Calibration-Aware Semi-Supervised Fetal Head Segmentation with Boundary-Positive Contrast

Ufaq Khan[*1] UFAQ.KHAN@MBZUAI.AC.AE
Umair Nawaz[1] UMAIR.NAWAZ@MBZUAI.AC.AE
Tajamul Ashraf[1] TAJAMUL.ASHRAF@MBZUAI.AC.AE
Tausifa Jan Saleem[1] TAUSIFA.SALEEM@MBZUAI.AC.AE
Massimo Caputo[2] M.CAPUTO@BRISTOL.AC.UK
Srinivas Ananth Narayan[3] SRINIVAS.NARAYAN@UHBW.NHS.UK
Muhammad Bilal[4] MUHAMMAD.BILAL@BCU.AC.UK
Junaid Qadir[5] JQADIR@QU.EDU.QA
Muhammad Haris[1] MUHAMMAD.HARIS@MBZUAI.AC.AE

[1] *Mohamed Bin Zayed University of Artificial Intelligence, UAE*
[2] *University of Bristol, UK*
[3] *University Hospital Bristol and Weston, UK*
[4] *Birmingham City University, UK*
[5] *Qatar University, Qatar*

Editors: Under Review for MIDL 2026

Abstract

Accurate fetal head segmentation in ultrasound is hard to scale as labels are scarce and most errors occur at the head–background interface under speckle, shadowing, and low contrast. We present *UltraSemiNet*, a teacher–student framework that makes cross–pseudo supervision (CPS) *selective* via temperature calibration and a dual gate requiring high confidence and test-time augmentation (TTA) stability. We also introduce two boundary-focused modules that complement CPS: **SAT**, a boundary-positive spatial contrast that learns *through* ambiguous edges using an entropy belt and a soft-IoU agreement test; and **PCM**, a prototype-guided curriculum that maintains uncertainty-weighted head/background prototypes and targets feature–prototype discrepancies. Across two datasets (FBUI and HC18), UltraSemiNet improves overlap and boundary metrics over a calibrated CPS baseline (e.g., Dice 0.927→0.971; HD95 7.9→6.8 px), with similar cross-dataset trends. Crucially, the calibrated gate reduces miscalibration of the *accepted* pseudo-labels: both expected calibration error (ECE) and Brier score decrease overall, with the largest gains within the 0–2 px boundary band, alongside improvements in pseudo-label accuracy. Ablations show CPS calibration, SAT, and PCM are complementary and concentrate improvements on boundary-sensitive metrics. In a blinded study, UltraSemiNet achieved greater agreement with an adjudicated rater consensus than either of two senior fetal medicine experts and met their quality checklist criteria more consistently, indicating the potential to reduce manual refinements. The code will be released upon acceptance.

Keywords: Fetal Segmentation, Self-Supervised Learning, Semi-Supervised Learning

[*] Corresponding Author

1. Introduction

The accurate delineation of the fetal head on ultrasound is a foundational step for biometric assessment (HC/BPD) and downstream clinical decisions (van den Heuvel et al., 2018a). Yet, obtaining pixel-accurate annotations is expensive and time-consuming. In addition, the boundary between the skull and the background is precisely where ultrasound is most challenging due to speckle, acoustic shadowing, and low local contrast (Noble and Boukerroui, 2006). These factors produce two practical obstacles: (i) label scarcity at scale, and (ii) boundary ambiguity even when labels exist, compounded by variation in annotation style (e.g., ellipse-like contours vs. fine-grained masks) (Zhang et al., 2020). As a result, fully supervised models often overfit limited labels or yield contours that are visually plausible yet misaligned, inflating boundary-sensitive errors (ASD/HD95) despite reasonable Dice (Shen et al., 2023).

Semi-supervised segmentation is attractive because it can leverage large unlabeled collections via self-training and consistency (Tarvainen and Valpola, 2017a; Chen et al., 2021b). However, in ultrasound, the pixels most likely to be mislabeled are those near the head boundary. If we trust raw softmax confidence, which is often miscalibrated, pseudo-labels can amplify edge errors. Similarly, enforcing consistency without checking the stability of the prediction can also reinforce artifacts. (Guo et al., 2017; Wang et al., 2019). A subtler issue is that many representation-learning strategies avoid uncertain regions when constructing contrastive pairs (e.g., by sampling confident pixels or class-memory exemplars), which sharpens features in easy interiors while leaving boundary features under-constrained (Wang and Kong, 2022; Wang et al., 2021; Alonso et al., 2021). Together, miscalibrated pseudo-labels and boundary avoidance create a gap between overlap metrics and the boundary fidelity required in practice.

We address semi-supervised *binary* segmentation of the fetal head in ultrasound with minimal pixel-level annotation. The technical challenge is to exploit unlabeled data without reinforcing boundary errors and to learn feature representations so that the two classes (head and background) remain well separated at the interface. Concretely, we seek a training recipe that (1) selects reliable pseudo-labels under explicit checks for confidence *and* augmentation stability; (2) learns discriminative features through ambiguous boundary neighborhoods only when local evidence agrees; and (3) regularizes the global feature space so that the two classes form compact, well-separated clusters, despite class imbalance and annotation style variability.

We propose *UltraSemiNet*, a teacher–student framework that combines cross–pseudo supervision (CPS) with two uncertainty-aware representation modules tailored to ultrasound boundaries. In practice, UltraSemiNet fits into the clinical workflow in two complementary ways. *(i)*, as a pre-measurement step, it produces a clean head mask that stabilizes ellipse-based HC/BPD tools and reduces manual contour editing, shortening scan time and standardizing measurements across operators. *(ii)*, as a real-time quality check, its calibrated confidence and augmentation-stability scores flag frames with unreliable boundaries for immediate reacquisition before measurements are finalized. Because the approach is semi-supervised, sites can leverage large unlabeled archives and a small labeled seed set, thereby easing deployment where expert annotation fundings are limited.

In terms of methodology, we first *calibrate* teacher probabilities via temperature scaling on a small labeled subset and gate pseudo-labels with a dual criterion, i.e., high binary confidence and stability under flip/rotation test-time augmentation (TTA). This converts CPS from a blind self-training signal into a selective supervision source targeted to trustworthy pixels. Second, we introduce **SAT**, a boundary-positive **S**p**AT**ial contrast that uses calibrated probabilities to detect an entropy-based "boundary belt" and admits positive pairs across edges only when the local probability fields agree (soft-IoU gate), with anchor weights that focus learning on clinically relevant ambiguity while avoiding extreme noise. Third, we introduce **PCM**, a prototype-guided curriculum miner that maintains *uncertainty-weighted* teacher prototypes for head and background and prioritizes pixels whose student features are far from their predicted class prototype, producing a simple global structure prior without brittle hard-patch heuristics.

