated by red dashed lines in Fig. 1. The overall training objective is a joint loss with four components: a supervised loss, $\mathcal{L}_{\mathrm{sup}}$; a cross-supervision loss, $\mathcal{L}_{\mathrm{cps}}$; an EMA-based interpolation consistency loss, $\mathcal{L}_{\mathrm{ict}}$; and a dual-agreement consistency loss, $\mathcal{L}_{\mathrm{dac}}$. The total loss is:

$$\mathcal{L}_{\mathrm{total}} = (\mathcal{L}_{\mathrm{sup}}^1 + \mathcal{L}_{\mathrm{sup}}^2) + \lambda(\mathcal{L}_{\mathrm{cps}}^1 + \mathcal{L}_{\mathrm{cps}}^2) + \tau\mathcal{L}_{\mathrm{ict}} + \beta\mathcal{L}_{\mathrm{dac}}$$

where $\{\lambda, \tau, \beta\}$ are linear trade-off hyper-parameters set to $\{2.0, 2.0, 10.0\}$, respectively. $\mathcal{L}_{\mathrm{sup}}^1$ and $\mathcal{L}_{\mathrm{sup}}^2$ are the supervision losses for $f_\theta^1(\cdot)$ and $f_\phi^2(\cdot)$ based on the labeled data $\boldsymbol{X}_{\mathrm{L}}$. They are designed with a combination of the Dice and CE losses, as follows:

$$\mathcal{L}_{\mathrm{sup}}^1 = \mathcal{H}\left(f_\theta^1\left(\boldsymbol{X}_{\mathrm{L}}\right), \boldsymbol{Y}_{\mathrm{gt}}\right) + \mathcal{D}\left(f_\theta^1\left(\boldsymbol{X}_{\mathrm{L}}\right), \boldsymbol{Y}_{\mathrm{gt}}\right)$$
$$\mathcal{L}_{\mathrm{sup}}^2 = \mathcal{H}\left(f_\phi^2\left(\boldsymbol{X}_{\mathrm{L}}\right), \boldsymbol{Y}_{\mathrm{gt}}\right) + \mathcal{D}\left(f_\phi^2\left(\boldsymbol{X}_{\mathrm{L}}\right), \boldsymbol{Y}_{\mathrm{gt}}\right)$$

cross-supervision losses $\mathcal{L}_{\mathrm{cps}}^1$ and $\mathcal{L}_{\mathrm{cps}}^2$ are given in (1). The loss $\mathcal{L}_{\mathrm{ict}}$ and $\mathcal{L}_{\mathrm{dac}}$ are defined in Section 2.2 and Section 2.3, respectively.

3 Experiments and Results

3.1 Datasets

We used two public datasets in this study: the **HC18** dataset, collected using General Electric US devices in the Netherlands [4], and the **F-Abd** dataset,

acquired by novice users using a low-cost portable probe connected to a smartphone in low-income countries [16]. The HC18 dataset contains fetal head standard planes annotated for head circumference measurement, whereas the F-Abd dataset contains images annotated with optimal planes for abdominal circumference measurement. To use the pre-trained Transformer models f_ϕ^2 and $f_{\bar{\phi}}^{\text{ema}}$ in our framework, we convert all US images to RGB format.

HC18: All images display a resolution of 800×540 pixels. It contains 500 training images (450 subjects), 50 validation images (50 subjects), and 449 test images (410 subjects). **F-Abd:** Each image has a resolution of 744×562 pixels. It includes 1084 training images (94 subjects), 139 validation images (18 subjects), and 922 test images (74 subjects).

