SPLITFED-CL: A SPLIT FEDERATED CO-LEARNING FRAMEWORK FOR MEDICAL IMAGE SEGMENTATION WITH INACCURATE LABELS

Zahra Hafezi Kafshgari, Hadi Hadizadeh and Parvaneh Saeedi

School of Engineering Science, Simon Fraser University, Burnaby, BC, Canada

ABSTRACT

Split Federated Learning (SplitFed) combines federated and split learning to preserve privacy while reducing client-side computation. However, in medical image segmentation, heterogeneous label quality across clients can significantly degrade performance. We propose SplitFed-CL, a co-learning framework where a global teacher guides local students to detect and refine unreliable annotations. Reliable labels supervise training directly, while unreliable labels are corrected via weighted student–teacher refinement. SplitFed-CL further incorporates consistency regularization for robustness to input perturbations and a trainable weighting module to balance loss terms adaptively. We also introduce a novel difficulty guided strategy to simulate human like boundary centric annotation errors, where the degree of perturbation is governed by shape complexity and the associated annotation difficulty. Experiments on two multiclass segmentation datasets with controlled synthetic noise, together with a binary segmentation dataset containing real-world annotation errors, demonstrate that SplitFed-CL consistently outperforms seven state-of-the-art baselines, yielding improved segmentation quality and robustness.

***Index Terms*—** SplitFed, label correction, co-learning

1 Introduction

Split Federated Learning (SplitFed) is a decentralized paradigm that integrates Split Learning (SL) [1] and Federated Learning (FL) [2] to achieve privacy-preserving, resource-efficient model training. In FL, clients retain data locally but must train full local models and transmit parameters for aggregation, which can require substantial on-device compute [3]. SL reduces this load by splitting the network between clients and server, exchanging intermediate activations instead of raw data [1]. SplitFed combines these strengths: data never leave the clients while the server executes the heavy layers, enabling participation by resource-constrained devices [4].

In medical image segmentation, distributed datasets often suffer from inconsistent label quality due to varying annotator expertise [5]. These inaccuracies degrade the global model, especially in decentralized settings [3].

In this paper, we propose SplitFed-CL, a co-learning framework for multiclass medical image segmentation that improves SplitFed training under heterogeneous and noisy annotations. Each client trains a student model guided by a global teacher that estimates label reliability. Reliable samples are used to directly update the model, while unreliable samples are refined through student–teacher predictions and incorporated with adaptive weighting. To systematically evaluate robustness under realistic annotation noise, we further introduce a difficulty-guided deformation strategy for simulating human-like segmentation errors. Moreover, we employ a consistency loss to enforce prediction invariance under input perturbations, together with a trainable loss-weighting module that automatically balances the reliable, unreliable, and consistency objectives, thereby eliminating manual tuning. Our main contributions are:

1. A difficulty-guided framework for simulating human-like annotation errors.
2. A global confidence-based mechanism for identifying reliable and unreliable annotations across clients.
3. A student–teacher strategy for refining unreliable labels.
4. A trainable loss-weighting module that automatically optimizes component contributions.

2 Related Work

Several FL methods proposed to mitigate the effect of annotation noise for *classification* tasks. For example, FedLN [6] improves robustness through interpolation-based regularization and energy-driven scoring, while FedNoiL [7] identifies reliable clients and corrects labels using global model predictions. FedCorr [8] further incorporates intrinsic dimensionality estimation and Gaussian Mixture Models to detect and revise noisy labels, and FNBench [9] offers a benchmarking framework for evaluating FL algorithms under label noise. FedNCL [10] introduces noise-robust aggregation by weighting clients based on sample reliability, while ARFL [11] dynamically adjusts client influence via data quality.

However, label correction in classification does not directly extend to segmentation, especially in medical imaging, where fuzzy boundaries hinder annotation accuracy. Recent works tackle limited supervision and noisy data in FL segmentation. FedMix [12] combines labeled and unlabeled data using consistency regularization and mix-up. FedA^{3}I [13] explicitly integrates annotation quality into ag-

gregation. FedDM [14] mitigates weak supervision and gradient conflicts through calibration and de-conflicting strategies. QA-SplitFed [15] addresses inaccurate annotations in split federated learning by adaptively weighting client contributions based on estimated label quality. Lastly, DHLC and CELC [16] refine lesion labels via student–teacher models, with CELC updating the global model using teacher parameters.

In addition, most prior works simulate label noise using simplistic perturbations such as random label shuffling or basic contour dilation/erosion [11, 15, 16]; in contrast, we propose a novel human-like annotation error method that generates realistic boundary deformations guided by a contour-based difficulty map.

3 Simulating Annotation Error using a Difficulty Map

To simulate human-like annotation errors in segmentation masks, we generate deformed version of accurate labels guided by a *difficulty map* that emphasizes boundary regions where annotators are more likely to disagree. Intuitively, pixels with a higher probability of misannotation receive higher difficulty scores.