Contributions. *(i)* We formulate a *calibrated* CPS scheme for binary ultrasound segmentation that accepts pseudo-labels only when both confident and TTA-stable, reducing confirmation bias near boundaries. *(ii)* We propose **SAT**, an uncertainty-aware, boundary-positive spatial contrast that admits cross-edge positives via a soft-IoU agreement test on probability crops. *(iii)* We propose **PCM**, a lightweight prototype-guided curriculum that replaces heuristic hard-patch mining with uncertainty-weighted teacher prototypes and a discrepancy-based selection of semantically hard pixels, improving global feature separation under class imbalance. *(iv)* We conduct a blinded reader study with two senior fetal medicine experts, demonstrating higher agreement with UltraSemiNet outputs.

2. Related work

Calibration-aware pseudo-labeling. The utility of pseudo-labels depends on how well predicted probabilities reflect correctness. For this purpose, simple temperature scaling improves this alignment and yields cleaner supervision when applied before thresholding or re-labeling (Choi et al., 2024; Joy et al., 2023). Beyond a single global cutoff, adaptive acceptance schedules either class-balanced self-training or curriculum thresholds, reduce bias toward majority regions, and progressively admit harder pixels. Stability checks under flip/rotation test-time augmentation (TTA) further filter unreliable pseudo-labels (Tan et al., 2024) that would otherwise reinforce boundary artifacts. In medical SSL, uncertainty-aware teachers reduce the weight of the ambiguous regions or gate them to curb error propagation near the interfaces (Alizadehsani et al., 2024; Vazhentsev et al., 2025; Chen et al., 2023).

UltraSemiNet adopts this calibration-aware view end-to-end, where we (i) temperature-scale the teacher on a small labeled split each epoch, (ii) enforce a dual gate that requires both high *calibrated* confidence and TTA-based stability before a pixel supervises CPS, and (iii) reuse the same calibrated probabilities to drive SAT's boundary-positive pairing and PCM's uncertainty-weighted prototype updates, ensuring selection and representation shaping are governed by a shared reliability criterion.

Classical Techniques in Fetal Ultrasound Segmentation. The segmentation of fetal structures in ultrasound images initially relied heavily on classical image processing methods. The techniques such as thresholding (Liu et al., 2019), region growth (Yuheng and Hao, 2017), and Active Contour Models (ACM) (Kass et al., 1988) formed the basis of early seg-

mentation efforts. While these methods provided initial frameworks for segmentation, they often required substantial manual intervention and struggled with the intrinsic challenges of ultrasound imaging, such as speckle noise and low contrast, limiting their effectiveness and reliability in clinical applications (Smith and Doe, 2010). The advent of deep-learning architectures has substantially advanced fetal ultrasound segmentation.

3. Methodology

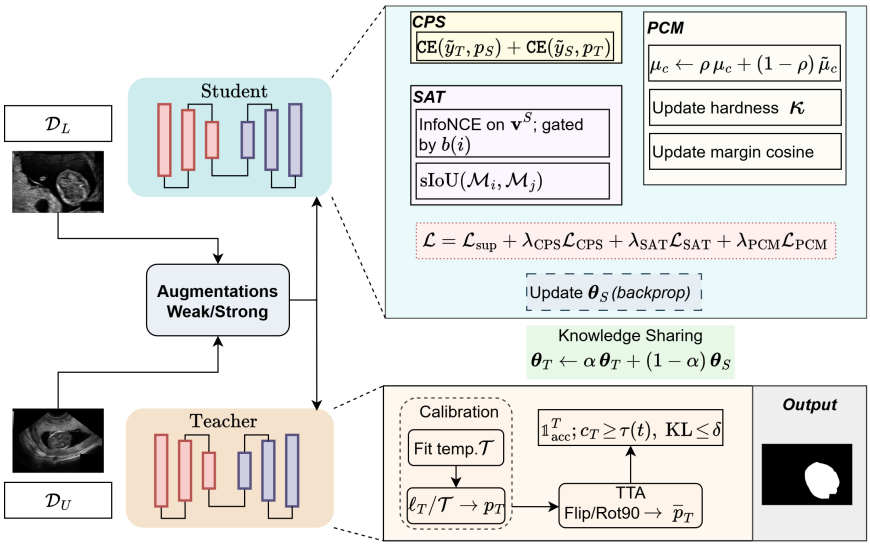


Figure 1: An overview of UltraSemiNet: weak/strong views feed teacher f_T (temperature-calibrated, TTA-averaged) and student f_S. The student minimizes the overall loss $\mathcal{L}$ using CPS, boundary-positive SAT, and PCM modules, while the teacher is updated via EMA.

3.1. Preliminaries and Notation

Let $\mathcal{D}_L = \{(x, y)\}$ be labeled images with binary masks $y \in \{0,1\}^{h\times w}$ for the fetal head (1=head, 0=background), and $\mathcal{D}_U = \{x\}$ unlabeled images, as depicted in Figure 1. A shared encoder–decoder backbone maps x to a feature map $H \in \mathbb{R}^{D\times h\times w}$. We maintain a *student* $f_S(\cdot;\theta_S)$ and an EMA *teacher* $f_T(\cdot;\theta_T)$, which can be updated as

$$\theta_T \leftarrow \alpha\,\theta_T + (1-\alpha)\,\theta_S, \qquad \alpha \in [0.99, 0.999]. \tag{1}$$

At pixel i, both networks output the *head probability* $p_S(i), p_T(i) \in [0,1]$ (sigmoid of the head logit). The teacher pseudo-label is $\hat{y}_T(i) = \mathbb{1}[\,p_T(i) \geq 0.5\,]$. We quantify uncertainty by the *binary entropy* which is:

$$u(i) = -\,p_T(i)\log p_T(i) \;-\; \big(1-p_T(i)\big)\log\big(1-p_T(i)\big), \tag{2}$$

which rises at $p_T(i) = 0.5$ and is small inside confident interior/background. For representation learning we obtain $\mathcal{L}_2$-normalized embeddings $v_i = \text{Norm}(\phi(H_i)) \in \mathbb{R}^d$ via a two-layer

projection head ϕ (default d=128). SAT uses a disk $\mathcal{N}_r(i) = \{j : \|j - i\| \le r\}$ for positives and an annulus $\mathcal{A}_{[d_{\min}, d_{\max}]}(i) = \{j : d_{\min} \le \|j - i\| \le d_{\max}\}$ for negatives. For local agreement tests, we extract a *single-channel* teacher probability crop $\mathcal{M}_i \in [0,1]^{s \times s}$ centered at i (window size s).

3.2. Calibrated Cross–Pseudo Supervision (CPS)

CPS leverages unlabeled data by letting teacher and student supervise each other. However, using raw probabilities directly can therefore risk reinforcing errors. We therefore calibrate teacher probabilities using calibration techniques (Kim et al., 2025; Forest and Fink, 2024) so that confidence aligns better with empirical correctness, and we restrict CPS to pixels that are both confident and stable under simple test-time augmentations.

Probability calibration. At the start of each epoch, we fit a temperature $\mathcal{T} > 0$ on a small labeled subset by minimizing negative log-likelihood. If $\ell_T(i)$ denotes the teacher head logit, the *calibrated* probability is $p_T(i) \leftarrow \sigma\big(\ell_T(i)/\mathcal{T}\big)$, where, $\sigma(z) = \frac{1}{1+e^{-z}}$.