3.2 Implementation Details

We built DACL by employing two segmentation model architectures, Swin-Unet [1] and UNeXt [20]. For the Swin-Unet model, we used pre-trained weights [1] to initialize it. We trained DACL for 400 epochs, with a labeled batch size of 1 and an unlabeled batch size of 4. A stochastic gradient descent optimizer was used with an initial learning rate of 0.01 and a momentum value of 0.9. The weight decay was set to 0.0001. Our code is implemented in Python (3.11.5) with PyTorch (2.1.2) and CUDA (12.2) on a single NVIDIA 4090 GPU. The model was evaluated on the validation set each epoch, and the best-performing UNeXt weights are saved. The above settings, including data splits, optimizer configuration, learning rate schedule, augmentations, and evaluation protocol, are directly applied to all baseline methods without modification.

Data Augmentation: In this study, data augmentations are applied to the labeled training data $\boldsymbol{X}_{\text{L}}$. The training and testing images are resized to 448×448 to facilitate computational resource demands. The detailed parameters of the data augmentation techniques are in our code repository on GitHub.

Evaluation Metrics: A range of quantitative metrics are used for testing methods, including the Dice Similarity Coefficient (DSC (%)), Jaccard (%), Average Surface Distance (ASD (mm)), and 95% Hausdorff Distance (HD95 (mm)). For fair comparison, all SSL frameworks use two networks: UNeXt and Swin-Unet for CNN-Transformer co-training, and two UNeXt networks otherwise.

3.3 Results

Quantitative Results. We compared DACL with several SSL baselines using 5% and 10% of the labeled training data. Table 1 shows that DACL achieves competitive performance across the compared methods. Notably, DACL consistently improves the performance of the lightweight UNeXt model and reveals that existing SSL frameworks remain less effective for lightweight segmentation networks under annotation scarcity. In particular, relative to LMCT (a strong baseline), DACL improves DSC by 1.07% ($p = 0.016$) on the HC18 dataset with 5% labeled data, while reducing HD95 by 4.47, indicating improved boundary

Table 1. The quantitative results on the HC18 dataset. The best results are in **bold**. The 2nd best results are in underline.

	Method	# Labeled	# Unlabeled	DSC ↑	Jaccard ↑	HD95 ↓	ASD ↓	*p*-value
FS	UNeXt [20]	500	-	93.35(8.58)	88.51(12.36)	39.11(41.87)	13.42(15.45)	0.000
	Swin-Unet [1]	(100%)		**96.96(2.88)**	**94.23(4.70)**	**9.54(11.91)**	**3.89(4.08)**	0.000
SSL	MT (2017) [19]	25 (5%)	475 (95%)	89.53(10.41)	82.37(14.22)	54.79(36.98)	18.90(15.26)	0.000
	DCT (2018) [13]			91.69(10.01)	85.92(13.76)	41.62(36.06)	14.57(14.69)	0.000
	CPS (2021) [2]			91.40(10.10)	85.42(13.70)	44.88(36.84)	15.59(15.10)	0.000
	ICT (2022) [21]			91.70(10.21)	85.97(13.87)	46.06(40.56)	15.80(15.77)	0.000
	CTCT (2022) [8]			91.57(9.84)	85.66(13.45)	43.45(36.00)	14.94(14.74)	0.000
	PCPCS (2024) [10]			92.30(9.07)	86.76(12.69)	41.13(37.76)	13.88(14.11)	0.000
	DSTCT (2024) [5]			93.97(7.78)	89.44(11.14)	26.49(30.86)	9.35(12.14)	0.000
	LMCT (2025) [22]			93.98(7.43)	89.39(10.80)	27.01(30.82)	9.49(11.96)	0.016
	DACL (Ours)			**95.05(5.41)**	**90.99(8.28)**	**22.54(28.02)**	**7.86(9.37)**	-
SSL	MT (2017) [19]	50 (10%)	450 (90%)	94.43(7.20)	90.15(10.53)	29.59(37.07)	10.08(12.45)	0.001
	DCT (2018) [13]			94.91(7.20)	91.02(10.30)	24.00(33.04)	8.55(11.56)	0.029
	CPS (2021) [2]			94.67(7.31)	90.61(10.51)	26.74(34.33)	9.18(12.02)	0.006
	ICT (2022) [21]			94.76(6.94)	90.70(10.06)	26.52(34.84)	9.42(12.40)	0.008
	CTCT (2022) [8]			95.15(6.35)	91.31(9.32)	22.60(29.93)	7.99(10.50)	0.084
	PCPCS (2024) [10]			94.26(7.90)	89.97(11.14)	27.26(34.46)	9.81(13.18)	0.000
	DSTCT (2024) [5]			95.17(6.42)	91.35(9.39)	20.37(27.15)	7.27(9.40)	0.097
	LMCT (2025) [22]			95.76(4.88)	**92.21(7.51)**	19.45(25.81)	6.78(8.17)	0.974
	DACL (Ours)			**95.77(4.16)**	92.15(6.66)	**19.42(25.29)**	**6.75(7.86)**	-