Given an input image x and its corresponding mask y, we compute the signed distance function (SDF) ϕ from y (negative inside the object and positive outside), and restrict all computations to a narrow boundary band as $B(u) = \mathbb{1}\big(|\phi(u)| \le w\big)$ where u denotes a pixel location and w is the band width. We set $w = 0.2\sqrt{\frac{A_{\text{obj}}}{\pi}}$, where A_{obj} is the object area in pixels and $\sqrt{\frac{A_{\text{obj}}}{\pi}}$ is the equivalent radius, enabling scale-adaptive boundary modeling.

Boundary difficulty cues: Within $B(u)$, we estimate three cues that capture boundary uncertainty:

- **Edge cue (D_{edge}):** Weak edges are harder to delineate consistently. We compute the normalized gradient magnitude as $g(u) = \|\nabla x(u)\| \in (0,1)$, where ∇ denotes the spatial derivative operator [17], and define the edge-based difficulty as

$$D_{\text{edge}}(u) = \big(1 - g(u)\big)\, B(u) \tag{1}$$

  Therefore, strong edges yield low difficulty scores, whereas weak edges are assigned higher scores.
- **Blur cue (D_{blur}):** Reduced sharpness along object contours increases annotation uncertainty. We quantify local sharpness using the normalized magnitude of the Laplacian operator, $\textit{Laplacian}(u) = \big|\nabla^2 x(u)\big| \in (0,1)$, where ∇^2 denotes the second-order spatial derivative [18], and define

$$D_{\text{blur}}(u) = \big(1 - \textit{Laplacian}(u)\big)\, B(u) \tag{2}$$

  where higher difficulty values are assigned to more blurred boundary regions.
- **Curvature cue (D_{curv}):** Boundary points with high-curvatures are more prone to inconsistent delineation. Using the SDF, normalized curvature is approximated as $\kappa(u) \approx \nabla \cdot \left(\frac{\nabla\phi(u)}{\|\nabla\phi(u)\|}\right) \in (0,1)$ following [19], and we define

$$D_{\text{curv}}(u) = |\kappa(u)|\, B(u) \tag{3}$$

  Accordingly, sharply changing contours are assigned higher difficulty than near-flat boundary segments.

Combined difficulty map ($D(u)$): We combine the cues into a normalized difficulty map $D(u) \in [0,1]$:

$$D(u) = \frac{D_{\text{edge}}(u) + D_{\text{blur}}(u) + D_{\text{curve}}(u)}{3} \tag{4}$$

Deformation magnitude and direction: We map difficulty to a bounded deformation magnitude as:

$$A(u) = \Big(a_{\min} + (a_{\max} - a_{\min}) D(u)^{\rho}\Big)\, B(u) \tag{5}$$

Where ρ emphasizes high-difficulty boundary regions. Parameters $a_{\min}$, $a_{\max}$ and σ controls the deformation rate. In this work, we set $\rho = 2$ to emphasize high-difficulty boundaries, $a_{\min} = 0$ and $a_{\max} = w$.

For the deformation direction, we compute the outward unit normal $\mathbf{n}(u) = \nabla\phi(u)/\|\nabla\phi(u)\|$ and compare the edge evidence across the contour, where $g_{\text{out}}(u) = g\big(u + \delta\,\mathbf{n}(u)\big)$ measures the edge strength outside the boundary and $g_{\text{in}}(u) = g\big(u - \delta\,\mathbf{n}(u)\big)$ measures the edge strength inside the boundary. The signed displacement field is then defined as:

$$b(u) = \text{clip}\left(\frac{g_{\text{in}}(u) - g_{\text{out}}(u)}{g_{\text{in}}(u) + g_{\text{out}}(u) + \varepsilon}, -1, 1\right), \tag{6}$$

where ε is a small constant added for numerical stability. Following $b(u)$, if $g_{\text{in}}(u) > g_{\text{out}}(u)$, stronger evidence lies inside and the boundary is encouraged to shrink, whereas if $g_{\text{in}}(u) < g_{\text{out}}(u)$, stronger evidence lies outside and the boundary is encouraged to expand.

Deformed mask: We deform the SDF and threshold it to obtain the noisy annotation as:

$$\phi'(u) = \phi(u) + \big(G_\sigma * (A\,b)\big)(u), \quad y_{\text{noisy}} = \mathbb{1}\big(\phi'(u) \le 0\big), \tag{7}$$

where $A(u)$ controls the deformation magnitude and $b(u)$ specifies its direction at pixel u. G_σ denotes a Gaussian smoothing kernel with standard deviation σ, which suppresses unnaturally jagged boundaries. we set $\sigma = w$ in this study. This pipeline generates realistic, boundary-centric annotation errors whose location, magnitude, and direction are guided by the image's learned difficulty structure. For multi-class segmentation, each class is deformed using its own class-specific difficulty map, resulting in more realistic corrupted labels than uniformly deforming the entire mask. Representative examples of the resulting deformed labels are shown in Fig. 2, row **(b)**.

