FIVA: Federated Inverse Variance Averaging for Universal CT Segmentation with Uncertainty Estimation

Asim Ukaye [1] ASIM.UKAYE@MBZUAI.AC.AE
Numan Saeed [1] NUMAN.SAEED@MBZUAI.AC.AE
Karthik Nandakumar [1,2] NANDAKUM@MSU.EDU

[1] *Department of Computer Vision, Mohamed bin Zayed University of Artificial Intelligence*
[2] *Department of Computer Science and Engineering, Michigan State University*

Abstract

Different CT segmentation datasets are typically obtained from different scanners under different capture settings and often provide segmentation labels for a limited and often disjoint set of organs. Using these heterogeneous data effectively while preserving patient privacy can be challenging. This work presents a novel federated learning approach to achieve universal segmentation across diverse abdominal CT datasets by utilizing model uncertainty for aggregation and predictive uncertainty for inference. Our approach leverages the inherent noise in stochastic mini-batch gradient descent to estimate a distribution over the model weights to provide an on-the-go uncertainty over the model parameters at the client level. The parameters are then aggregated at the server using the additional uncertainty information using a Bayesian-inspired inverse-variance aggregation scheme. Furthermore, the proposed method quantifies prediction uncertainty by propagating the uncertainty from the model weights, providing confidence measures essential for clinical decision-making. In line with recent work shown, predictive uncertainty is utilized in the inference stage to improve predictive performance. Experimental evaluations demonstrate the effectiveness of this approach in improving both the quality of federated aggregation and uncertainty-weighted inference compared to previously established baselines. The code for this work is made available at: https://github.com/asimukaye/fiva

1. Introduction

Medical image segmentation plays a crucial role in clinical diagnostics, with applications ranging from tumor detection to organ delineation. While deep learning models have achieved remarkable success in this domain, data privacy regulations restrict patient data sharing, thereby limiting their full potential. Federated Learning (FL) enables collaborative model training across distributed datasets while preserving data privacy (McMahan et al. (2017)). In a typical horizontal FL setting, each client trains a local model on its private dataset and sends the model to a central server. The server aggregates these models to form a global model, which is then redistributed to the clients for the next training round. However, heterogeneity in client data due to differences in data quantity, quality, or acquisition equipment is known to degrade FL performance (Li et al. (2020)). Real-world settings such as medical imaging naturally exhibit this data heterogeneity due to the diversity of patients (e.g., region, ethnicity) and differences in imaging equipment, data collection, and

labeling practices at different healthcare centers. This makes the direct application of FL algorithms to medical imaging, and healthcare in general, an open problem (Rauniyar et al. (2024)).

On the other hand, uncertainty estimation in medical image segmentation is paramount due to its potential to enhance the reliability and interpretability of automated diagnostic tools. Kendall and Gal (2017) formulated uncertainty estimation in the context of deep learning to differentiate the impact of irreducible, data-dependent (aleatoric) uncertainty and reducible, model-dependent (epistemic) uncertainty. In medical image segmentation, capturing epistemic uncertainty helps to identify ambiguous regions and potential out-of-distribution samples (Jalal et al. (2024)). This motivates the need to develop methods that enable federated training of medical image segmentation models while providing reliable uncertainty estimates of predictions.

Recently, Tölle et al. (2024) proposed a novel approach to federated learning for medical image segmentation (FUNAvg) that leverages predictive uncertainty at the inference stage. By performing uncertainty-weighted averaging of output channels, FUNAvg improves predictive performance. We build on FUNAvg and draw inspiration from statistical meta-analysis such as inverse variance weighting for aggregating random variables (Hartung (2008), Borenstein (2013)). Based on this, we propose a novel strategy: **FIVA**: Federated Inverse Variance Averaging, that embeds uncertainty in the model and effectively utilizes it to improve aggregation during the federated learning step.

In this work, we propose using parameter uncertainty estimation to enhance federated learning for abdominal CT image segmentation. Each client estimates the mean and variance of the model parameters during local training by tracking the gradients and parameters at each iteration of mini-batch stochastic gradient descent. The clients then send these model parameters and their variances to the server. The central server aggregates these parameters, considering both their mean values and associated uncertainties. The resulting global model represents a posterior distribution over parameters, allowing inference on test clients by sampling from this distribution. We illustrate this approach in Figure 1.

Our key contributions are as follows.

- Leverage the stochasticity in mini-batch SGD to estimate a distribution over model parameters during client training. This enables each client to capture epistemic uncertainty locally efficiently.
- Introduce a Bayesian aggregation framework compatible with existing FL pipelines that utilizes this parameter distribution to improve both server-side aggregation and predictive uncertainty estimation during inference.
- Demonstrate improved performance over standard baselines through both quantitative and qualitative evaluations. We benchmark our method on abdominal CT segmentation tasks and show its robustness under varying data distributions.

Generalizable insights and clinical significance. This approach can significantly enhance diagnostic confidence through uncertainty-aware predictions. By quantifying uncertainty, healthcare professionals can better assess the reliability of segmentation predictions and identify cases that require expert review. Federated learning allows for privacy-