Table 2. The quantitative results on the F-Abd dataset. The best results are in **bold**. The 2nd best results are in underline.

	Method	# Labeled	# Unlabeled	DSC ↑	Jaccard ↑	HD95 ↓	ASD ↓	*p*-value
FS	UNeXt	1084	-	88.74(7.11)	80.42(10.47)	44.87(42.34)	14.94(13.20)	0.000
	Swin-Unet	(100%)		**93.24(3.79)**	**87.55(6.23)**	**15.64(19.05)**	**6.24(7.30)**	0.000
SSL	MT (2017) [19]	29 (5%)	1055 (95%)	61.78(13.69)	46.10(14.26)	112.61(46.91)	46.12(19.00)	0.000
	DCT (2018) [13]			61.22(15.89)	45.98(16.45)	113.36(42.26)	47.17(17.78)	0.000
	CPS (2021) [2]			64.43(12.92)	48.87(14.14)	101.61(38.44)	43.05(16.06)	0.000
	ICT (2022) [21]			63.14(15.02)	47.87(15.90)	125.01(52.69)	50.59(21.34)	0.000
	CTCT (2022) [8]			63.77(15.90)	48.80(17.11)	100.05(42.00)	42.62(18.81)	0.000
	PCPCS (2024) [10]			63.64(17.49)	49.02(18.51)	90.13(42.45)	38.96(18.21)	0.000
	DSTCT (2024) [5]			58.89(13.85)	43.09(13.85)	90.86(30.69)	43.11(15.41)	0.000
	LMCT (2025) [22]			62.06(13.41)	46.34(13.98)	99.34(41.52)	42.25(17.22)	0.000
	DACL (Ours)			**66.53(13.67)**	**51.43(15.54)**	**74.43(32.17)**	**35.14(15.22)**	-
SSL	MT (2017) [19]	93 (10%)	991 (90%)	73.51(21.63)	61.76(21.71)	58.42(36.44)	23.07(15.05)	0.050
	DCT (2018) [13]			73.43(18.40)	60.73(18.97)	59.17(32.33)	23.95(14.19)	0.021
	CPS (2021) [2]			**76.32(15.53)**	**63.93(18.06)**	75.38(42.83)	29.19(18.06)	0.081
	ICT (2022) [21]			71.59(20.30)	58.96(20.59)	72.40(34.64)	29.81(16.01)	0.000
	CTCT (2022) [8]			72.88(13.92)	59.19(17.05)	75.89(37.19)	31.93(16.94)	0.000
	PCPCS (2024) [10]			70.61(14.53)	56.48(17.13)	82.46(37.03)	35.17(17.31)	0.000
	DSTCT (2024) [5]			64.63(19.24)	50.45(19.32)	61.58(31.85)	29.28(15.18)	0.000
	LMCT (2025) [22]			73.21(17.36)	60.21(18.32)	73.43(40.92)	29.20(18.23)	0.002
	DACL (Ours)			75.15(12.98)	61.84(15.90)	**49.49(27.48)**	**21.64(11.93)**	-