Calibration is re-estimated periodically (e.g., every 3 epochs) to track the evolving teacher. All downstream decisions in the epoch use the calibrated p_T. For binary segmentation, a natural confidence is $c_T(i) = \max\{\, p_T(i),\, 1 - p_T(i) \,\}$. A pixel is eligible for CPS if it is confident and its probability is stable under flips/rotations. Let $\overline{p}_T^{\mathrm{TTA}}(i)$ be the average teacher probability at i over horizontal/vertical flips and 90° rotations (mapped back to the original coordinates). For scalars $p, q \in (0,1)$, $\mathrm{Bern}(p)$ denotes the Bernoulli distribution with success probability p. Here, we have an acceptance indicator as:

$$\mathbb{1}_{\mathrm{acc}}^{T}(i) = \mathbb{1}[c_T(i) \ge \tau(t)] \;\times\; \mathbb{1}\Big[\mathrm{KL}\Big(\mathrm{Bern}\big(p_T(i)\big) \,\Big\|\, \mathrm{Bern}\big(\overline{p}_T^{\mathrm{TTA}}(i)\big)\Big) \le \delta\Big] \tag{3}$$

where $\tau(t)$ anneals linearly from 0.95 to 0.80 over training and δ is a small KL cutoff (default 0.15). Now, the accepted teacher labels are:

$$\tilde{y}_T(i) = \begin{cases} \hat{y}_T(i), & \text{if } \mathbb{1}_{\mathrm{acc}}^{T}(i) = 1, \\ \text{ignore}, & \text{otherwise.} \end{cases} \tag{4}$$

Symmetrically, we obtain $\tilde{y}_S(i)$ with $p_S(i)$ and $c_S(i) = \max\{p_S(i), 1 - p_S(i)\}$, defining $\mathbb{1}_{\mathrm{acc}}^{S}(i)$ via (3). Here, $\hat{y}_T(i)$ denotes the *raw teacher pseudo-label* at pixel i, obtained from the calibrated probability $p_T(i)$ whereas $\tilde{y}_T(i)$ represent the *accepted* pseudo-label. We then supervise the student using teacher-accepted pixels $\Omega_T = \{i : \mathbb{1}_{\mathrm{acc}}^{T}(i) = 1\}$, and supervise the teacher (EMA) using student-accepted pixels $\Omega_S = \{i : \mathbb{1}_{\mathrm{acc}}^{S}(i) = 1\}$:

$$\mathcal{L}_{\mathrm{CPS}} = \frac{1}{|\Omega_T|} \sum_{i \in \Omega_T} \mathrm{CE}\big(\tilde{y}_T(i),\, p_S(i)\big) + \frac{1}{|\Omega_S|} \sum_{i \in \Omega_S} \mathrm{CE}\big(\tilde{y}_S(i),\, p_T(i)\big). \tag{5}$$

By restricting CPS to confident and augmentation-stable pixels, we reduce confirmation bias where the model is most error-prone (near the head boundary) and provides consistent, calibrated probabilities $p_T(i)$ and uncertainty $u(i)$ as depicted in Figure A(b) of supplementary material.

3.3. SAT: Uncertainty-Aware Boundary-Positive Spatial Contrast

The SAT module is designed around a simple observation: in binary fetal head segmentation, most of the ambiguity lives exactly where the head meets the background. Rather than discarding these pixels as "noisy," SAT turns them into useful supervision, but only when the nearby evidence supports doing so. Concretely, SAT uses the teacher's *calibrated* head probabilities $p_T(i) \in [0,1]$ (Sec. 3.2) to identify potentially ambiguous locations and then shapes the student's representations to be both consistent across those locations and well-separated from the background. We begin by quantifying local uncertainty using the binary entropy: $u(i) = -p_T(i)\log p_T(i) - \big(1-p_T(i)\big)\log\big(1-p_T(i)\big)$, and normalize it by its maximum, $\tilde{u}(i) = u(i)/\log 2 \in [0,1]$. Pixels with intermediate $\tilde{u}$ are likely to lie on (or near) the head boundary. We therefore form a *boundary belt* by marking pixels whose normalized entropy falls in $[\epsilon_1, \epsilon_2]$, so $b(i) = \mathbb{1}\big[\epsilon_1 \le \tilde{u}(i) \le \epsilon_2\big]$,where $\epsilon_1 = 0.40,\ \epsilon_2 = 0.95$.

Now, ambiguous pixels are not automatically trusted. Before we let them act as "positives" in a contrastive pair, we ask whether the *local probability fields* around two candidate pixels agree. For that we extract a single-channel probability crop $\mathcal{M}_i \in [0,1]^{s\times s}$ centered at i and compute a *soft* intersection-over-union with a neighbor j,

$$\mathrm{sIoU}(\mathcal{M}_i, \mathcal{M}_j) = \frac{\sum_\omega \min\{\mathcal{M}_i(\omega), \mathcal{M}_j(\omega)\}}{\sum_\omega \max\{\mathcal{M}_i(\omega), \mathcal{M}_j(\omega)\}}, \tag{6}$$

which measures how similar the surrounding head probabilities are within an $s \times s$ window.

Positive/negative sampling with CPS consistency. With these ingredients, SAT constructs contrastive pairs as follows. For an anchor pixel i with teacher pseudo-label $\hat{y}_T(i) = \mathbb{1}[p_T(i) \ge 0.5]$, we look for a *positive* $p(i)$ in a small disk $\mathcal{N}_r(i)$ that shares the same pseudo-label. If the anchor is away from the boundary ($b(i) = 0$), we simply require that both pixels are eligible under the CPS acceptance rule (Eq. (3)), ensuring they are confident and stable. If the anchor lies in the boundary belt ($b(i) = 1$), we only admit a positive with probability $q_b(t)$ *and* when the local fields agree, i.e., $\mathrm{sIoU}(\mathcal{M}_i, \mathcal{M}_{p(i)}) > \tau_{\mathrm{bIoU}}$. Moreover, at least one of the two pixels must be CPS-accepted to guard against spurious matches. Now, the negatives are drawn from an annulus $\mathcal{A}_{[d_{\min}, d_{\max}]}(i)$ and must carry the opposite pseudo-label. Here, to avoid background overwhelming the loss in this binary setting, we sample negatives so that head and background contribute in a balanced way whenever possible.