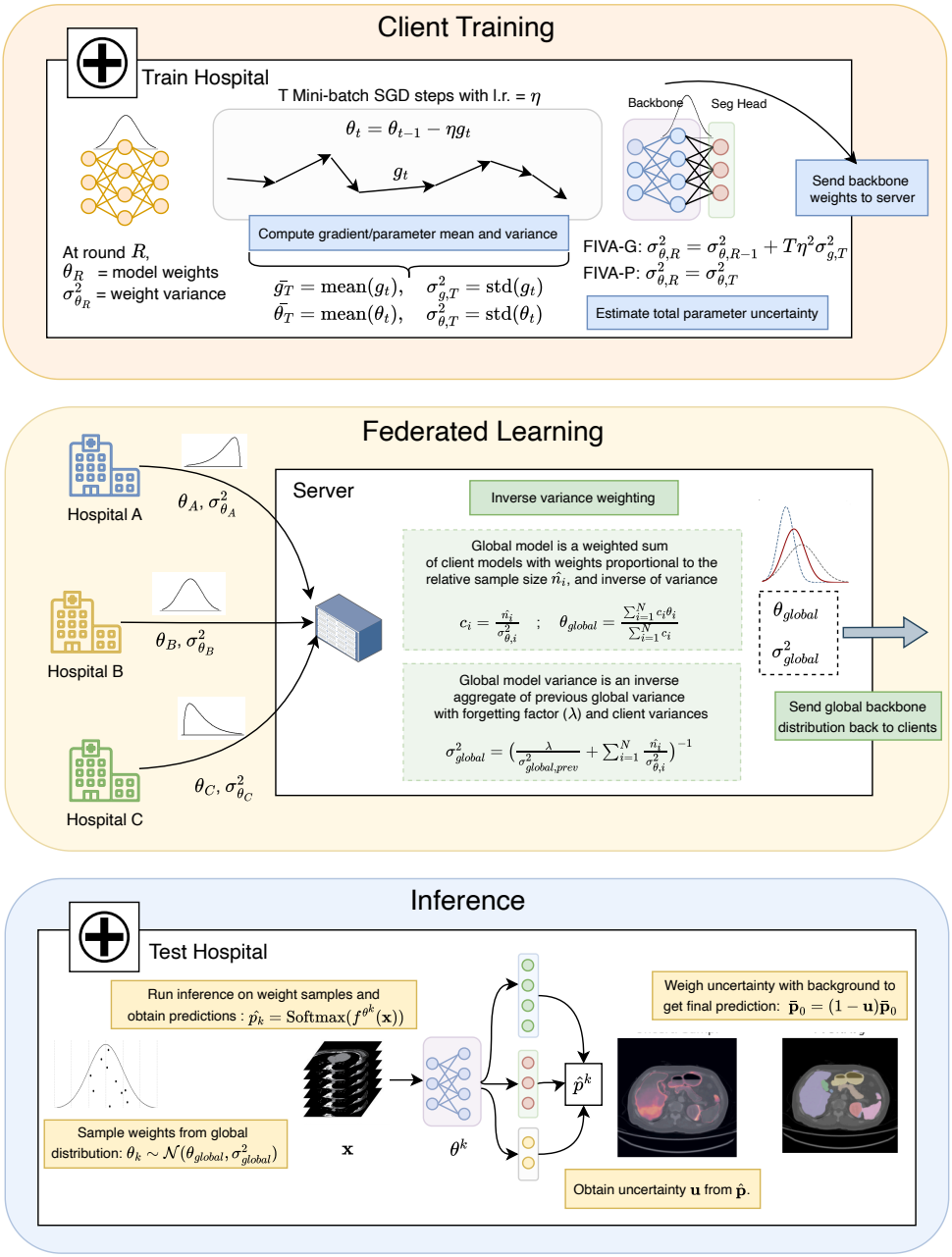


Figure 1: Overview of the proposed FIVA approach. Each client induces a distribution over model parameters during local training. These parameter distributions are aggregated on the server using inverse-variance weighting to form a global distribution. At inference time, the test client samples model parameters from the global distribution to obtain predictive uncertainty. Final predictions are computed by averaging stochastic forward passes and reweighting the background class using total uncertainty, following Tölle et al. (2024).

preserving collaboration across institutions, improving model generalization across diverse patient populations without centralized data collection.

2. Related Work

2.1. Universal Segmentation

Universal Segmentation aims to train models that generalize across datasets with diverse organ coverage, label availability, and acquisition settings. Recently, Butoi et al. (2023) introduced *UniverSeg*, a model capable of adapting to unseen medical segmentation tasks without requiring additional training or fine-tuning. Gao et al. (2024) propose *Hermes*, a context-prior learning approach that integrates task and modality priors into the segmentation process. Another line of work focuses on creating highly optimized end-to-end pipelines to enable generalization across diverse medical imaging tasks. For example, Isensee et al. (2021) introduced *nnU-Net*, a self-configuring pipeline built on the U-Net architecture (Ronneberger et al. (2015)) that is adaptable to various segmentation tasks. Wasserthal et al. (2023) propose *TotalSegmentator* which can segment 104 anatomical structures covering organs, bones, muscles, and major vessels. Other works such as *Multi-Talent*, (Ulrich et al. (2023)), and *Med3D* (Chen et al. (2019)) attempt to learn a shared representation between datasets while preserving label-specific heads.

2.2. Federated Learning in Medical Imaging

Federated Learning (FL) has emerged as a popular paradigm for training models on decentralized medical data while preserving patient privacy. Zhang et al. (2022) propose *pFedBayes* that performs personalized federated learning by treating both client and server neural networks as Bayesian Neural Networks. In the context of medical imaging, Xu et al. (2023) introduces *Fed-MENU*, which utilizes sub-networks that specialize in extracting features for specific organs. Bernecker et al. (2022) address variations in imaging modalities and scanner types in their work *FedNORM* for liver segmentation in a federated learning setting. Jiang et al. (2023) propose *FedCE* that addresses fairness in medical image segmentation. They estimated client contributions in an FL setting using gradient updates and server-side validation. Asokan et al. (2024) integrate Low-Rank Adapters into the *Segment Anything Model (SAM)* architecture to enable parameter-efficient fine-tuning of of 3D medical image segmentation under an FL setting. Xu et al. (2024) introduce *FedCross* to train the global model in a round-robin manner eliminating the need for aggregation. They also show a novel way to estimate the predictive uncertainty for image segmentation using their approach.

2.3. Uncertainty Estimation in Medical Imaging

Uncertainty quantification enhances model interpretability and safety in medical AI. Gal and Ghahramani (2016) show that applying Monte Carlo Dropout (MCD) at test time enables estimation of epistemic uncertainty in deep learning models. Lakshminarayanan et al. (2017) similarly show that deep ensembles provide well-calibrated predictive uncertainty estimates. Baumgartner et al. (2019) introduce *PHiSeg*, which uses variational autoencoders to model segmentation uncertainty for medical images. Zhang et al. (2023) study various

uncertainty quantification techniques within federated learning frameworks and show how heterogeneous data affects the calibration and reliability of uncertainty estimates. Judge et al. (2022) use contrastive learning and anatomical priors to generate well-calibrated and anatomically consistent uncertainty maps for medical image segmentation. Koutsoubis et al. (2024) provided a comprehensive review of federated learning and uncertainty estimation methods specific to medical imaging.

3. Methodology

We follow a training approach similar to the methodology used for multi-task segmentation as described in Chen et al. (2019), Ulrich et al. (2023), and Tölle et al. (2024). We retain a common model backbone for all clients, and allocate a separate segmentation head to each client, with output channels corresponding to the labels present in their dataset. We run the training phase of the experiments in three settings. 1) In the **centralized** setting, the common backbone is jointly trained using the data from all clients. For each batch, the loss is propagated through the segmentation head corresponding to the dataset the batch was drawn from. This setting typically serves as an upper bound for model performance. 2) In the **federated** setting, each client locally trains the common backbone along with its own segmentation head, and shares the updated backbone with the central server at the end of each communication round. The server aggregates the client models with a chosen aggregation scheme and sends the global backbone to the clients. 3) In the **standalone** setting, each client trains a model independently on its own dataset. This setting is used to assess how well a client-specific model generalizes to data from other clients.

Inference is primarily performed by averaging the logits obtained from each segmentation head. However, one of the key insights from FUNAvg is that predictive performance can be significantly improved by reweighting the background channel of the predicted logits using an uncertainty estimate. This adjustment accounts for the tendency to overestimate the background class during aggregation. They argue that aggregation from multiple heads can overestimate the background class, since the target class may not be present in all segmentation heads. They used MC Dropout as their predictive uncertainty estimation technique. In our work, we retain this uncertainty-weighted averaging scheme. However, we employ a different method for quantifying uncertainty, as described in Section 3.3.