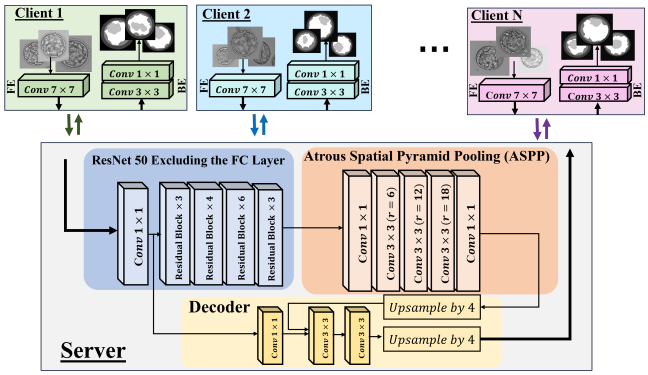

Fig. 1: SplitFed DeepLabV3+ architecture: client-side FE/BE sub-models with a server-side sub-model for heavy computation. ResNet50 [20] is used as the encoder backbone.

4 Proposed SplitFed Method

We adopt DeepLabV3+ [21] and partition it across client and server (Fig. 1). On each client, a lightweight front-end (FE) sub-model $f_{\theta_{\text{FE}}}$ encodes the input and transmits activations to the server sub-model $f_{\theta_{\text{S}}}$ for heavy computation; the returned feature maps are decoded by the client back-end (BE) sub-model $f_{\theta_{\text{BE}}}$ to produce the final prediction. This preserves data locality while offloading most compute to the server. At the start of each global round, the server distributes updated sub-models to clients; after local training, clients return updated weights for aggregation. We adopt a student–teacher setup: each client trains a student $\mathcal{F} = \{f_{\theta_{\text{FE}}}, f_{\theta_{\text{S}}}, f_{\theta_{\text{BE}}}\}$, while the globally averaged model serves as a teacher $\bar{\mathcal{F}} = \{\bar{f}_{\theta_{\text{FE}}}, \bar{f}_{\theta_{\text{S}}}, \bar{f}_{\theta_{\text{BE}}}\}$ that guides label reliability and provides a stable target for consistency.

For a batch $X = \{x_k\}_{k=1}^{K}$ with labels $Y = \{y_k\}_{k=1}^{K}$, the student and teacher produce predictions $f_{\theta_{\text{BE}}}(X)$ and $\bar{f}_{\theta_{\text{BE}}}(X)$ and per-sample region losses $\mathcal{L}_R = [\ell_1, \ldots, \ell_K]$, $\bar{\mathcal{L}}_R = [\bar{\ell}_1, \ldots, \bar{\ell}_K]$. Using a global threshold τ, define the index sets

$$\begin{aligned} X_{\text{re}}, Y_{\text{re}} &= \{\, x_k \in X, y_k \in Y \mid \ell_k \le \tau \text{ and } \bar{\ell}_k \le \tau \,\}, \\ X_{\text{un}}, Y_{\text{un}} &= \{\, x_k \in X, y_k \in Y \mid \ell_k > \tau \text{ and } \bar{\ell}_k > \tau \,\}. \end{aligned} \tag{8}$$

As detailed in Sec. 4.2, τ starts high (most samples treated as reliable) and is updated each round using cross-client statistics and performance.

Unreliable labels Y_{un} are progressively refined to $\tilde{Y}_{\text{un}}$ using student–teacher predictions (Sec. 4.3), so they contribute proportionally while limiting noise. Robustness is further improved via a co-training–style consistency term that encourages the student on perturbed inputs X' to match the teacher on original inputs X. The student is trained with:

$$\begin{aligned} \mathcal{L}_{\text{total}} = \mathcal{L}_R\big(f_{\theta_{\text{BE}}}(X_{\text{re}}), Y_{\text{re}}\big) + \alpha\, \mathcal{L}_R\big(f_{\theta_{\text{BE}}}(X_{\text{un}}), \tilde{Y}_{\text{un}}\big) \\ + \beta\, \mathcal{L}_C\big(f_{\theta_{\text{BE}}}(X'), \tilde{f}_{\theta_{\text{BE}}}(X)\big), \end{aligned} \tag{9}$$

where $\mathcal{L}_R$ is the region loss (applied to reliable and refined–unreliable labels) and $\mathcal{L}_C$ is a consistency term [22, 23]. The coefficients α and β automatically weight these components (see Sec. 4.4). After each local update on the student, the teacher sub-models are updated via an exponential moving average (EMA) of the student parameters, yielding a stable and progressively refined guidance signal [22].

4.1 SplitFed Averaging Method

After local training, each client transmits its student parameters together with summary statistics (total samples, reliable-sample count, and reliable-subset loss) to the server. This enables reliability-aware aggregation that accounts for both data quantity and quality.