To focus learning where it matters, each anchor is given a weight that increases with local ambiguity but does not reward extreme uncertainty. Mathematically, it can be shown as: $w(i) = \Big(\gamma_1\big(1-\tilde{u}(i)\big) + \gamma_2\,\overline{\tilde{u}}_{\mathcal{N}_r(i)}\Big)\,\big(1+\lambda_b\,b(i)\big)$, where $\overline{\tilde{u}}_{\mathcal{N}_r(i)}$ averages $\tilde{u}$ in the positive radius, γ_1=0.7, γ_2=0.3, and a small boost λ_b=0.5 nudges attention toward the belt. Using $\mathcal{L}_2$-normalized student embeddings, SAT minimizes a cosine InfoNCE loss as:

$$\mathcal{L}_{\mathrm{SAT}} = \textstyle\sum_i w(i)\left[-\log \frac{\exp\big(\langle v_i^S, v_{p(i)}^S\rangle/\beta\big)}{\exp\big(\langle v_i^S, v_{p(i)}^S\rangle/\beta\big) + \sum_{n\in\mathcal{N}(i)}\exp\big(\langle v_i^S, v_n^S\rangle/\beta\big)}\right],\quad \beta = 0.07 \tag{7}$$

In effect, SAT asks the representation to be continuous across genuinely consistent boundary neighborhoods and separate them decisively from the background nearby. Now, as the admission and weighting depend on the *same calibrated probabilities and acceptance*

rule used in CPS, the contrastive signal remains logically consistent with supervision selection i.e, ambiguous regions shape the features only when the evidence supports it, leading to sharper, more reliable contours (improved HD95/ASD) without sacrificing Dice. In practice, we use r=5, $[d_{\min}, d_{\max}]$=[7, 11], s=15, K_-=64 negatives per anchor, ramp $q_b(t)$ from 0 to 0.4 during the first half of training, and choose $\tau_{\text{bIoU}} \in \{0.5, 0.6, 0.7\}$.

3.4. PCM: Prototype-Guided Curriculum Miner

Where SAT shapes features *locally* around the boundary, PCM provides a lightweight *global* structure by maintaining running prototypes for "head" and "background" and asks the student's features to organize themselves around these references. This is done in a way that is robust to label noise, i.e., prototypes are updated only from pixels that the teacher deems reliable under the CPS acceptance rule, and those contributions are further down-weighted when uncertainty is high.

Specifically, for each class $c \in \{0, 1\}$ we maintain an EMA prototype $\mu_c \in \mathbb{R}^d$. At a training step, teacher embeddings v_i^T with pseudo-label $\hat{y}_T(i) = c$ are pooled into a temporary estimate $\tilde{\mu}_c$ using weights: $\omega(i) = \mathbb{1}_{\text{acc}}^T(i)\,\big(1 - \tilde{u}(i)\big)$, so that only confident, augmentation-stable pixels contribute and highly uncertain cases have little influence. The class prototype is then updated by exponential averaging,

$$\tilde{\mu}_c = \frac{\sum_{i:\,\hat{y}_T(i)=c}\ \omega(i)\, v_i^T}{\sum_{i:\,\hat{y}_T(i)=c}\ \omega(i)}, \quad \mu_c \leftarrow \rho\,\mu_c + (1-\rho)\,\tilde{\mu}_c, \quad \rho = 0.99. \tag{8}$$

We warm-start μ_c from the labeled set. If a batch contains no reliable pixels for a class, the EMA simply carries the previous prototype forward. Not all pixels are equally informative for organizing the feature space. To determine where PCM should act most strongly, we define a *semantic hardness* score at each pixel by measuring how far the student's feature is from the prototype of its *own* predicted class. Let $p_S(i)$ be the student's head probability and set $\pi_1(i) = p_S(i)$, $\pi_0(i) = 1 - p_S(i)$. We compute $\kappa(i) = \max\Big\{\pi_1(i)\big(1-\cos\langle v_i^S, \mu_1\rangle\big), \pi_0(i)\big(1-\cos\langle v_i^S, \mu_0\rangle\big)\Big\}$, which is large when a pixel is predicted as head (or background) but its feature is inconsistent with the corresponding prototype. A curriculum then selects the top $\gamma(t)$ fraction of pixels by $\kappa(i)$ (with γ annealed from 0.3 to 0.7 as training stabilizes) and forms the active set Ω_γ. Because background typically occupies more area than head, we subsample background within Ω_γ so that both classes contribute comparably.

Finally, PCM pulls features toward the prototype of their predicted class while pushing them away from the other class, using a marginized cosine loss:

$$\mathcal{L}_{\text{PCM}} = \frac{1}{|\Omega_\gamma|}\sum_{i\in\Omega_\gamma}\Big[\pi_1(i)\,\big(m - \cos\langle v_i^S, \mu_1\rangle + \cos\langle v_i^S, \mu_0\rangle\big)_+ + \ \pi_0(i)\,\big(m - \cos\langle v_i^S, \mu_0\rangle + \cos\langle v_i^S, \mu_1\rangle\big)_+\Big], m = 0.2. \tag{9}$$

with $(\cdot)_+ = \max(0, \cdot)$, the effect is to produce compact, well-separated head and background clusters in the embedding space, guided by prototypes that are themselves estimated from reliable pixels. Because prototype updates use the *same* acceptance indicator and entropy as CPS, PCM's global organization is aligned with the supervision that trains the classifier head and with the local boundary shaping performed by SAT. In practice, we

compute $\kappa(i)$ at 1/2 and 1/4 resolution for efficiency and upsample the resulting mask. Here, if prototypes begin to drift or collapse, increasing the margin to m=0.3 or slowing the curriculum ramp mitigates the issue.

4. Experimental Setup

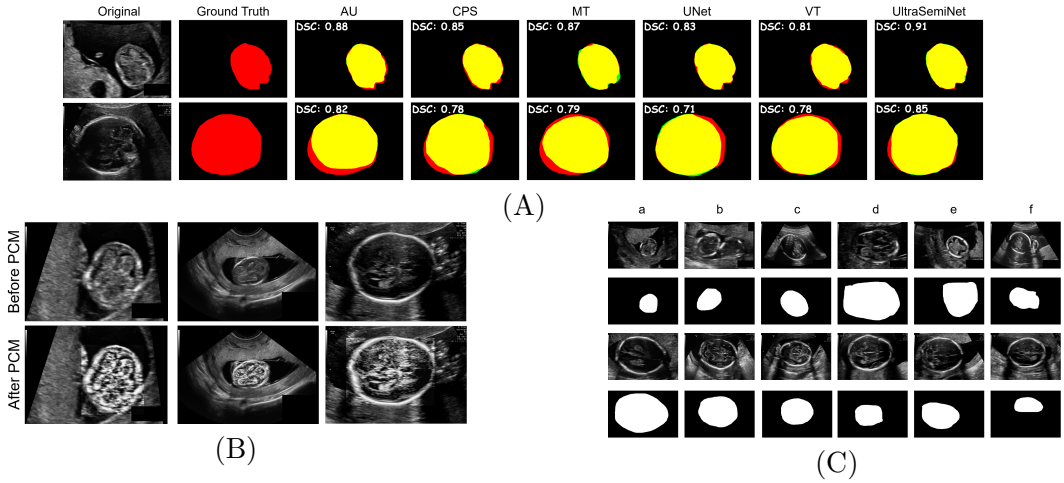


Figure 2: (A) Original ultrasound images, ground truth segmentations, and model outputs with Dice scores; (B) Before and after PCM; (C) Best (a-c) and worst (d-f) UltraSemiNet predictions.