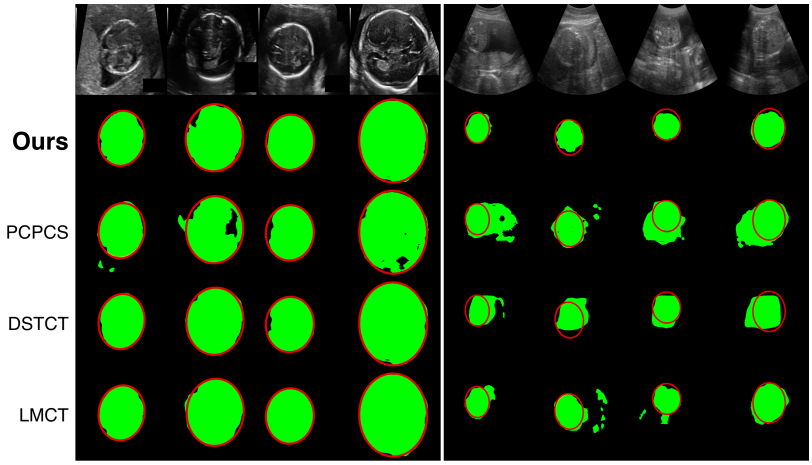

Fig. 2. Visual comparison of recent SSL methods using 10% labeled data. From left to right, they are four HC18 and four F-Abd samples. These examples are representative samples selected to reflect the overall quantitative trends rather than best-case results.

accuracy. Under the 10% labeled setting, DACL and LMCT achieved comparable performance ($p = 0.974$), indicating that both methods achieved a similar level of performance. Additionally, DACL outperforms the fully supervised (FS) UNeXt model across all metrics ($p < 0.05$), demonstrating its ability to exploit unlabeled data more effectively than FS training.

On the more challenging F-Abd dataset (Table 2), DACL achieves competitive segmentation accuracy among semi-supervised methods and yields state-of-the-art (SOTA) performance on boundary-sensitive metrics ($p < 0.05$) with 5% labeled data. In particular, DACL achieved statistically significant improvements over LMCT ($p < 0.05$) across all metrics. However, under the 10% labeled setting, DACL performs slightly below the best competing method in DSC and Jaccard. Nevertheless, a performance gap remains between semi-supervised and FS methods. This gap can be attributed to the extremely limited number of labeled samples and the use of low-cost portable ultrasound probes, which produce noisier images and are prone to both intra- and inter-operator variability.

Qualitative Results. Fig. 2 compares the segmentation results of DACL with GT masks and the three latest semi-supervised methods. On the fetal head dataset, DACL shows performance comparable to the competing methods, with only modest visual differences. In contrast, on the fetal abdomen dataset, DACL generates significantly more accurate and complete abdomen Region of Interest (ROI) delineations, with fewer missing regions and smoother boundaries, indicating a clearer advantage under the more challenging imaging conditions.

Ablation Study. Table 3 presents the results of our ablation study. The second row indicates that CPS provides effective supervision signals from unlabeled data by adding $\mathcal{L}_{\text{cps}}$ to the baseline. Incorporating the interpolation consistency loss

Table 3. Ablation studies for each loss function on the HC18 and F-Abd datasets with 10% of the data labeled. The best results are in **bold**.

Method	HC18			F-Abd		
	DSC ↑	HD95 ↓	ASD ↓	DSC ↑	HD95 ↓	ASD ↓
Baseline	92.36(9.28)	36.82(36.01)	13.22(14.07)	70.88(21.46)	75.02(39.70)	30.18(17.83)
+ $\mathcal{L}_{\text{cps}}$	93.64(7.17)	30.48(31.94)	10.79(11.83)	72.24(14.85)	72.26(36.50)	31.08(17.01)
+ $\mathcal{L}_{\text{cps}}$ + $\mathcal{L}_{\text{ict}}$	93.84(6.90)	30.35(31.24)	10.55(11.11)	74.05(17.70)	61.83(36.70)	26.18(17.55)
+ $\mathcal{L}_{\text{cps}}$ + $\mathcal{L}_{\text{dac}}$	94.95(6.31)	23.44(30.29)	8.26(10.85)	70.47(20.33)	52.34(38.01)	22.08(15.46)
+ $\mathcal{L}_{\text{cps}}$ + $\mathcal{L}_{\text{ict}}$ + $\mathcal{L}_{\text{dac}}$	**95.77(4.16)**	**19.42(25.29)**	**6.75(7.86)**	**75.15(12.98)**	**49.49(27.48)**	**21.64(11.93)**