Let d_{re}^i be the number of reliable samples for client $i = 1, .., N$ and $\mathcal{L}_{\text{re}}^i$ its mean reliable loss. Define the data ratio $\mathbf{d} = [d_{\text{re}}^i]_{i=1}^N / \sum_{j=1}^N d_{\text{re}}^j$ and the performance ratio $\mathbf{q} = \text{softmax}(-\gamma\, \mathbf{L})$ with $\mathbf{L} = [\mathcal{L}_{\text{re}}^i]_{i=1}^N$, where γ is initialized small and increased over global epochs to emphasize lower-loss clients. The contribution ratios are finally computed as $\mathbf{r} = \frac{\mathbf{q} \odot \mathbf{d}}{\mathbf{q}^\top \mathbf{d}} = [\, r_1, \ldots, r_N \,]^\top$. Using $\mathbf{r}$, global sub-model parameters are updated via a reliability-weighted average:

$$\{\bar{\theta}_{\text{FE}}, \bar{\theta}_{\text{S}}, \bar{\theta}_{\text{BE}}\} \leftarrow \left\{ \sum_{i=1}^{N} r_i\, \theta_{\text{FE}}^i,\ \sum_{i=1}^{N} r_i\, \theta_{\text{S}}^i,\ \sum_{i=1}^{N} r_i\, \theta_{\text{BE}}^i \right\} \tag{10}$$

The updated teacher and student parameters are then broadcast for the next round. By weighting clients based on the amount and quality of *reliable* data, rather than total volume, this aggregation reduces the influence of inaccurate annotations and improves global performance.

4.2 Updating Global Threshold

Motivated by the intuition that reliable samples can be identified through their training-loss behavior, we update the global threshold as a reliability-weighted statistic, $\tau = \mathbf{b}^\top \mathbf{q}$ where $\mathbf{b} = \big[\mu_1 + \lambda\sigma_1, \ldots, \mu_N + \lambda\sigma_N\big]$ and $\mathbf{q}$ is the performance-ratio vector (Sec. 4.1) and μ_i, σ_i denote the mean and standard deviation of per-sample training losses for client i. The scalar $\lambda > 0$ controls the upper bound of this confidence statistic and is decayed over global epochs, making τ increasingly selective.

Leveraging reliable-sample identification, τ uses $\mathbf{q}$ as reliability weights to apply the client-specific bounds in $\mathbf{b}$ when filtering all samples. Initially, τ is set to a high value (e.g., 10) so local models leverage most data; as training proceeds and λ shrinks, τ tightens to focus learning on more reliable samples, improving robustness.

4.3 Label Correction

After identifying reliable and unreliable labels in Y, unreliable labels, Y_{un}, are locally modified to $\bar{Y}_{un}$ using predictions from both the student and teacher models. A difference mask R is generated for each unreliable label to highlight areas of uncertainty between the predictions of the student model, teacher model, and the ground truth label, defined by $R = P_{\text{un}} \triangle \bar{P}_{\text{un}} \triangle Y_{\text{un}}$. Here, $\triangle$ is the symmetric difference, $P = \arg\max_{c \in \{1,\ldots,C\}} f_{\theta_{\text{BE}}}(X_{un})$ and $\overline{P} = \arg\max_{c \in \{1,\ldots,C\}} \bar{f}_{\theta_{\text{BE}}}(X_{un})$ are the hard-labeled

masks with C segmentation classes. Subsequently, the modified Y_{un} is computed as follows:

$$\tilde{Y}_{un}(D) = \begin{cases} P_{un}(R) & \text{if } f_{\theta_{\mathrm{BE}}}(X_{un}) > T \\ \bar{P}_{un}(R) & \text{if } \bar{f}_{\theta_{\mathrm{BE}}}(X_{un}) > T \\ Y_{un} & \text{otherwise} \end{cases}. \tag{11}$$

The modified labels $\tilde{Y}_{un}$ are then utilized as the ground truth to compute the region loss, as described in this section.

4.4 Learning Loss Coefficients

As defined in Eq. (9), the coefficients α and β weight the unreliable-subset region loss and the consistency loss, respectively. Manual tuning is costly and brittle. Inspired by Kendall–Gal [24], we replace fixed coefficients with a trainable weighting module and optimize them during training. Concretely, we use:

$$\mathcal{L}_{\mathrm{total}} = w_{\mathrm{re}}\, \widehat{\mathcal{L}}_R\big(f_{\theta_{\mathrm{BE}}}(X_{\mathrm{re}}), Y_{\mathrm{re}}\big) + w_{\mathrm{un}}\, \widehat{\mathcal{L}}_R\big(f_{\theta_{\mathrm{BE}}}(X_{\mathrm{un}}), \tilde{Y}_{\mathrm{un}}\big) + w_{\mathrm{cons}}\, \widehat{\mathcal{L}}_C\big(f_{\theta_{\mathrm{BE}}}(X'), \tilde{f}_{\theta_{\mathrm{BE}}}(X)\big) + \mathcal{R}(u), \tag{12}$$

with $w_{\mathrm{re}} = 1$ and $w_{un}(t) = \sigma(u_{\mathrm{un}})\, s(t) \in (0,1)$ and $w_{cons}(t) = \sigma(u_{\mathrm{cons}})\, s(t) \in (0,1)$. Here, $\sigma(\cdot)$ is the logistic sigmoid and $s(t)$ is a linear warm-up schedule that keeps w_{un} and w_{cons} small in early epochs. The logits u_{un} and u_{cons} are trainable; the terms $\widehat{\mathcal{L}}$ are EMA-normalized to maintain comparable, stable scales. We initialize $u_{\mathrm{un}}, u_{\mathrm{cons}} \ll 0$ (e.g., -10), so both weights start near zero $w_{\mathrm{cons}}, w_{\mathrm{noisy}} \approx 0$ and are allowed to grow toward 1 as training progresses. To mitigate overconfident saturation, we include a small ℓ_2 regularizer, $\mathcal{R}(u) = \eta\left(u_{\mathrm{un}}^2 + u_{\mathrm{cons}}^2\right)$. At each global round, local logits u_{un} and u_{cons} are aggregated using the reliability-weighted averaging rule described in Sec. 4.2. In all experiments, we set $\eta = 5 \times 10^{-4}$.