Datasets. We evaluate on two fetal head ultrasound collections. (1) *FBUI* (Alzubaidi et al., 2023): 3,832 images spanning 18–40 weeks GA, split **by subject** into 60%/20%/20% train/val/test. (2) *HC18 subset* (van den Heuvel et al., 2018b): It contains 999 fetal brain images with binary head annotations. Unless stated, FBUI provides labeled supervision, and HC18 contributes additional unlabeled images for semi-supervised training and cross-dataset evaluation. Further details are provided in the supplementary material.

Table 1: FBUI (in-distribution) results under 5-fold CV.

Method	DSC ↑	ASD (px) ↓	HD95 (px) ↓
U–Net (Ronneberger et al., 2015)	0.913 ± 0.015	2.41 ± 1.25	9.8 ± 4.2
Attention U–Net (Oktay et al., 2018)	0.921 ± 0.013	2.12 ± 1.19	8.7 ± 3.9
ViT (Dosovitskiy et al., 2020)	0.906 ± 0.018	2.78 ± 1.43	11.3 ± 5.1
2D nnU–Net (Isensee et al., 2021)	0.930 ± 0.011	1.94 ± 1.07	8.2 ± 3.6
Mean Teacher (MT) (Tarvainen and Valpola, 2017b)	0.918 ± 0.014	2.23 ± 1.18	9.3 ± 4.0
CPS-only (Chen et al., 2021a)	0.927 ± 0.012	1.86 ± 1.09	7.9 ± 3.8
MedSAM (Ma et al., 2024)	0.821 ± 0.061	6.85 ± 3.72	26.4 ± 11.7
SAMUS (Lin et al., 2024)	0.846 ± 0.049	5.72 ± 3.11	22.7 ± 10.5
UltraSemiNet (ours)	**0.971 ± 0.010**	**1.07 ± 0.92**	**6.8 ± 3.2**

Table 2: Cross-dataset results: train on FBUI, test on HC18 (no retuning).

Method	DSC ↑	ASD (px) ↓	HD95 (px) ↓
U–Net (Ronneberger et al., 2015)	0.887 ± 0.028	3.21 ± 1.92	14.2 ± 6.3
Attention U–Net (Oktay et al., 2018)	0.896 ± 0.026	2.98 ± 1.78	13.1 ± 5.8
2D nnU–Net (Isensee et al., 2021)	0.905 ± 0.022	2.75 ± 1.63	12.4 ± 5.5
Mean Teacher (MT) (Tarvainen and Valpola, 2017b)	0.898 ± 0.025	2.91 ± 1.71	12.8 ± 5.7
CPS-only (Chen et al., 2021a)	0.907 ± 0.022	2.63 ± 1.56	11.7 ± 5.1
MedSAM (Ma et al., 2024)	0.803 ± 0.072	7.41 ± 4.10	28.9 ± 12.6
SAMUS (Lin et al., 2024)	0.828 ± 0.064	6.32 ± 3.65	24.7 ± 11.2
UltraSemiNet (ours)	**0.925 ± 0.019**	**1.27 ± 1.44**	**10.3 ± 4.7**

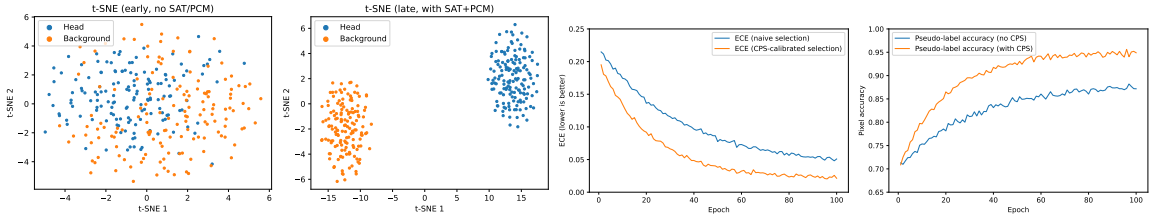
Baselines. We compare *UltraSemiNet* against: U-Net (Ronneberger et al., 2015), Attention U-Net (Oktay et al., 2018), Vision Transformer (ViT) (Dosovitskiy et al., 2020), Mean Teacher (MT) (Tarvainen and Valpola, 2017b), and CPS-only (Chen et al., 2021a). To contextualize results with widely used references, we also include 2D nnU-Net (Isensee et al., 2021) (default config) and promptable foundation models MedSAM (Ma et al., 2024) and SAMUS (Lin et al., 2024).

Implementation details. Images are resized to 224×224 and intensity-normalized. The teacher uses *weak* views (flips, 90° rotations), whereas the student uses weak+strong views (brightness/contrast/gamma jitter, mild elastic, light speckle). Models are trained on 1×A100 with total batch size 32 for 10,000 iterations (five-fold CV unless noted). We use AdamW (lr 1×10^{-4}, weight decay 1×10^{-4}), cosine decay with 5% warmup, gradient clipping at 1.0. The student is optimized with $\mathcal{L}_{\text{sup}}+\lambda_{\text{CPS}}\mathcal{L}_{\text{CPS}}+\lambda_{\text{SAT}}\mathcal{L}_{\text{SAT}}+\lambda_{\text{PCM}}\mathcal{L}_{\text{PCM}}$ (defaults: λ_{CPS}=1.0, λ_{SAT}=0.3, λ_{PCM}=0.2). The teacher is updated by EMA with α=0.996. SAT uses one positive and $K_- = 64$ negatives per anchor (radius r=5, annulus $[7, 11]$, gate τ_{bIoU}=0.6), whereas PCM maintains head/background prototypes with EMA ρ=0.99 and a curriculum keep ratio $\gamma(t) : 0.3 \to 0.7$. Early stopping was also applied on validation DSC with patience of 30.

5. Results

Main Results on FBUI (5-fold CV). Table 1 reports results on FBUI under five-fold, patient-level cross-validation with 1,500 labeled images and the remaining training images treated as unlabeled. UltraSemiNet achieves the best Dice while yielding the largest gains on boundary-sensitive metrics (ASD, HD95). Improvements over CPS-only are modest in Dice (as expected for a relatively clear binary target) but consistent and more pronounced for HD95, indicating sharper and more reliable contours.

Cross-Dataset Generalization (Train on FBUI, Test on HC18 subset). We next assess robustness under distribution shift by training on FBUI and evaluating on the HC18 subset without retuning thresholds. All methods degrade, but UltraSemiNet maintains the strongest performance and the largest boundary gains (Table 2).



(a) t-SNE (early, no SAT/PCM) (b) t-SNE (late, with SAT+PCM) (c) ECE over epochs (lower is better) (d) Pseudo-label accuracy w/ and w/o CPS

Figure 3: **UltraSemiNet dynamics.** a–b) SAT+PCM sharpen feature separation in t−SNE (compact, well−separated classes). (c) Calibrated CPS (with TTA stability) yields lower ECE than naive selection. (d) CPS achieves higher pseudo−label accuracy and faster convergence.