We propose novel enhancements to the standard federated learning pipeline at three stages: a) local client training, b) server aggregation, and c) client inference. In the local client training stage, we introduce an online variance estimation technique for constructing the client parameter distribution. This distribution is sent to the server, which uses an inverse variance aggregation scheme to obtain the global model parameters. Finally, during inference, we employ a sampling strategy to estimate the predictive uncertainty over the test samples. To describe the problem setup, we define the standard variables used in federated learning. Let N be the total number of participating clients and R the total number of federated rounds. Each client performs local training for T mini-batch SGD steps. Let n_i denote the number of data samples held by client i.

3.1. Parameter Distribution Estimation

Each client trains its local model using mini-batch stochastic gradient descent (SGD) and estimates parameter uncertainty by tracking the parameter or gradient statistics across SGD iterations. Let t represent the current mini-batch SGD step, $\boldsymbol{\theta}_t$ the model parameters at step t, and $\boldsymbol{g}_t$ the corresponding gradient. The client computes the parameter update using the standard mini-batch SGD update with η as the learning rate.

$$\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} - \eta \boldsymbol{g}_t \tag{1}$$

We propose two different approaches for estimating the parameter uncertainty:
Gradient-based estimation. In this approach, we estimate the parameter uncertainty by first estimating the gradient variance and then accumulating them to construct the parameter variance. To reduce the memory overhead from storing all T gradient vectors, we compute the gradient mean and variance in an online fashion using Welford's algorithm (Welford (1962)), as shown below.

$$\begin{aligned} \bar{\boldsymbol{g}}_t &= \bar{\boldsymbol{g}}_{t-1} + \frac{\boldsymbol{g}_t - \bar{\boldsymbol{g}}_{t-1}}{t} \\ \boldsymbol{M}_{2,t} &= \boldsymbol{M}_{2,t-1} + (\boldsymbol{g}_t - \bar{\boldsymbol{g}}_{t-1})(\boldsymbol{g}_t - \bar{\boldsymbol{g}}_t) \\ \boldsymbol{\sigma}^2_{g,T} &= \frac{\boldsymbol{M}_{2,T}}{T} \end{aligned} \tag{2}$$

$\boldsymbol{M}_{2,t}$ is the intermediate sum of squared errors used to compute the gradient variance at the end of T steps, $\boldsymbol{\sigma}^2_{g,T}$. The estimated gradient variance is added to the previous estimate of the total parameter variance $\boldsymbol{\sigma}^2_{\theta,0}$ to get the new estimated parameter variance:

$$\boldsymbol{\sigma}^2_{\theta,T} = \boldsymbol{\sigma}^2_{\theta,0} + T\eta^2 \boldsymbol{\sigma}^2_{g,T} \tag{3}$$

The above equation makes a simplifying assumption of independence of gradient updates after each round to avoid computing the full gradient covariance. We refer to this variant of the algorithm as 'FIVA-G' in the upcoming sections.
Parameter-based estimation. In this variant, we directly estimate the parameter variance by tracking the running parameter mean and variance using Welford's algorithm as shown below.

$$\begin{aligned} \bar{\boldsymbol{\theta}}_t &= \bar{\boldsymbol{\theta}}_{t-1} + \frac{\boldsymbol{\theta}_t - \bar{\boldsymbol{\theta}}_{t-1}}{t} \\ \boldsymbol{P}_{2,t} &= \boldsymbol{P}_{2,t-1} + (\boldsymbol{\theta}_t - \bar{\boldsymbol{\theta}}_{t-1})(\boldsymbol{\theta}_t - \bar{\boldsymbol{\theta}}_t) \\ \boldsymbol{\sigma}^2_{\theta,T} &= \frac{\boldsymbol{P}_{2,T}}{T} \end{aligned} \tag{4}$$

$\boldsymbol{P}_{2,t}$ is the sum of squared errors for the parameters similar to $\boldsymbol{M}_{2,t}$. We refer to the parameter-based estimation variant of the overall algorithm as 'FIVA-P' in the upcoming sections.
The model parameters obtained after T mini-batch SGD steps represent the parameters for client i in global federation round r, denoted as $\boldsymbol{\theta}_{i,r} = \boldsymbol{\theta}_T$. Similarly, the parameter variance for client i at round r is denoted as $\boldsymbol{\sigma}^2_{\theta,i,r} = \boldsymbol{\sigma}^2_{\theta,T}$. For readability, we drop the θ subscript and represent the parameter variance for client i in round r as $\boldsymbol{\sigma}^2_{i,r}$ in the following sections.

3.2. Inverse Variance Aggregation

For a given federation round r, each client i transmits its parameter update and variance $(\boldsymbol{\theta}_{i,r}, \boldsymbol{\sigma}^2_{i,r})$ to the central server. The server then aggregates the received parameters using a weighted averaging scheme to obtain the global parameter vector and its variance, denoted by $(\boldsymbol{\theta}_{global,r}, \boldsymbol{\sigma}^2_{global,r})$. Typically, the relative client sample sizes $\hat{n_i} = n_i / \sum^N n_i$ are used as weights for averaging (McMahan et al. (2017)). We retain the relative sample size $\hat{n_i}$ as a scaling factor and incorporate the inverse of the parameter variance (i.e., the parameter precision), yielding the client-specific weight $\boldsymbol{c}_i$. This approach is inspired by the inverse-variance aggregation scheme widely used in statistical meta-analysis, where it has been shown to yield optimal weights for combining independent random variables (Borenstein (2013)). All client models are initialized using He initialization (He et al. (2015)), with parameter variances uniformly initialized to one for both the clients and the server. The overall aggregation scheme for round r is given as follows.

$$\boldsymbol{c}_{i,r} = \frac{\hat{n_i}}{\boldsymbol{\sigma}^2_{i,r}} \quad , \quad \hat{n_i} = \frac{n_i}{\sum_{i=1}^N n_i} \tag{5}$$

$$\boldsymbol{\theta}_{global,r} = \frac{\sum_{i=1}^N \boldsymbol{c}_{i,r} \boldsymbol{\theta}_{i,r}}{\sum_{i=1}^N \boldsymbol{c}_{i,r}} \tag{6}$$

$$\boldsymbol{\sigma}^2_{global,r} = \left(\lambda \cdot \frac{1}{\boldsymbol{\sigma}^2_{global,r-1}} + \sum_{i=1}^N \frac{\hat{n_i}}{\boldsymbol{\sigma}^2_{i,r}} \right)^{-1} \tag{7}$$

$$\boldsymbol{\theta}_{i,r+1} = \boldsymbol{\theta}_{global,r}, \quad \boldsymbol{\sigma}^2_{i,r+1} = \boldsymbol{\sigma}^2_{global,r} \quad \forall i \in \{1, \ldots, N\} \tag{8}$$

Variances are aggregated in a Bayesian update setting. To stabilize the updates, we incorporate the previous round's global variance scaled by a forgetting factor λ. We choose $\lambda = 0.95$ in all our experiments.