5 Experiments

5.1 Datasets

To evaluate the effectiveness of the proposed SplitFed-CL method, we utilize two multi-class and one binary medical image segmentation dataset. The *Human Embryo* dataset comprises 594 microscopic images of human blastocysts, with labels segmenting four structures: Zona Pellucida (ZP), Trophectoderm (TE), Inner Cell Mass (ICM), and Blastocoel (BL) [25]. In the SplitFed setup, the data are distributed heterogeneously across four clients (100, 150, 200, and 50 images, respectively), while 94 precisely annotated images are reserved for testing. The second dataset, Pubic Symphysis and Fetal Head Segmentation (*PSFHS*) [26], introduced in the MICCAI 2023 challenge, contains 1,358 annotated intrapartum transperineal ultrasound images for segmenting the pubic symphysis (PS) and fetal head (FH). Although newer samples have been released, we use only the initial subset due to its higher annotation reliability. The PSFHS data are allocated across four clients (100, 200, 450, and 250 images, respectively) with 358 images retained for evaluation. To model heterogeneous annotation quality, we corrupt a subset of each client's labels using the deformation strategy in Sec. 3, with corruption ratios of 20%, 50%, 80%, and 0% for Clients 1–4, respectively.

Table 1: Performance comparison on ***PSFHS*** dataset for **PS** & **FH** segments.

Method	Accuracy	Dice Loss	Mean IOU	PS IOU	FH IOU
All Accurate Labels (FedAVG [2])	0.9635	0.0426	0.9605	0.8701	0.9196
All Accurate Labels (***SplitFed-CL***)	0.9810	0.0418	0.9630	0.8630	0.9290
FedAVG [2]	0.9430	0.0585	0.9475	0.8191	0.8970
FedMix [12]	0.9412	0.0844	0.9430	0.6912	0.8984
FedNCL-V2 [10]	0.9326	0.0793	0.9456	0.7065	0.9055
ARFL [11]	0.9312	0.0856	0.9420	0.6913	0.8920
QA-SplitFed [15]	0.9578	0.0574	0.9493	0.8235	0.8982
CELC [16]	0.9127	0.1620	0.8701	0.5488	0.7547
DHLC [16]	0.9111	0.1650	0.8625	0.5570	0.7339
SplitFed-CL (No Label Correction)	0.9715	0.0425	0.9537	0.8561	0.9330
SplitFed-CL (No Consistency Loss)	0.9725	0.0399	0.9556	0.8660	0.9330
SplitFed-CL (full)	**0.9788**	**0.0363**	**0.9592**	**0.8805**	**0.9424**

To evaluate the proposed SplitFed-CL framework under a more realistic annotation noise condition, we included a third dataset, *ISIC*, a dermoscopic skin-lesion segmentation benchmark from the *ISIC* Archive [27]. Specifically, we used ISIC MultiAnnot++ (IMA++), which contains 12,573 single-annotation masks and 2,394 multi-annotator masks with tool/skill metadata [28, 29].

Using the IMA++ metadata, we selected 600 multi-annotator samples annotated using a semi-automated flood-fill tool (T2) with expert-defined parameters and verified by a novice (S2), denoted as *T2/S2*, and treated them as real-world unreliable annotations. We also selected 5,124 single-annotation masks generated via manual polygon tracing by an expert (T1) and verified by an expert (S1), denoted as *T1/S1*, and treated them as accurate samples. We distributed these samples across five clients containing 500, 800, 800, 1500, and 1000 images, respectively, while retaining 1,124 *T1/S1* images for evaluation. The corruption ratios for clients 1–5 were 0%, 50%, 50%, 80%, and 60%, respectively. The corrupted labels in client 5 corresponded to real-world noisy annotations (*T2/S2*), whereas those in client 4 were synthetically generated using the proposed difficulty-guided mask deformation strategy (Sec. 3).

5.2 Results and Discussion

We evaluated our SplitFed method against seven state-of-the-art baselines adapted to the SplitFed setting: FedAvg [2] (server aggregation $\overline{\mathbf{W}} = \left(\sum_{i=1}^{N} m_i\right)^{-1} \sum_{i=1}^{N} m_i\, \mathbf{W}_i$, with m_i the client sample counts); FedMix [12] ($\beta = 1.5$, $\lambda = 10$); FedNCL-V2 [10] ($\alpha = \beta = 2$); ARFL [11] (with $M_P = \lambda$ as in the original); Quality-Adaptive SplitFed (QA-SplitFed) [15] (aggregation tailored for noisy clients); and two label-correction methods, DHLC and CELC [16].