Qualitative Analysis. Figure 2A shows qualitative results along with a DSC score for each method where UltraSemiNet reduces boundary leakage. Figure 2B shows that PCM further refines contrast in high-entropy areas, whereas Figure 2C shows the best and worst predictions for our model. Figure 3 summarizes why UltraSemiNet improves downstream accuracy. Here, SAT+PCM turns entangled early features into compact, separated clusters (a–b). CPS reduces ECE across training (c) and produces higher pseudo-label accuracy

throughout (d). We also show some further analysis in Figure 4 of the appendix by presenting **dynamics** where CPS acceptance rises while TTA KL falls (a), boundary pixels steadily catch up to interiors as SAT focuses learning at ambiguous edges (b), and near-diagonal reliability at late epochs (c).

Clinical and Practical Implications Table 3a compares UltraSemiNet predictions with two medical experts (ME1, ME2) on total 60 samples from the FBUI and HC18 datasets, revealing that UltraSemiNet consistently outperforms medical experts. In terms of role and experience, ME1 is a board-certified fetal medicine specialist with over 10 years of experience in obstetric ultrasound, while ME2 is a senior sonographer with over 13 years of dedicated practice in high-risk pregnancy imaging. Each expert independently delineated the fetal head boundary on the set of 30 images from each dataset. These annotations were then compared with the ground truth masks provided by the source dataset as well as the predicted masks by our UltraSemiNet model.

Ablations and Calibration We isolate the contributions of calibration/gating, SAT, and PCM as depicted in Table 3b. All ablations use the same backbone, schedules, and labeled fraction. We quantify probability reliability using Expected Calibration Error (ECE; 15 bins) and Brier score on the student's probability map. Table 3c shows that calibrating CPS and using SAT/PCM reduces ECE and Brier versus naive CPS.

Table 3: Comparative performance, ablations, and probability calibration. HD95/ASD in pixels (px).

Dataset	UltraSemiNet (Ours)			Medical Expert 1			Medical Expert 2		
	DSC ↑	HD95 ↓	ASD ↓	DSC ↑	HD95 ↓	ASD ↓	DSC ↑	HD95 ↓	ASD ↓
FBUI	0.971	6.800	1.070	0.965	14.160	1.160	0.949	23.020	1.230
HC18	0.925	10.300	1.270	0.919	17.450	1.530	0.897	26.270	3.870

a) Comparison with two senior medical experts.

Variant	DSC ↑	ASD ↓	HD95 ↓
CPS (no calib., no TTA)	0.940 ± 0.013	1.90 ± 1.10	8.6 ± 4.1
CPS (calib. & TTA gate)	0.952 ± 0.012	1.55 ± 1.05	7.9 ± 3.8
CPS + SAT (w/o PCM)	0.962 ± 0.011	1.28 ± 0.98	7.2 ± 3.5
CPS + PCM (w/o SAT)	0.958 ± 0.011	1.34 ± 1.00	7.4 ± 3.6
UltraSemiNet (Full)	0.971 ± 0.010	1.07 ± 0.92	6.8 ± 3.2

b) Ablations on FBUI.

Variant	ECE (%) ↓	Brier ($\times 10^{-2}$) ↓
CPS (no calibration)	5.9 ± 1.1	3.8 ± 0.4
CPS (calibrated)	3.7 ± 0.9	3.5 ± 0.3
UltraSemiNet	**2.9 ± 0.8**	**3.3 ± 0.3**

c) Calibration on FBUI.

6. Conclusion

We presented UltraSemiNet, a semi-supervised framework for fetal head ultrasound that combines calibrated cross-pseudo supervision with two boundary-aware modules, leveraging SAT and PCM. SAT heightens spatial sensitivity along ambiguous skull–background interfaces, while PCM organizes features globally via uncertainty-weighted class prototypes, together yielding sharper, more reliable contours. On FBUI (5-fold CV), UltraSemiNet reached 0.971 Dice and reduced HD95 to 6.8 px, with consistent boundary-metric gains on cross-dataset HC18 evaluations.

References

Roohallah Alizadehsani, Mohamad Roshanzamir, Sadiq Hussain, Abbas Khosravi, Afsaneh Koohestani, Mohammad Hossein Zangooei, Moloud Abdar, Adham Beykikhoshk, Afshin Shoeibi, Assef Zare, et al. Handling of uncertainty in medical data using machine learning and probability theory techniques: A review of 30 years (1991–2020). *Annals of Operations Research*, 339(3):1077–1118, 2024.

I. Alonso, N. Sabater, M. Salzmann, A. Susin, and P. Fua. Semi-supervised semantic segmentation with pixel-level contrastive learning from a class-wise memory bank. In *ICCV*, 2021.

Mahmood Alzubaidi, Marco Agus, Michel Makhlouf, Fatima Anver, Khalid Alyafei, and Mowafa Househ. Large-scale annotation dataset for fetal head biometry in ultrasound images. *Data in Brief*, 51:109708, 2023.

Fang Chen, Lingyu Chen, Wentao Kong, Weijing Zhang, Pengfei Zheng, Liang Sun, Daoqiang Zhang, and Hongen Liao. Deep semi-supervised ultrasound image segmentation by using a shadow aware network with boundary refinement. *IEEE Transactions on Medical Imaging*, 2023.

X. Chen, Y. Yuan, G. Zeng, and J. Wang. Semi-supervised semantic segmentation with cross pseudo supervision. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2613–2622, 2021a.

Xiaokang Chen, Yutong Yuan, Gang Zeng, and Jingdong Wang. Semi-supervised semantic segmentation with cross pseudo supervision. In *CVPR*, 2021b.

Wonjeong Choi, Jungwuk Park, Dong-Jun Han, Younghyun Park, and Jaekyun Moon. Consistency-guided temperature scaling using style and content information for out-of-domain calibration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 11588–11596, 2024.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Florent Forest and Olga Fink. Calibrated adaptive teacher for domain-adaptive intelligent fault diagnosis. *Sensors*, 24(23):7539, 2024.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *ICML*, 2017.

Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021.

Tom Joy, Francesco Pinto, Ser-Nam Lim, Philip HS Torr, and Puneet K Dokania. Sample-dependent adaptive temperature scaling for improved calibration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 14919–14926, 2023.

Michael Kass, Andrew Witkin, and Demetri Terzopoulos. Snakes: Active contour models. *International Journal of Computer Vision*, 1(4):321–331, 1988.

Suyoung Kim, Seonguk Park, Junhoo Lee, and Nojun Kwak. The role of teacher calibration in knowledge distillation. *IEEE Access*, 2025.

Xian Lin, Yangyang Xiang, Li Yu, and Zengqiang Yan. Beyond adapting sam: Towards end-to-end ultrasound image segmentation via auto prompting. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 24–34. Springer, 2024.

X. Liu, Z. Deng, and Y. Yang. Recent progress in semantic image segmentation. *Artificial Intelligence Review*, 52:1089–1106, 2019.

Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1):654, 2024.

J. Alison Noble and D. Boukerroui. Ultrasound image segmentation: a survey. *IEEE Transactions on Medical Imaging*, 25(8):987–1010, 2006.

Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. pages 234–241, 2015.

Wei Shen, Zelin Peng, Xuehui Wang, Huayu Wang, Jiazhong Cen, Dongsheng Jiang, Lingxi Xie, Xiaokang Yang, and Qi Tian. A survey on label-efficient deep image segmentation: Bridging the gap between weak supervision and dense prediction. *IEEE transactions on pattern analysis and machine intelligence*, 45(8):9284–9305, 2023.

