# FUNAvg: Federated Uncertainty Weighted Averaging for Datasets with Diverse Labels

Malte Tölle[1,2,3,4][0000−0002−0804−5794], Fernando Navarro[4][0000−0001−8906−9079], Sebastian Eble[1][0009−0000−3736−4316], Ivo Wolf[1,5][0000−0002−6519−6484], Bjoern Menze[4*][0000−0003−4136−5690], Sandy Engelhardt[1,2,3*][0000−0001−8816−7654]

[1] Department of Internal Medicine III, Heidelberg University Hospital, Germany
[2] Informatics for Life Institute, Ruprecht-Karls University Heidelberg, Germany
[3] DZHK (German Centre for Cardiovascular Research), partner site Heidelberg/Mannheim
[4] Department of Quantitative Biomedicine, University of Zurich, Switzerland
[5] Department of Computer Science, Mannheim University of Applied Sciences, Germany
malte.toelle@med.uni-heidelberg.de

**Abstract.** Federated learning is one popular paradigm to train a joint model in a distributed, privacy-preserving environment. But partial annotations pose an obstacle meaning that categories of labels are heterogeneous over clients. We propose to learn a joint backbone in a federated manner, while each site receives its own multi-label segmentation head. By using Bayesian techniques we observe that the different segmentation heads although only trained on the individual client's labels also learn information about the other labels not present at the respective site. This information is encoded in their predictive uncertainty. To obtain a final prediction we leverage this uncertainty and perform a weighted averaging of the ensemble of distributed segmentation heads, which allows us to segment "locally unknown" structures. With our method, which we refer to as FUNAvg, we are even on-par with the models trained and tested on the same dataset on average. The code is publicly available[1].

**Keywords:** Federated Learning · Bayesian Neural Networks · Partial Labels.

## 1 Introduction

The ability of deep neural networks to accurately segment anatomical structures promises precise quantitative analysis and clinical diagnosis, and has the potential to significantly improve medical decision making [24]. However, due to the required effort for accurately labelling medical images every institution only creates annotations for their particular research endeavour [22]. Despite being from

---

* These authors contributed equally.
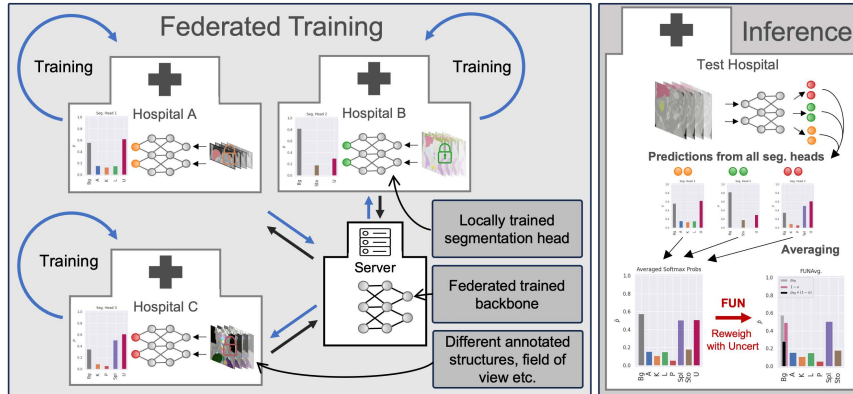[1] https://github.com/Cardio-AI/FUNAvg

Fig. 1: Our proposed training and inference scheme. During training each site optimizes its own segmentation head according to the number of present labels at the respective site. On the central server only the backbone is averaged in a federated fashion. During inference all segmentation heads are gathered and an average of the softmax probabilities is computed weighted by the number of sites the individual label was present. By utilizing the predictive uncertainty of the classifiers with FUNAvg the predictions can be improved, which is especially benefitial for underrepresented labels across the federated sites.

the same modality and capturing a joint field of view, the cumulative information is hardly exploited. This leads to partially annotated, distributed datasets in a sense that some structures might be segmented in one dataset (e.g. liver, kidney) and other structures in the other dataset (e.g., spleen, spine) that can not be gathered on a central server because of privacy constraints.