$\mathcal{L}_{\text{ict}}$ further boosts accuracy on both datasets, suggesting that teacher-guided mixup regularization stabilizes training and improves generalization. Interestingly, replacing $\mathcal{L}_{\text{ict}}$ with the proposed dual-agreement loss $\mathcal{L}_{\text{dac}}$ yields the largest boundary-quality gains (HD95/ASD), particularly on F-Abd, although it can trade off some overlap accuracy (DSC). Finally, the last row shows that combining all three losses yields the best performance on both fetal head and abdomen segmentation, confirming the overall effectiveness of our method.

4 Conclusion

We introduced DACL, a lightweight semi-supervised framework that stabilizes cross pseudo supervision via the proposed DAC loss, jointly enforcing probabilistic agreement and uncertainty-aware confidence agreement between heterogeneous models. This strategy improves pseudo-label reliability under extreme annotation scarcity and challenging ultrasound imaging conditions while maintaining a lightweight architecture suitable for practical deployment. Experiments on fetal head and abdomen data show consistent improvements in both overlap and boundary accuracy compared with recent semi-supervised approaches, supporting more reliable downstream fetal biometry and quantitative assessment.

Acknowledgments. This work was funded by Taighde Éireann – Research Ireland through the Research Ireland Centre for Research Training in Machine Learning (18/CRT/6183).

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

References

1. Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M.: Swin-Unet: Unet-Like pure transformer for medical image segmentation. In: ECCV Workshops. pp. 205–218 (2022)
2. Chen, X., Yuan, Y., Zeng, G., Wang, J.: Semi-supervised semantic segmentation with cross pseudo supervision. In: CVPR. pp. 2613–2622 (June 2021)
3. Espinoza, J., Good, S., Russell, E., Lee, W.: Does the use of automated fetal biometry improve clinical work flow efficiency? Journal of Ultrasound in Medicine **32**(5), 847–850 (2013). https://doi.org/10.7863/jum.2013.32.5.847