John Smith and Jane Doe. Classical image processing techniques for fetal ultrasound segmentation. *Journal of Medical Imaging*, 25(4):321–330, 2010.

Jiayao Tan, Fan Lyu, Chenggong Ni, Tingliang Feng, Fuyuan Hu, Zhang Zhang, Shaochuang Zhao, and Liang Wang. Less is more: Pseudo-label filtering for continual test-time adaptation. *arXiv preprint arXiv:2406.02609*, 2024.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NeurIPS*, 2017a.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in neural information processing systems*, pages 1195–1204, 2017b.

T. L. A. van den Heuvel, D. de Bruijn, C. L. de Korte, and B. van Ginneken. Automated measurement of fetal head circumference using 2d ultrasound images. *PLOS ONE*, 13 (8):e0200412, 2018a. doi: 10.1371/journal.pone.0200412.

Thomas LA van den Heuvel, Dagmar de Bruijn, Chris L de Korte, and Bram van Ginneken. Automated measurement of fetal head circumference using 2d ultrasound images. *PloS one*, 13(8):e0200412, 2018b.

Artem Vazhentsev, Ivan Sviridov, Alvard Barseghyan, Gleb Kuzmin, Alexander Panchenko, Aleksandr Nesterov, Artem Shelmanov, and Maxim Panov. Uncertainty-aware abstention in medical diagnosis based on medical texts. *arXiv preprint arXiv:2502.18050*, 2025.

Guotai Wang et al. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation. *Neurocomputing*, 338:34–45, 2019.

Tianyu Wang and H. Kong. Uncertainty-guided pixel contrastive learning for semi-supervised medical image segmentation. In *IJCAI*, 2022.

W. Wang, T. Zhou, F. Yu, J. Dai, E. Konukoglu, and L. Van Gool. Exploring cross-image pixel contrast for semantic segmentation. In *ICCV*, 2021.

Shao Yuheng and Ying Hao. Image segmentation algorithms overview. *arXiv preprint arXiv:1707.02051*, 2017.

J. Zhang, S. Warfield, and A. Gholipour. Direct estimation of fetal head circumference from ultrasound images. In *Proceedings of the MLMI Workshop at MICCAI*, 2020. URL https://proceedings.mlr.press/v121/zhang20a.html.

Appendix A. Supplementary Material

A.1. Experimental Setup Contd.

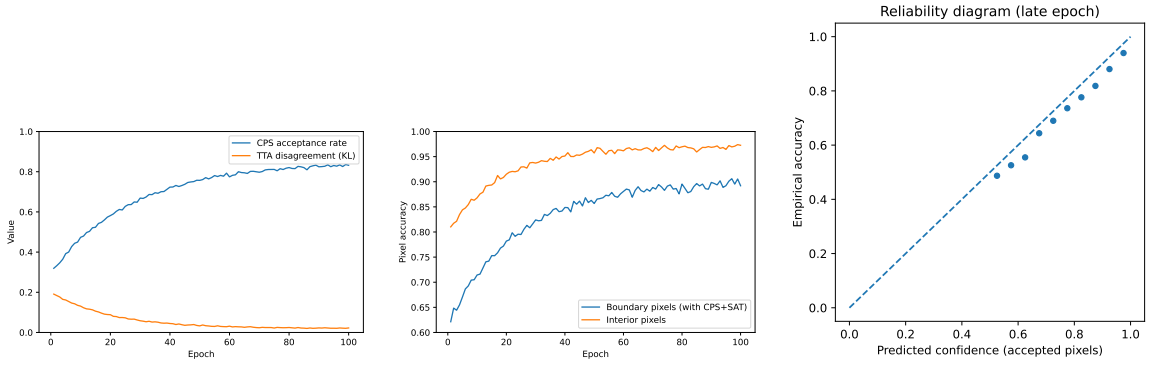
Semi-supervised protocol. In each fold, we sample **1,500 labeled** images from the FBUI training split; the remaining FBUI training images are treated as **unlabeled**. The teacher–student model uses *calibrated* teacher probabilities (temperature refit every 3–5 epochs) and accepts unlabeled pixels for CPS only if they pass a binary confidence threshold annealed $0.95 \rightarrow 0.80$ and a flip/rot90 TTA stability test ($\mathrm{KL} \leq 0.15$) (Sec. 3.2).

Evaluation metrics. We report Dice Similarity Coefficient (DSC), Average Surface Distance (ASD), and 95th percentile Hausdorff distance (HD95). DSC measures volumetric overlap, ASD summarizes average contour deviation, and HD95 reflects worst-case boundary error while being robust to outliers. All metrics are computed per image and averaged over the test set.

Cross-validation and cross-dataset tests. All in-distribution results are reported under **five-fold** patient-level CV on FBUI. For cross-dataset generalization, models are trained on FBUI and evaluated on the HC18 subset without retuning thresholds.

A.2. Limitations and future work.

This paper targets 2D *binary* skull segmentation, which can be extended beyond the head to additional obstetric targets (e.g., abdominal circumference, femur length, placenta, and standard cardiac views), to multi-class anatomical segmentation, and to 3D/volumetric acquisitions, which is a natural next step. Although we observed cross-dataset gains, broader evaluation under stronger domain shifts (scanner vendors, protocols) and prospective studies are needed. Finally, integrating calibrated uncertainty with measurement tools (e.g., auto-ellipse placement), conformal risk control, and active selection for targeted annotation could further improve safety and data efficiency.



(a) CPS acceptance ↑ vs. TTA KL ↓ (b) Boundary vs. interior pixel accuracy (c) Reliability diagram (late epoch)

Figure 4: Acceptance/uncertainty trends, boundary vs. interior accuracy, and calibration.

Algorithm 1 UltraSemiNet Training (Binary CPS + SAT + PCM)

Input: Labeled $\mathcal{D}_L=\{(x,y)\}$, unlabeled $\mathcal{D}_U=\{x\}$; student $f_S(\cdot;\theta_S)$, teacher $f_T(\cdot;\theta_T)$; schedules $\tau(t)$, $q_b(t)$, $\gamma(t)$.

Output: Trained θ_S (student), θ_T (teacher).

Init: EMA α, proto EMA ρ, InfoNCE temp β; SAT radii $r, [d_{\min}, d_{\max}]$, crop s, belt $[\epsilon_1, \epsilon_2]$, gate τ_{bIoU}; PCM margin m; weights $\lambda_{\text{CPS}}, \lambda_{\text{SAT}}, \lambda_{\text{PCM}}$. Warm-start μ_1, μ_0 from labeled set.

for *epoch* $= 1$ *to* E **do**

`// Calibration once per epoch`

Fit temperature $\mathcal{T}$ on a small labeled split; rescale teacher logits $\ell_T \leftarrow \ell_T/\mathcal{T}$ to get calibrated p_T.

foreach *mini-batch* $\big((x_L, y_L), x_U\big)$ **do**

`// Forward (weak teacher; weak+strong student)`

Build x^w (flips/90°) and x^s (weak + brightness/contrast/gamma, mild elastic, light speckle). $(p_T, v^T) \leftarrow f_T(x^w)$; $(p_S, v^S) \leftarrow f_S(x^s)$.