To our knowledge, QA-SplitFed [15] is the only SplitFed framework explicitly targeting noisy labels. Accordingly, for a fair comparison we ported the remaining FL baselines to the SplitFed protocol (same client–server partition, communication schedule, and comparable hyperparameter budgets).

Input images are resized to 352×352, and augmentation is performed using horizontal and vertical flipping, and rotation of up to $\pm 35^{\circ}$. The system is trained with the Adam optimizer

Table 2: Performance comparison on ***Human Embryo*** dataset for **ZP**, **TE**, **ICM** and **BL** segments.

Method	Accuracy	Dice Loss	Mean IoU	ZP IoU	TE IoU	ICM IoU	BL IoU
All Accurate Labels (FedAVG [2])	0.9368	0.0960	0.8661	0.7684	0.7339	0.8576	0.8605
All Accurate Labels (***SplitFed-CL***)	0.9448	0.0753	0.8960	0.8147	0.7816	0.8811	0.8898
FedAVG [2]	0.9234	0.1074	0.8557	0.7883	0.6846	0.8372	0.8405
FedMix [12]	0.9289	0.1033	0.8608	0.7780	0.7047	0.8423	0.8514
FedNCL-V2 [10]	0.9420	0.0776	0.8907	**0.8281**	0.7828	0.8702	0.8814
ARFL [11]	0.9365	0.0830	0.8839	0.8212	0.7713	0.8573	0.8732
QA-SplitFed [15]	0.9312	0.0945	0.8695	0.8077	0.7318	0.8460	0.8632
CELC [16]	0.9178	0.1387	0.8209	0.7296	0.5846	0.8156	0.8315
DHLC [16]	0.9232	0.1349	0.8225	0.7205	0.6037	0.8185	0.8416
SplitFed-CL (No Label Correction)	0.9326	0.0806	0.8825	0.8106	0.7533	0.8612	0.8655
SplitFed-CL (No Consistency Loss)	0.9434	0.0775	0.8933	0.8250	0.7651	0.8705	0.8799
SplitFed-CL (full)	**0.9451**	**0.0725**	**0.8995**	0.8265	**0.7862**	**0.8780**	**0.8923**

Table 3: Performance comparison on ***ISIC*** dataset.

Method	Accuracy	Dice Loss	FG IOU	Precision	Recall
All Accurate Labels (FedAVG [2])	0.9860	0.1320	0.7641	0.9073	0.9156
All Accurate Labels (***SplitFed-CL***)	0.9850	0.1377	0.7580	0.8974	0.9149
FedAVG [2]	0.9801	0.1616	0.7199	0.9044	0.8348
FedMix [12]	0.9811	0.1488	0.7326	0.8916	0.8467
FedNCL-V2 [10]	0.9822	0.1389	0.7401	0.8910	0.8532
ARFL [11]	0.9800	0.1411	0.7320	0.8901	0.8442
QA-SplitFed [15]	0.9827	0.1402	0.7432	0.9002	0.8732
CELC [16]	0.9555	0.1890	0.6912	0.8555	0.8321
DHLC [16]	0.9619	0.1853	0.7001	0.8584	0.8333
SplitFed-CL (No Label Correction)	0.9814	0.1544	0.7413	**0.9393**	0.8432
SplitFed-CL (No Consistency Loss)	0.9826	0.1438	0.7544	0.9330	0.8665
SplitFed-CL (full)	**0.9830**	**0.1320**	**0.7641**	0.9074	**0.9157**

with the learning rate set to 10^{-4}. All models are trained for 5 local epochs per round and validated over 100 global epochs. Threshold T in Equation 11 is set to 0.9. For evaluation, we report pixel-wise Accuracy ($\frac{TP+TN}{TP+TN+FP+FN}$), Dice loss ($1-\frac{2TP}{2TP+FP+FN}$), and IoU ($\frac{TP}{TP+FP+FN}$) computed from (TP, TN, FP, FN) for *Human Embryo* and *PSFHS*. For *ISIC*, we additionally report Precision ($\frac{TP}{TP+FP}$) and Recall ($\frac{TP}{TP+FN}$).

We define $\alpha = w_{\text{un}}(t)$ and $\beta = w_{\text{cons}}(t)$ with trainable logits $u_{\text{un}}, u_{\text{cons}}$. The logits are optimized with Adam (lr $= 10^{-3}$) and initialized as $u_{\text{un}} = u_{\text{cons}} = -10$, so $w_{\text{un}}(0), w_{\text{cons}}(0) \approx 0$. The temperature parameter γ in Sec. 4.1 is linearly warmed up from 1 to 5 over the first 50 global epochs, while λ in Sec. 4.2 decays linearly from 3 to 0 over the same period and then kept constant thereafter.

Tables 1–3 report quantitative results on *PSFHS*, *Human Embryo*, and *ISIC*. In each table, the first two rows represent the clean-label reference setting (all clients have accurate annotations), using FedAvg and ***SplitFed-CL*** for aggregation, respectively. For *ISIC*, this baseline is obtained by replacing *T2/S2* annotations with the corresponding *T1/S1* masks. Best results are shown in bold. Across all datasets, SplitFed-CL consistently surpasses prior SOTA methods and achieves performance closest to the clean-label reference. Ablation results further show that removing either label correction or consistency loss degrades performance, confirming the importance of both components.