3.3. Inference with Uncertainty Estimation

The central insight of FUNAvg is to leverage the predictive uncertainty for performance improvement at the inference stage. Instead of relying on Monte Carlo Dropout to obtain model samples, we leverage the global parameter variances learned during federated training. We treat the global model as a Gaussian-distributed parameter vector and sample from this distribution by adding a perturbation $\boldsymbol{\epsilon}$ drawn from $\mathcal{N}(\mathbf{0}, \boldsymbol{\sigma}^2_{global,R})$, yielding model samples $\boldsymbol{\theta}_k$.

$$\boldsymbol{\theta}_k = \boldsymbol{\theta}_{global,R} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\sigma}^2_{global,R}) \tag{9}$$

We perform K stochastic forward passes using the sampled models $\boldsymbol{\theta}_k$, computing class probabilities $\hat{\boldsymbol{p}}_k$ via the softmax function. These are then averaged to obtain the final class probabilities $\bar{\boldsymbol{p}}$ for each pixel, for each segmentation head.

$$\hat{\boldsymbol{p}}_k = \text{Softmax}(f^{\theta_k}(\mathbf{x})) \quad ; \quad \bar{\boldsymbol{p}} = \frac{1}{K} \sum_{k=1}^K \hat{\boldsymbol{p}}_k \tag{10}$$

Table 1: Overview of the abdominal CT datasets used in our study. We use five datasets to serve as training clients: TotalSegmentator (TS) Wasserthal et al. (2023), Liver Tumor Segmentation (LiTS) Bilic et al. (2023), Beyond The Cranial Vault (BTCV) Landman et al. (2015), AbdomenCT-1k (A1k) Ma et al. (2022), and Learn2Reg (L2R) Xu et al. (2016). AMOS Ji et al. (2022) is used as the hold-out test client. Each dataset has a different number of samples and labeled organs. The test client includes all organ classes present in the training datasets.

Dataset	# Samples	# Labels	DD	Eso	GB	LK	Li	Pan	RK	Spl	Sto	UB
TS	1139	7	✓	✓	✓	✓	✗	✗	✓	✗	✓	✓
LiTS	131	1	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗
BTCV	30	6	✗	✗	✗	✓	✓	✓	✓	✓	✓	✗
A1k	1000	3	✗	✗	✗	✗	✓	✓	✗	✓	✗	✗
L2R	30	8	✗	✓	✓	✓	✗	✓	✓	✓	✓	✗
AMOS	300	10	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Following the approach of Kwon et al. (2022) and Tölle et al. (2024), we estimate the predictive uncertainty after K sampling steps as:

$$\boldsymbol{u} = \underbrace{\frac{1}{K}\left(\sum_{k=1}^{K} \operatorname{diag}(\hat{\boldsymbol{p}}_k) - \hat{\boldsymbol{p}}^{\otimes 2}\right)}_{\text{aleatoric}} + \underbrace{\frac{1}{K}\sum_{k=1}^{K}(\hat{\boldsymbol{p}}_k - \bar{\boldsymbol{p}})^{\otimes 2}}_{\text{epistemic}} \tag{11}$$

where $\hat{\boldsymbol{p}}^{\otimes 2}$ denotes the outer product $\hat{\boldsymbol{p}}\hat{\boldsymbol{p}}^T$. This decomposition separates the data-dependent (aleatoric) and model-based (epistemic) uncertainty. Finally, outputs across all segmentation heads are aggregated. We reweigh the predictions from the background class using the estimated uncertainty $\boldsymbol{u}$, as proposed in FUNAvg. We refer the reader to Tölle et al. (2024) for further details on this step. The general methodology for our proposed approach is illustrated in Figure 1.

3.4. Experimental Setup

We use a total of six publicly available abdominal CT datasets, each containing a distinct subset of ten foreground organ labels[1]. We treat each dataset as an individual client in a federated learning setup, with five clients participating in training and one serving as the hold-out test client. Dataset details and label availability are summarized in Table 1. The label distribution across clients is visualized in Figure 2. Each training dataset is further split into an 80-20 train-validation split.

1. Organ labels (abbr.): Duodenum (DD), Esophagus (Eso), Gall Bladder (GB), Left Kidney (LK), Liver (Liv), Pancreas (Pan), Right Kidney (RK), Spleen (Spl), Stomach (Sto), Urinary Bladder (UB).

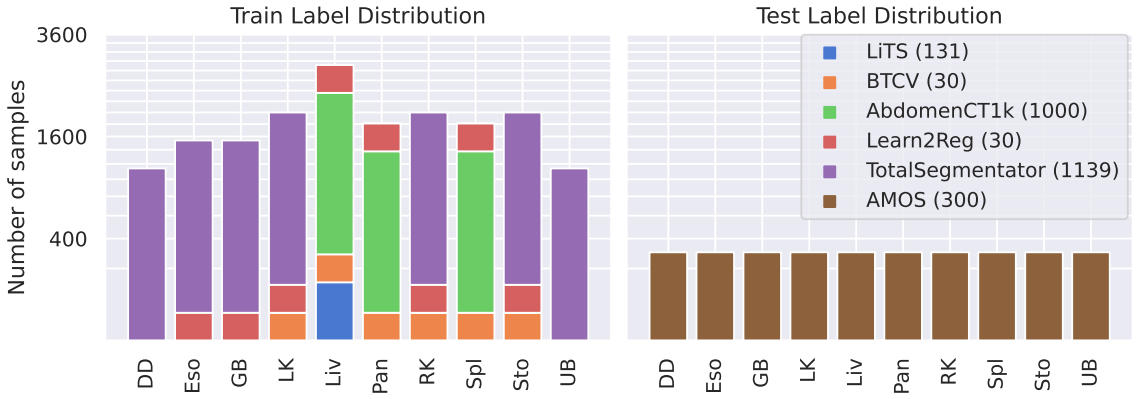


Figure 2: Label distribution across the six client datasets from Table 1, illustrating the variation in organ classes and sample sizes. Each training client was assigned a subset of available organ labels such that the test client (AMOS) contains the union of all labels used during training.

We use the nnUNetv2 framework (Isensee et al. (2021)) in its 2D configuration as the segmentation model. All preprocessing and architecture settings follow nnUNet's built-in auto-configuration. The model architecture generated for the TotalSegmentator dataset is used uniformly across all clients. A fixed patch size of 256×256 is used for all clients. We train the federated models for 1500 rounds using SGD with momentum=0.99 and Nesterov acceleration, and a PolyLR learning rate scheduler with initial learning rate=0.01 and the polynomial exponent=0.9. The loss function is a sum of Cross-Entropy Loss and Dice Loss. We evaluated three configurations: standalone (per-client), centralized, and federated. For the federated setup, we first compare both variants of our proposed algorithm, FIVA-G and FIVA-P (without uncertainty-weighted inference), against FedAvg. This establishes a baseline to observe the effects of inverse-variance aggregation alone. We then evaluated both proposed variants with uncertainty-weighted inference (denoted FIVA-G+UN and FIVA-P+UN) and compared them with FUNAvg. Plotting code was adapted from Tölle et al. (2024).