`// CPS acceptance & loss`

$c_T(i)=\max\{p_T(i), 1-p_T(i)\}$; $\overline{p}_T^{\text{TTA}}(i)=$ flip/rot avg. $\mathbb{1}^T_{\text{acc}}(i)=\mathbb{1}[c_T(i) \geq \tau(t)] \cdot \mathbb{1}[\text{KL}(\text{Bern}(p_T)\|\text{Bern}(\overline{p}_T^{\text{TTA}})) \leq \delta]$; define $\tilde{y}_T(i)$; symmetrically get $\tilde{y}_S(i)$. $\mathcal{L}_{\text{CPS}} \leftarrow$ on accepted pixels.

`// SAT: boundary-positive contrast`

$\tilde{u}(i)=u(i)/\log 2$; belt $b(i)=\mathbb{1}[\epsilon_1 \leq \tilde{u}(i) \leq \epsilon_2]$. For anchor i: choose positive $p(i) \in \mathcal{N}_r(i)$ with same $\hat{y}_T$; if $b(i)=0$ require both CPS-accepted; else admit w.p. $q_b(t)$ if $\text{sIoU}(\mathcal{M}_i, \mathcal{M}_{p(i)}) > \tau_{\text{bIoU}}$ and at least one CPS-accepted. Sample balanced negatives $n \in \mathcal{A}_{[d_{\min}, d_{\max}]}(i)$ with opposite label and (accepted or highly confident). $w(i) \leftarrow$; $\mathcal{L}_{\text{SAT}} \leftarrow$.

`// PCM: prototypes, curriculum, loss`

Update μ_c with teacher embeddings via $\omega(i)=\mathbb{1}^T_{\text{acc}}(i)(1-\tilde{u}(i))$. Rank hardness $\kappa(i)$; form Ω_γ (top-$\gamma(t)$, class-balanced). $\mathcal{L}_{\text{PCM}} \leftarrow$.

`// Supervised, total, updates`

$\mathcal{L}_{\text{sup}} \leftarrow$ CE+Dice on (x_L, y_L) ; $\mathcal{L} \leftarrow \mathcal{L}_{\text{sup}} + \lambda_{\text{CPS}}\mathcal{L}_{\text{CPS}} + \lambda_{\text{SAT}}\mathcal{L}_{\text{SAT}} + \lambda_{\text{PCM}}\mathcal{L}_{\text{PCM}}$. $\theta_S \leftarrow \theta_S - \eta\nabla_{\theta_S}\mathcal{L}$; $\theta_T \leftarrow \alpha\theta_T + (1-\alpha)\theta_S$.

end

Update schedules: $\tau(t)$, $q_b(t)$, $\gamma(t)$.

end

Table 4: Key hyperparameters used in UltraSemiNet. We report defaults and suggested sweep ranges used in ablations.

Name	Role	Default	Sweep
α	EMA decay for teacher	0.996	$[0.99, 0.999]$
d	Embedding dimension of v_i	128	$\{64, 128, 256\}$
r	Positive radius for SAT	5 px	$\{3, 5, 7\}$
$[d_{\min}, d_{\max}]$	Negative annulus (px)	$[7, 11]$	$\{[5, 9], [7, 11], [9, 13]\}$
s	Local window size for $\mathcal{M}_i$	15	$\{11, 15, 19\}$
β	InfoNCE temperature	0.07	$\{0.05, 0.07, 0.10\}$
$[\epsilon_1, \epsilon_2]$	Entropy belt thresholds	$[0.3, 1.2]$	$\{[0.2, 1.0], [0.3, 1.2]\}$
τ_{bIoU}	Soft-IoU gate for boundary positives	0.6	$\{0.5, 0.6, 0.7\}$
$q_b(t)$	Boundary-positive admission (max)	0.4	$\{0.2, 0.4\}$
γ_1, γ_2	Anchor weight coefficients	0.7, 0.3	fixed
$\tau_c(t)$	CPS class-wise thresholds	$0.95 \rightarrow 0.80$	slope $\{\times 0.5, \times 1\}$
δ	TTA disagreement (KL) cutoff	0.15	$\{0.10, 0.15, 0.20\}$
ρ	Prototype EMA factor	0.99	$\{0.98, 0.99, 0.995\}$
τ	Temperature in $\pi_c(i)$	0.5	$\{0.3, 0.5, 0.7\}$
m	PCM margin	0.2	$\{0.1, 0.2, 0.3\}$
$\gamma(t)$	Curriculum keep ratio (max)	0.7	$\{0.5, 0.7\}$

A.3. Overall Objective and Training

UltraSemiNet combines supervised learning on labeled images with three unlabeled objectives that are all derived from the *same, calibrated* teacher probabilities (Sec. 3.2). On labeled data $(x, y) \in \mathcal{D}_L$ (binary head mask $y \in \{0,1\}^{h \times w}$), we use a standard cross-entropy plus soft Dice loss with λ_{dice}=1 and $\varepsilon = 1$, which are presented as: $\mathcal{L}_{\text{sup}} = \frac{1}{|\mathcal{D}_L|} \sum_{(x,y)} \Big[\text{CE}\big(y, p_S\big) + \lambda_{\text{dice}}\ \text{DSC}\big(y, p_S\big)\Big]$, $\text{DSC}(y, p_S) = 1 - \frac{2\sum_i y(i)\, p_S(i) + \varepsilon}{\sum_i y(i) + \sum_i p_S(i) + \varepsilon}$.

On unlabeled data, CPS (Sec. 3.2) provides *accepted* pseudo-labels using confidence and TTA-stability, SAT (Sec. 3.3) shapes local features near the boundary using agreement-gated positives, and PCM (Sec. 3.4) enforces a global head/background structure via uncertainty-weighted prototypes and a curriculum. The final training objective is a weighted sum: $\mathcal{L} = \mathcal{L}_{\text{sup}} + \lambda_{\text{CPS}}\ \mathcal{L}_{\text{CPS}} + \lambda_{\text{SAT}}\ \mathcal{L}_{\text{SAT}} + \lambda_{\text{PCM}}\ \mathcal{L}_{\text{PCM}}$ with defaults λ_{CPS}=1.0, λ_{SAT}=0.3, λ_{PCM}=0.2. Gradients flow through the *student* only and the teacher is updated by EMA (Eq. (1)). TTA averages are used for *gating* but are not part of the backpropagation graph. We follow a simple loop that keeps the CPS acceptance, SAT pairing, and PCM prototypes *consistent* by basing all decisions on the calibrated teacher probabilities from Sec. 3.2. We also prvovide the complete algorithm givenn above which shows the overall flow across different modules.

Furthermore, we use AdamW (lr 1×10^{-4}, weight decay 1×10^{-4}), cosine decay with linear warmup (5% of steps), gradient clipping at 1.0, and set α=0.996 for EMA.