To demonstrate the impact of label correction, Fig. 2 demonstrates some examples of accurate, corrupted, and modified labels from the three datasets. The mean IoU is calculated for corrupted and corrected labels against the accurate labels. For the last samples highlighted by the red box, corrupted labels are real *ISIC T2/S2* annotations; the rest of noisy labels are synthetically generated as described in Sec. 3. As shown in Fig. 2, SplitFed-CL refines corrupted labels, improving local training and ultimately enhancing the global model, consistent with Tables 1–3.

Fig. 3 tracks client-level quantities on *PSFHS* dataset. The learnable logits u_{un} and u_{cons} in Eq. (12) are optimized locally and aggregated per Sec. 4.1. By the end of training, the global logits converge to $u_{\text{un}} \approx 0$ and $u_{\text{cons}} \approx -1.77$, yielding $w_{\text{un}} = \sigma(u_{\text{un}}) = 0.5$

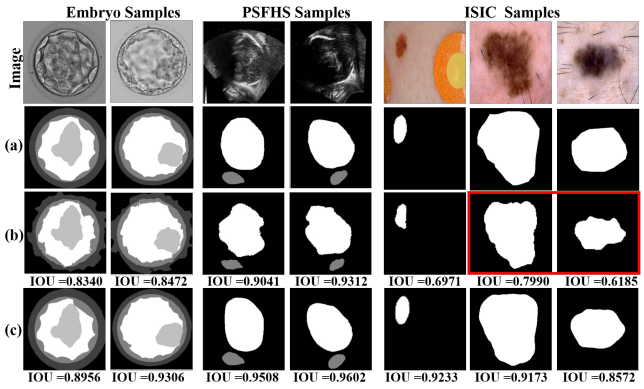

Fig. 2: Qualitative label-correction results on representative samples from three datasets under the proposed SplitFed-CL framework, showing **(a)** accurate, **(b)** inaccurate, and **(c)** corrected labels. Labels in the red box correspond to real-world *ISIC* T2/S2 annotations.

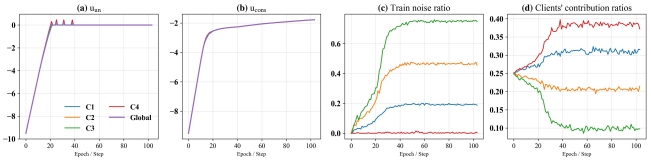

Fig. 3: *PSFHS*: Per-client training dynamics—**(a)** local u_{un}, **(b)** local u_{cons}, **(c)** train-noise ratio, **(d)** aggregation contribution; curves: C1 to C4

and $w_{\text{cons}} = \sigma(u_{\text{cons}}) \approx 0.1455$ for the unreliable-label and consistency losses, respectively. The model's estimated training-noise ratio (Fig. 3c) closely matches the injected corruption: ground truth 20%, 50%, 80%, 0% for C1–C4 versus detected 18.8%, 45.6%, 74.8%, 0.4% using the threshold τ from Sec. 4.3. Finally, the reliability-aware aggregation adapts client contributions over rounds (Fig. 3d): with γ initialized small, the ratios start similar and then shift toward lower-loss clients as γ increases (Sec. 4.1).

6 Conclusion

We introduce SplitFed-CL, a Split Federated co-learning framework for medical image segmentation under noisy and inconsistent labels. The method integrates student–teacher learning with reliability-aware aggregation to identify, refine, and reweight unreliable annotations. A global confidence metric separates reliable samples, while label correction, consistency regularization, and adaptive loss weighting jointly enhance robustness and stability. Experiments on two medical segmentation datasets with controlled label noise show that SplitFed-CL consistently outperforms existing FL and SplitFed baselines, achieving performance close to clean-label training. These results highlight its effectiveness in improving model reliability and noise resilience in decentralized medical imaging.

7 References

[1] P. Vepakomma, O. Gupta, T. Swedish, and R. Raskar, "Split learning for health: Distributed deep learning without sharing raw patient data," *arXiv preprint arXiv:1812.00564*, 2018.

[2] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial intelligence and statistics*. PMLR, 2017, pp. 1273–1282.

[3] H. Guan, P. T. Yap, A. Bozoki, and M. Liu, "Federated learning for medical image analysis: A survey," *Pattern Recognition*, p. 110424, 2024.

[4] Ch. Thapa, P. Ch. M. Arachchige, S. Camtepe, and L. Sun, "Splitfed: When federated learning meets split learning," in *AAAI Conference on Artificial Intelligence*, 2022, vol. 36, pp. 8485–8493.

[5] R. R. Kumar and R. Priyadarshi, "Denoising and segmentation in medical image analysis: A comprehensive review on machine learning and deep learning approaches," *Multimedia Tools and Applications*, vol. 84, no. 12, pp. 10817–10875, 2025.