4. Results

Table 2 summarizes the performance of our method compared to the baselines across training clients and the held-out AMOS test client. Both variants, FIVA-G and FIVA-P, improve on FedAvg for all but one client, demonstrating that inverse-variance aggregation offers a standalone improvement during training. The parameter-based variant of our proposed method, FIVA-P, improves over FedAvg by nearly 11 percentage points in mean Dice score (54.16 vs. 65.49). When combined with the sampling-based uncertainty weighting strategy (FIVA-G+UN and FIVA-P+UN), we observe similar improvements as compared to FUNAvg, which uses MC Dropout-based uncertainty. We also observe that uncertainty-

Table 2: DICE Scores ($mean_{std.}$% from runs over five different sets of image slices) evaluated on the hold-out test sets of each of the training clients and on the hold-out test client (AMOS). The top two rows show the standalone training performance on each client's own test set (Same-client) and on the test sets of other clients (Cross-client). Centralized training is performed by jointly learning a model from all clients' data. Baseline results are first compared without uncertainty-weighted inference (FedAvg vs. FIVA-G/FIVA-P), where the best results are underlined. The next group of results incorporates uncertainty-weighted inference (FUNAvg vs. FIVA-G+UN/FIVA-P+UN), with the best results shown in bold.

Method	**TS**	**LiTS**	**BTCV**	**A1k**	**L2R**	**AMOS**	**Mean**
Same-client	$87.99_{0.27}$	$90.10_{0.05}$	$72.50_{0.36}$	$90.90_{0.03}$	$70.82_{0.30}$	$85.80_{0.33}$	$83.02_{0.09}$
Cross-client	$42.89_{1.04}$	$65.08_{0.16}$	$64.37_{0.41}$	$29.07_{0.23}$	$29.33_{0.25}$	$59.39_{0.24}$	$48.35_{0.21}$
Centralized	$72.42_{0.22}$	$89.68_{0.07}$	$62.44_{0.54}$	$65.55_{0.16}$	$67.03_{0.36}$	$46.77_{0.51}$	$67.31_{0.10}$
FedAvg	$\underline{90.72_{0.13}}$	$57.98_{0.26}$	$41.81_{0.59}$	$48.54_{0.68}$	$48.67_{1.30}$	$37.23_{0.87}$	$54.16_{0.18}$
FIVA-G	$79.40_{0.18}$	$70.67_{0.15}$	$\underline{59.33_{0.69}}$	$60.51_{0.08}$	$53.48_{0.33}$	$\underline{53.03_{0.77}}$	$62.74_{0.12}$
FIVA-P	$88.02_{0.19}$	$\underline{83.40_{0.12}}$	$55.92_{0.41}$	$\underline{61.70_{0.15}}$	$\underline{56.57_{0.25}}$	$47.36_{0.89}$	$\underline{65.49_{0.19}}$
FUNAvg	$\mathbf{89.02_{0.08}}$	$80.38_{0.29}$	$54.57_{0.18}$	$53.23_{0.74}$	$52.83_{0.90}$	$53.93_{0.60}$	$63.99_{0.20}$
FIVA-G+UN	$78.08_{0.20}$	$81.50_{0.15}$	$60.76_{0.72}$	$60.66_{0.07}$	$54.33_{0.22}$	$\mathbf{55.53_{0.68}}$	$65.14_{0.13}$
FIVA-P+UN	$87.03_{0.17}$	$\mathbf{86.45_{0.10}}$	$\mathbf{56.97_{0.53}}$	$\mathbf{61.92_{0.14}}$	$\mathbf{57.03_{0.28}}$	$49.15_{0.82}$	$\mathbf{66.42_{0.21}}$

weighted inference generally improves performance over naive inference, consistent with the findings of Tölle et al. (2024).

The qualitative results in Figure 3 further illustrate the benefit of uncertainty-driven inference. FIVA+UN captures subtle boundary ambiguities that are overlooked by the MC dropout-based method, especially in anatomically complex regions or small organs. Additional qualitative results on the AMOS test set are shown in Figure 5 in Appendix A. Reliability diagrams (Figure 4) show that FIVA models exhibit better-calibrated predictions with lower Expected Calibration Error (ECE) compared to FedAvg and FUNAvg. These results establish FIVA's accuracy and calibration reliability in federated segmentation under real-world heterogeneity.

5. Discussion

The results we obtained highlight some important takeaways. First, we propose a sampling-based method for modeling uncertainty that improves both the training and inference stages in real-world federated learning. Second, we observe consistent improvements in Dice scores when using sampling-based uncertainty estimates compared to MC Dropout. Although MC Dropout is a widely used approach for predictive uncertainty, it reduces model capacity during inference by randomly disabling neurons, potentially removing those most salient for

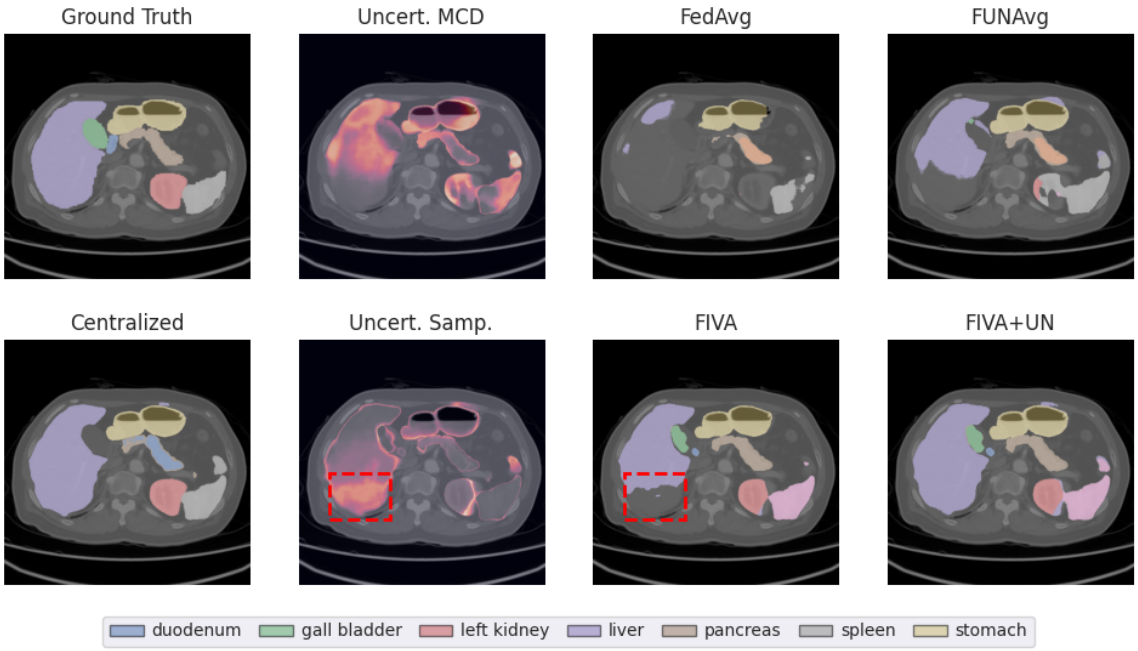
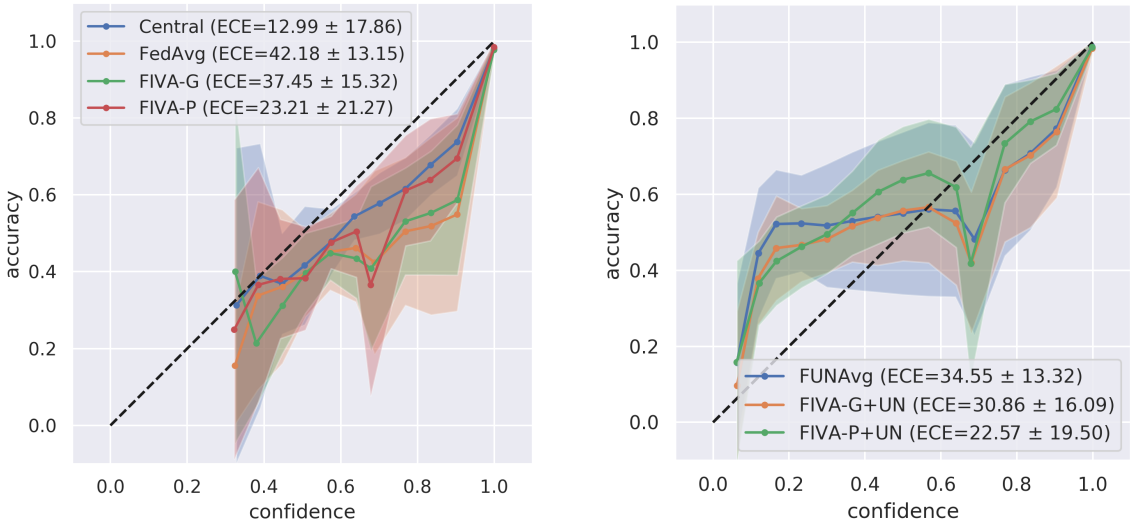