4. van den Heuvel, T.L.A., de Bruijn, D., de Korte, C.L., van Ginneken, B.: Automated measurement of fetal head circumference using 2d ultrasound images. PLOS ONE **13**(8), 1–20 (08 2018)
5. Jiang, J., Wang, H., Bai, J., Long, S., Chen, S., Campello, V.M., Lekadir, K.: Intrapartum ultrasound image segmentation of pubic symphysis and fetal head using dual student-teacher framework with cnn-vit collaborative learning. In: MICCAI. pp. 448–458 (2024)
6. Joyce, J.M.: Kullback-Leibler divergence. In: International Encyclopedia of Statistical Science, pp. 720–722. Springer Berlin Heidelberg (2011)
7. Li, J., Gao, Z., Wang, C., Pu, B., Li, K.: A rule-guided interpretable lightweight framework for fetal standard ultrasound plane capture and biometric measurement. Neurocomputing **621**, 129290 (2025). https://doi.org/10.1016/j.neucom.2024.129290
8. Luo, X., Hu, M., Song, T., Wang, G., Zhang, S.: Semi-supervised medical image segmentation via cross teaching between CNN and transformer. In: MIDL. pp. 820–833 (2022)
9. Lyu, C., Han, K., Liu, L., Chen, J., Ma, L., Pang, Z., Liu, Z.: Bidirectional prototype-guided consistency constraint for semi-supervised fetal ultrasound image segmentation. IEEE Journal of Biomedical and Health Informatics pp. 1–13 (2025)
10. Ma, C., Wang, Z.: Semi-Mamba-UNet: Pixel-level contrastive and cross-supervised visual mamba-based UNet for semi-supervised medical image segmentation. Knowledge-Based Systems **300**, 112203 (2024)
11. Mongelli, M., Ek, S., Tambyrajia, R.: Screening for fetal growth restriction: A mathematical model of the effect of time interval and ultrasound error. Obstetrics & Gynecology **92**(6) (1998)
12. Papageorghiou, A.T., Ohuma, E.O., Altman, D.G., Todros, T., Ismail, L.C., L., A., Jaffer, Y.A., Bertino, E., Gravett, M.G., Purwar, M., Noble, J.A., Pang, R., Victora, C.G., Barros, F.C., Carvalho, M., Salomon, L.J., Bhutta, Z.A., Kennedy, S.H., Villar, J.: International standards for fetal growth based on serial ultrasound measurements: the fetal growth longitudinal study of the intergrowth-21st project. The Lancet **384**(9946), 869–879 (2014)
13. Qiao, S., Shen, W., Zhang, Z., Wang, B., Yuille, A.: Deep co-training for semi-supervised image recognition. In: ECCV. pp. 135–152 (September 2018)
14. Ramirez Zegarra, R., Ghi, T.: Use of artificial intelligence and deep learning in fetal ultrasound imaging. Ultrasound in Obstetrics & Gynecology **62**(2), 185–194 (2023). https://doi.org/10.1002/uog.26130
15. Salomon, L.J., Alfirevic, Z., Berghella, V., Bilardo, C., Hernandez-Andrade, E., Johnsen, S.L., Kalache, K., Leung, K.Y., Malinger, G., Munoz, H., Prefumo, F., Toi, A., Lee, W., on behalf of the ISUOG Clinical Standards Committee: Practice guidelines for performance of the routine mid-trimester fetal ultrasound scan. Ultrasound in Obstetrics & Gynecology **37**(1), 116–126 (2011). https://doi.org/10.1002/uog.8831
16. Sappia, M.S., de Korte, C.L., van Ginneken, B., Ninalga, D., Kondo, S., Kasai, S., Hirasawa, K., Akumu, T., Martín-Isla, C., Lekadir, K., Campello, V.M., Fabila, J., Beverdam, A., van Dillen, J., Neff, C., Murphy, K.: Acouslic-ai challenge report: Fetal abdominal circumference measurement on blind-sweep ultrasound data from low-income countries. Medical Image Analysis **105**, 103640 (2025). https://doi.org/10.1016/j.media.2025.103640

17. Sarris, I., Ioannou, C., Chamberlain, P., Ohuma, E., Roseman, F., Hoch, L., Altman, D.G., Papageorghiou, A.T., International Fetal and Newborn Growth Consortium for the 21st Century (INTERGROWTH-21st): Intra- and interobserver variability in fetal ultrasound measurements. Ultrasound Obstet. Gynecol. **39**(3), 266–273 (2012)
18. Shannon, C.E.: A mathematical theory of communication. The Bell System Technical Journal **27**(3), 379–423 (1948). https://doi.org/10.1002/j.1538-7305.1948.tb01338.x
19. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. NeurIPS **30** (2017)
20. Valanarasu, J.M.J., Patel, V.M.: UNeXt: MLP-Based rapid medical image segmentation network. In: MICCAI. pp. 23–33 (2022)
21. Verma, V., Kawaguchi, K., Lamb, A., Kannala, J., Solin, A., Bengio, Y., Lopez-Paz, D.: Interpolation consistency training for semi-supervised learning. Neural Networks **145**, 90–106 (2022)
22. Wang, F., Curran, K.M., Silvestre, G.: Semi-supervised cervical segmentation on ultrasound by a dual framework for neural networks. In: IEEE 22nd International Symposium on Biomedical Imaging (2025)