[6] V. Tsouvalas, A. Saeed, T. Özçelebi, and N. Meratnia, "Federated learning with noisy labels: Achieving generalization in the face of label noise," in *First Workshop on Interpolation Regularizers and Beyond at NeurIPS 2022*, 2022.

[7] Zh. Wang, T. Zhou, G. Long, B. Han, and J. Jiang, "Fednoil: a simple two-level sampling method for federated learning with noisy labels," *arXiv preprint arXiv:2205.10110*, 2022.

[8] J. Xu, Z. Chen, T. QS. Quek, and K. F. E. Chong, "Fedcorr: Multi-stage federated learning for label noise correction," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10184–10193.

[9] X. Jiang, J. Li, N. Wu, Zh. Wu, X. Li, Sh. Sun, G. Xu, Y. Wang, Q. Li, and M. Liu, "Fnbench: Benchmarking robust federated learning against noisy labels," *arXiv preprint arXiv:2505.06684*, 2025.

[10] K. Tam, L. Li, B. Han, Ch. Xu, and H. Fu, "Federated noisy client learning," *arXiv preprint arXiv:2106.13239*, 2021.

[11] Sh. Li, E. Ngai, F. Ye, and Th. Voigt, "Auto-weighted robust federated learning with corrupted data sources," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 13, no. 5, pp. 1–20, 2022.

[12] J. Wicaksana, Z. Yan, D. Zhang, X. Huang, H. Wu, X. Yang, and K. T. Cheng, "Fedmix: Mixed supervised federated learning for medical image segmentation," *IEEE Transactions on Medical Imaging*, 2022.

[13] N. Wu, Zh. Sun, Z. Yan, and L. Yu, "Feda3i: annotation quality-aware aggregation for federated medical image segmentation against heterogeneous annotation noise," in *AAAI Conference on Artificial Intelligence*, 2024, vol. 38, pp. 15943–15951.

[14] M. Zhu, Zh. Chen, and Y. Yuan, "Feddm: Federated weakly supervised segmentation via annotation calibration and gradient de-conflicting," *IEEE Transactions on Medical Imaging*, vol. 42, no. 6, pp. 1632–1643, 2023.

[15] Z. H. Kafshgari, Ch. Shiranthika, P. Saeedi, and I. Bajić, "Quality-adaptive split-federated learning for segmenting medical images with inaccurate annotations," in *20th International Symposium on Biomedical Imaging*. IEEE, 2023, pp. 1–5.

[16] L. Bai, D. Wang, H. Wang, M. Barnett, M. Cabezas, W. Cai, F. Calamante, K. Kyle, D. Liu, L. Ly, et al., "Improving multiple sclerosis lesion segmentation across clinical sites: A federated learning approach with noise-resilient training," *Artificial Intelligence in Medicine*, vol. 152, pp. 102872, 2024.

[17] A. Pratondo, Ch. K. Chui, and S. H. Ong, "Robust edge-stop functions for edge-based active contour models in medical image segmentation," *IEEE Signal Processing Letters*, vol. 23, no. 2, pp. 222–226, 2016.

[18] F. Crété, T. Dolmiere, P. Ladret, and M. Nicolas, "The blur effect: Perception and estimation with a new no-reference perceptual blur metric," in *Proc. SPIE*, 2007.

[19] J. Ma, J. He, and X. Yang, "Learning geodesic active contours for embedding object global information in segmentation cnns," *IEEE Transactions on Medical Imaging*, vol. 40, no. 1, pp. 93–104, 2020.

[20] K. He, X. Zhang, Sh. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[21] L. Ch. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *the European conference on computer vision (ECCV)*, 2018, pp. 801–818.

[22] J. Peng, G. Estrada, M. Pedersoli, and Ch. Desrosiers, "Deep co-training for semi-supervised image segmentation," *Pattern Recognition*, vol. 107, pp. 107269, 2020.

[23] G. Sudhamsh, S. Girisha, and R. Rashmi, "Semi-supervised tissue segmentation from histopathological images with consistency regularization and uncertainty estimation," *Scientific Reports*, vol. 15, no. 1, pp. 6506, 2025.

[24] A. Kendall, Y. Gal, and R. Cipolla, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7482–7491.

[25] P. Saeedi, D. Yee, J. Au, and J. Havelock, "Automatic identification of human blastocyst components via texture," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 12, pp. 2968–2978, 2017.

[26] G. Chen, J. Bai, Zh. Ou, Y. Lu, and H. Wang, "Psfhs: intrapartum ultrasound image dataset for ai-based segmentation of pubic symphysis and fetal head," *Scientific Data*, vol. 11, no. 1, pp. 436, 2024.

[27] N. C. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler, and A. Halpern, "Skin lesion analysis toward melanoma detection: A challenge at isbi 2017," in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. IEEE, 2018, pp. 168–172.

[28] K. Abhishek, J. Kawahara, and Gh. Hamarneh, "Segmentation style discovery: Application to skin lesion images," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2024, pp. 24–34.

[29] K. Abhishek, J. Kawahara, and Gh. Hamarneh, "What can we learn from inter-annotator variability in skin lesion segmentation?," in *MICCAI Workshop on Deep Generative Models*. Springer, 2025, pp. 23–33.