Figure 3: Qualitative results on the AMOS test set highlight the core operating principle of uncertainty weighting. Sampling-based uncertainty (Uncert. Samp.) captures uncertainties near organ boundaries and in partially segmented regions. In contrast, MC Dropout-based uncertainty (Uncert. MCD) fails to capture the lower regions of the liver (highlighted in red). The last column shows false background predictions suppressed in areas of high uncertainty.

segmentation. In contrast, our sampling-based approach preserves full model capacity as long as parameter variances remain bounded. Third, qualitative results show that the uncertainty estimates derived from our sampling method correctly emphasize organ boundaries, regions that naturally exhibit high uncertainty. Finally, our results reinforce earlier findings from FUNAvg: uncertainty-weighted logit averaging at inference time helps suppress false-positive background predictions, improving segmentation accuracy.
Although the proposed approach shows an improvement in segmentation performance, it incurs an additional computation cost. Accounting for the added overhead of estimating the variances and uncertainties would be crucial for deployment in compute-limited environments. We briefly discuss this overhead below.

Time complexity. Let the model have M trainable parameters, and assume each input data point is D-dimensional. For a total of T mini-batch SGD updates using batches of size B, the computational time complexity of standard training is $\mathcal{O}(TMBD)$, accounting for both forward and backward passes. Our proposed method, FIVA, introduces two additional vectorized operations per update: one for updating the running mean of gradients/parameters and another for computing the sum of squared errors, both using Welford's algorithm. These operations introduce an added computational cost of $\mathcal{O}(TM)$. Given that



(a) Model calibration **without** uncertainty weighted inference.

(b) Model calibration **with** uncertainty weighted inference.

Figure 4: Reliability diagrams for the tested methods with the label-wise aggregated ECE (mean ± std). FIVA and FIVA+UN show improvement in calibration error over the baselines. Notably, uncertainty-weighted inference (b) leads to a more balanced model calibration compared to naive averaging (a).

$M \ll MBD$, and noting that additional computations are parallelizable with a GPU, the asymptotic complexity remains $\mathcal{O}(TMBD)$. Therefore, the runtime overhead introduced by FIVA is negligible in practical settings, particularly when training deep models on large datasets.

Space complexity. In standard training, each client typically maintains three tensors of size $\mathcal{O}(M)$: model parameters, gradient estimates, and optimizer states such as momentum, leading to a total space complexity of $\mathcal{O}(3M)$. FIVA augments this with three additional $\mathcal{O}(M)$ tensors per client to store the running means, the sum of squared errors, and the variances of the gradients/parameters. This results in a total space complexity of $\mathcal{O}(6M)$ per client. Although FIVA requires double the memory compared to standard training, it is significantly more memory efficient than storing full gradient/parameter histories that require $\mathcal{O}(TM)$ space, where T is the number of training iterations. This is due to Welford's algorithm, which uses constant memory per parameter for the variance estimation.

Limitations. While the inverse-variance aggregation scheme is well grounded in statistical meta-analysis, its use in machine learning remains relatively unexplored. For this method to function reliably, parameter variances must remain upper- and lower-bounded throughout the training. In low-data regimes, variance estimates can become unreliable, potentially destabilizing training. Although our approach improves upon existing baselines, a significant gap remains between federated and standalone performance, limiting immedi-

ate clinical applicability. Furthermore, calibration errors, while improved, are still relatively high compared to typical values reported in classical machine learning settings.
Overall, this work bridges two important paradigms in machine learning for healthcare: uncertainty estimation and federated learning, demonstrating how insights from one can meaningfully improve the other. We hope that future research will focus on variance stabilization techniques, improved calibration under heterogeneity, and more efficient approximations of uncertainty to make these methods viable for broader adoption.

References

Mothilal Asokan, Joseph Geo Benjamin, Mohammad Yaqub, and Karthik Nandakumar. A federated learning-friendly approach for parameter-efficient fine-tuning of sam in 3d segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 226–235. Springer, 2024.

Christian F. Baumgartner, Kerem C. Tezcan, Krishna Chaitanya, Andreas M. Hötker, Urs J. Muehlematter, Khoschy Schawkat, Anton S. Becker, Olivio Donati, and Ender Konukoglu. PHiSeg: Capturing Uncertainty in Medical Image Segmentation, July 2019.

Tobias Bernecker, Annette Peters, Christopher L Schlett, Fabian Bamberg, Fabian Theis, Daniel Rueckert, Jakob Weiß, and Shadi Albarqouni. Fednorm: Modality-based normalization in federated learning for multi-modal liver segmentation. *arXiv preprint arXiv:2205.11096*, 2022.

Patrick Bilic, Patrick Christ, Hongwei Bran Li, Eugene Vorontsov, Avi Ben-Cohen, Georgios Kaissis, Adi Szeskin, Colin Jacobs, Gabriel Efrain Humpire Mamani, Gabriel Chartrand, Fabian Lohöfer, Julian Walter Holch, Wieland Sommer, Felix Hofmann, Alexandre Hostettler, Naama Lev-Cohain, Michal Drozdzal, Michal Marianne Amitai, Refael Vivanti, Jacob Sosna, Ivan Ezhov, Anjany Sekuboyina, Fernando Navarro, Florian Kofler, Johannes C. Paetzold, Suprosanna Shit, Xiaobin Hu, Jana Lipková, Markus Rempfler, Marie Piraud, Jan Kirschke, Benedikt Wiestler, Zhiheng Zhang, Christian Hülsemeyer, Marcel Beetz, Florian Ettlinger, Michela Antonelli, Woong Bae, Míriam Bellver, Lei Bi, Hao Chen, Grzegorz Chlebus, Erik B. Dam, Qi Dou, Chi-Wing Fu, Bogdan Georgescu, Xavier Giró i Nieto, Felix Gruen, Xu Han, Pheng-Ann Heng, Jürgen Hesser, Jan Hendrik Moltz, Christian Igel, Fabian Isensee, Paul Jäger, Fucang Jia, Krishna Chaitanya Kaluva, Mahendra Khened, Ildoo Kim, Jae-Hun Kim, Sungwoong Kim, Simon Kohl, Tomasz Konopczynski, Avinash Kori, Ganapathy Krishnamurthi, Fan Li, Hongchao Li, Junbo Li, Xiaomeng Li, John Lowengrub, Jun Ma, Klaus Maier-Hein, Kevis-Kokitsi Maninis, Hans Meine, Dorit Merhof, Akshay Pai, Mathias Perslev, Jens Petersen, Jordi Pont-Tuset, Jin Qi, Xiaojuan Qi, Oliver Rippel, Karsten Roth, Ignacio Sarasua, Andrea Schenk, Zengming Shen, Jordi Torres, Christian Wachinger, Chunliang Wang, Leon Weninger, Jianrong Wu, Daguang Xu, Xiaoping Yang, Simon Chun-Ho Yu, Yading Yuan, Miao Yue, Liping Zhang, Jorge Cardoso, Spyridon Bakas, Rickmer Braren, Volker Heinemann, Christopher Pal, An Tang, Samuel Kadoury, Luc Soler, Bram van Ginneken, Hayit Greenspan, Leo Joskowicz, and Bjoern Menze. The liver tumor segmentation benchmark (lits). *Medical Image Analysis*, 84:

102680, 2023. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2022.102680. URL https://www.sciencedirect.com/science/article/pii/S1361841522003085.

Michael Borenstein, editor. *Introduction to Meta-Analysis*. Wiley, Chichester, nachdr. edition, 2013. ISBN 978-0-470-05724-7.

Victor Ion Butoi, Jose Javier Gonzalez Ortiz, Tianyu Ma, Mert R Sabuncu, John Guttag, and Adrian V Dalca. Universeg: Universal medical image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21438–21451, 2023.

Sihong Chen, Kai Ma, and Yefeng Zheng. Med3d: Transfer learning for 3d medical image analysis. *arXiv preprint arXiv:1904.00625*, 2019.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.

Yunhe Gao, Zhuowei Li, Di Liu, Mu Zhou, Shaoting Zhang, and Dimitris N. Metaxas. Training Like a Medical Resident: Context-Prior Learning Toward Universal Medical Image Segmentation, April 2024.

Joachim Hartung. *Statistical Meta-Analysis with Applications*. Wiley, Hoboken, NJ, 2008. ISBN 978-0-470-29089-7 978-0-470-38633-0.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

Fabian Isensee, Paul F. Jaeger, Simon A. A. 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, February 2021. ISSN 1548-7091, 1548-7105. doi: 10.1038/s41592-020-01008-z.

Nyaz Jalal, Małgorzata Śliwińska, Wadim Wojciechowski, Iwona Kucybała, Miłosz Rozynek, Kamil Krupa, Patrycja Matusik, Jarosław Jarczewski, and Zbisław Tabor. Evaluating Uncertainty Quantification in Medical Image Segmentation: A Multi-Dataset, Multi-Algorithm Study. *Applied Sciences*, 14(21):10020, November 2024. ISSN 2076-3417. doi: 10.3390/app142110020.

Yuanfeng Ji, Haotian Bai, Chongjian Ge, Jie Yang, Ye Zhu, Ruimao Zhang, Zhen Li, Lingyan Zhanng, Wanling Ma, Xiang Wan, and Ping Luo. AMOS: A Large-Scale Abdominal Multi-Organ Benchmark for Versatile Medical Image Segmentation. *Advances in Neural Information Processing Systems*, 35:36722–36732, December 2022.

Meirui Jiang, Holger R Roth, Wenqi Li, Dong Yang, Can Zhao, Vishwesh Nath, Daguang Xu, Qi Dou, and Ziyue Xu. Fair Federated Medical Image Segmentation via Client Contribution Estimation. *CVPR*, 2023.

Thierry Judge, Olivier Bernard, Mihaela Porumb, Agis Chartsias, Arian Beqiri, and Pierre-Marc Jodoin. CRISP - Reliable Uncertainty Estimation for Medical Image Segmentation, June 2022.

Alex Kendall and Yarin Gal. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?, October 2017.

Nikolas Koutsoubis, Yasin Yilmaz, Ravi P Ramachandran, Matthew Schabath, and Ghulam Rasool. Privacy preserving federated learning in medical imaging with uncertainty estimation. *arXiv preprint arXiv:2406.12815*, 2024.

Yongchan Kwon, Joong-Ho Won, Beom Joon Kim, and Myunghee Cho Paik. Uncertainty quantification using Bayesian neural networks in classification: Application to ischemic stroke lesion segmentation. In *Medical Imaging with Deep Learning*, July 2022.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Bennett Landman, Zhoubing Xu, Juan Igelsias, Martin Styner, Thomas Langerak, and Arno Klein. Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In *Proc. MICCAI multi-atlas labeling beyond cranial vault—workshop challenge*, volume 5, page 12. Munich, Germany, 2015.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated Optimization in Heterogeneous Networks. *Proceedings of Machine Learning and Systems*, 2:429–450, March 2020.

Jun Ma, Yao Zhang, Song Gu, Cheng Zhu, Cheng Ge, Yichi Zhang, Xingle An, Congcong Wang, Qiyuan Wang, Xin Liu, Shucheng Cao, Qi Zhang, Shangqing Liu, Yunpeng Wang, Yuhui Li, Jian He, and Xiaoping Yang. AbdomenCT-1K: Is Abdominal Organ Segmentation a Solved Problem? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6695–6714, October 2022. ISSN 0162-8828, 2160-9292, 1939-3539. doi: 10.1109/TPAMI.2021.3100536.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, pages 1273–1282. PMLR, April 2017.

Ashish Rauniyar, Desta Haileselassie Hagos, Debesh Jha, Jan Erik Håkegård, Ulas Bagci, Danda B. Rawat, and Vladimir Vlassov. Federated Learning for Medical Applications: A Taxonomy, Current Trends, Challenges, and Future Research Directions. *IEEE Internet of Things Journal*, 11(5):7374–7398, March 2024. ISSN 2327-4662, 2372-2541. doi: 10.1109/JIOT.2023.3329061.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted*

intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18, pages 234–241. Springer, 2015.

Malte Tölle, Fernando Navarro, Sebastian Eble, Ivo Wolf, Bjoern Menze, and Sandy Engelhardt. FUNAvg: Federated Uncertainty Weighted Averaging for Datasets with Diverse Labels, July 2024.

Constantin Ulrich, Fabian Isensee, Tassilo Wald, Maximilian Zenk, Michael Baumgartner, and Klaus H. Maier-Hein. MultiTalent: A Multi-Dataset Approach to Medical Image Segmentation. volume 14222, pages 648–658. 2023. doi: 10.1007/978-3-031-43898-1_62.

Jakob Wasserthal, Hanns-Christian Breit, Manfred T. Meyer, Maurice Pradella, Daniel Hinck, Alexander W. Sauter, Tobias Heye, Daniel T. Boll, Joshy Cyriac, Shan Yang, Michael Bach, and Martin Segeroth. TotalSegmentator: Robust Segmentation of 104 Anatomic Structures in CT Images. *Radiology: Artificial Intelligence*, 5(5):e230024, September 2023. ISSN 2638-6100. doi: 10.1148/ryai.230024.

B. P. Welford. Note on a method for calculating corrected sums of squares and products. *Technometrics : a journal of statistics for the physical, chemical, and engineering sciences*, 4(3):419–420, 1962. doi: 10.1080/00401706.1962.10490022.

Xuanang Xu, Hannah H. Deng, Jamie Gateno, and Pingkun Yan. Federated Multi-Organ Segmentation With Inconsistent Labels. *IEEE Transactions on Medical Imaging*, 42(10): 2948–2960, October 2023. ISSN 0278-0062, 1558-254X. doi: 10.1109/TMI.2023.3270140.

Xuanang Xu, Hannah H Deng, Tianyi Chen, Tianshu Kuang, Joshua C Barber, Daeseung Kim, Jaime Gateno, James J Xia, and Pingkun Yan. Federated cross learning for medical image segmentation. In *Medical Imaging with Deep Learning*, pages 1441–1452. PMLR, 2024.

Zhoubing Xu, Christopher P. Lee, Mattias P. Heinrich, Marc Modat, Daniel Rueckert, Sebastien Ourselin, Richard G. Abramson, and Bennett A. Landman. Evaluation of Six Registration Methods for the Human Abdomen on Clinically Acquired CT. *IEEE Transactions on Biomedical Engineering*, 63(8):1563–1572, August 2016. ISSN 0018-9294, 1558-2531. doi: 10.1109/TBME.2016.2574816.

Xu Zhang, Yinchuan Li, Wenpeng Li, Kaiyang Guo, and Yunfeng Shao. Personalized Federated Learning via Variational Bayesian Inference. In *Proceedings of the 39th International Conference on Machine Learning*, pages 26293–26310. PMLR, June 2022.

Yuwei Zhang, Abhirup Ghosh, Tong Xia, and Cecilia Mascolo. Uncertainty Quantification in Federated Learning for Heterogeneous Health Data. 2023.

Appendix A. Additional Qualitative Results

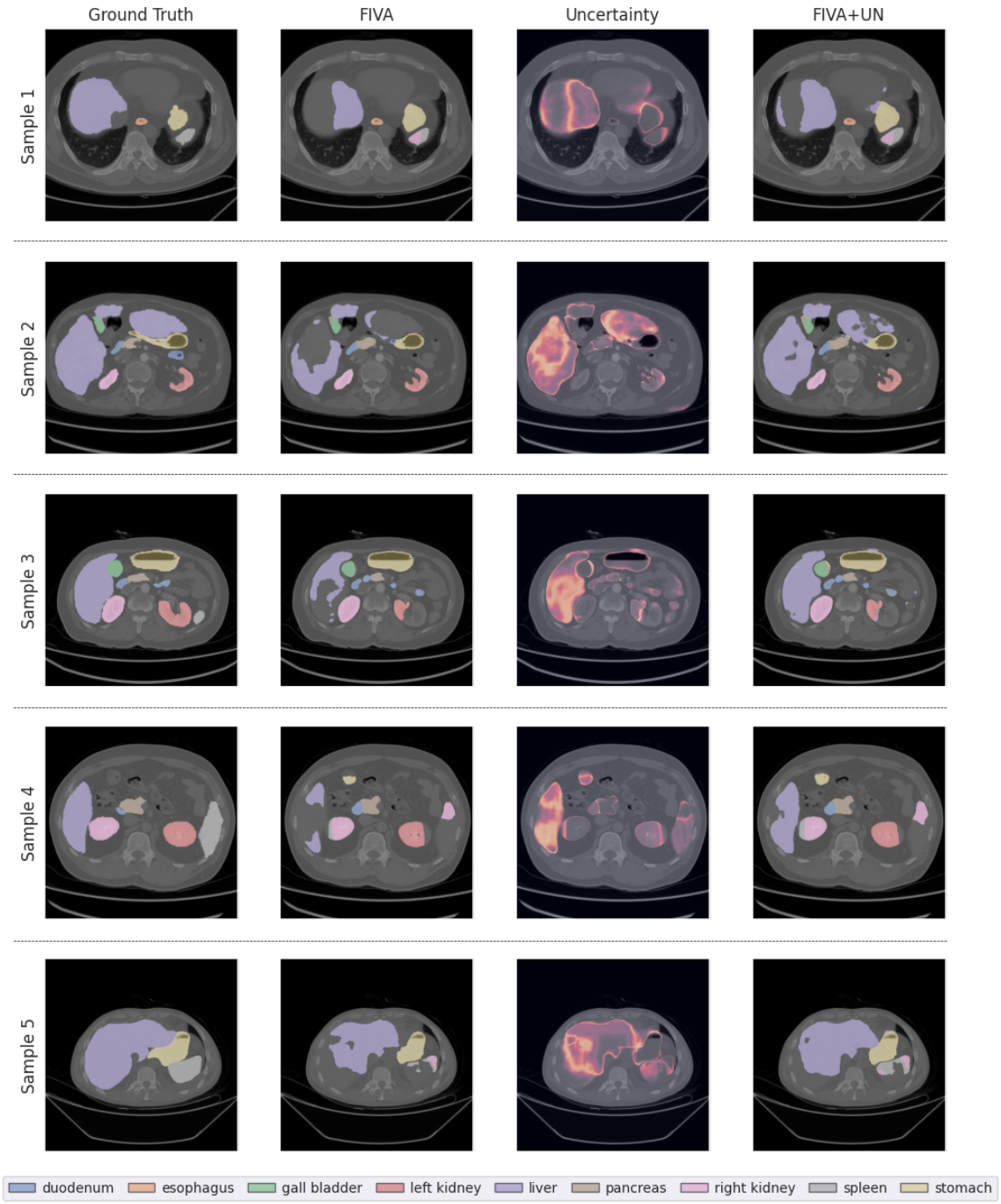


Figure 5: Additional qualitative results on different samples from the AMOS test set.