The standard workflow includes training a model on dataset A that has the structure of interest annotated and applying it to dataset B, which due to domain gaps leads to subpar performance. Recently, self-supervised methods have gained attention which use an unsupervised pre-training on a large corpus of data similar to dataset B, where the model is usually trained to minimize a reconstruction loss [7]. Federated Learning (FL) is one renowned method that reverts the common paradigm of central data storage to circumvent privacy constraints [20]. In FL the model is sent to each data holding institution and subsequently averaged on a central server during model training. Our method combines both approaches by jointly learning all annotated structures across a multitude of datasets while showing that not-annotated information in the data is learned as well and therefore can be explored for improving predictions during inference similar to self-supervised methods. The approach most related to ours, MultiTalent [24], also employs training multiple segmentation heads across different datasets but discards the possible information learned from not-annotated structures, partly because they do not employ a Bayesian approach [24].
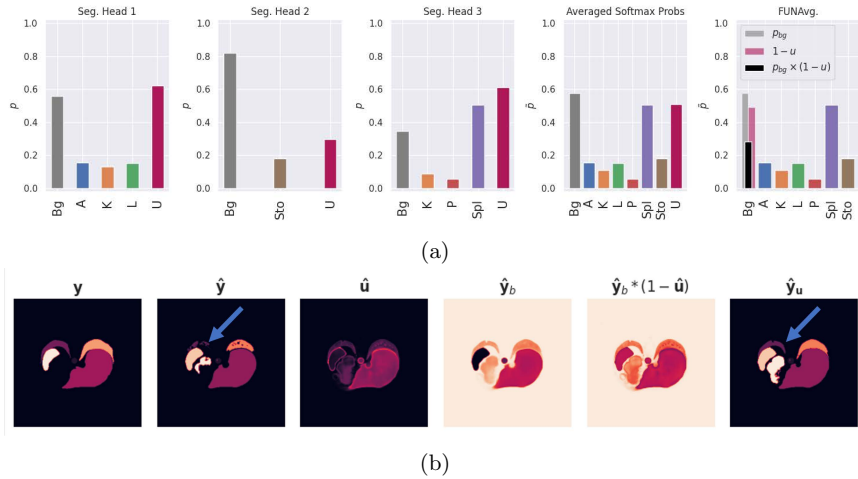
(a)



(b)

Fig. 2: Proposed federated uncertainty weighted averaging for (a) one pixel and (b) an entire image. (a) When averaging the softmax probabilities of sites 1-3 the final prediction would be "Background" (Bg). By reweighting the probability for background by the uncertainty (U) we obtain the right label of "Spleen" in this example. (b) After averaging the logits the final prediction $\hat{y}$ is fragmentary especially in the area of the lung in above example, while the lung is perfectly visible in the uncertainty estimation $\hat{u}$. We therefore multiply the background channel with the inverse of the uncertainty $\hat{y}_b \times (1 - \hat{u})$ to obtain $\hat{y}_u$.

We train one model in a federated setting across several partially labelled clients, which have different number of pre-segmented images and structures as well as annotation protocols. Our method trains a backbone in a federated manner across all clients, while each client obtains a separate segmentation head. The final prediction is then obtained by averaging in an ensemble-like manner. Suprisingly, also those structures are represented in the uncertainty maps for models trained at sites where the corresponding labels were not available. We show that this uncertainty can be leveraged to segment previously unknown structures at this location to improve performance especially for structures that are underrepresented in the training sets i.e. are only present at one or few clients.

## 2    Related Work

*Learning from Partially Annotated Datasets.* Lately, there has been an increasing amount of studies focusing on multi-organ segmentation using data that is partially labeled. One line of research leverages the inherent homogeneous anatomy of the human body in terms of shape, size, and locations of anatomical struc-
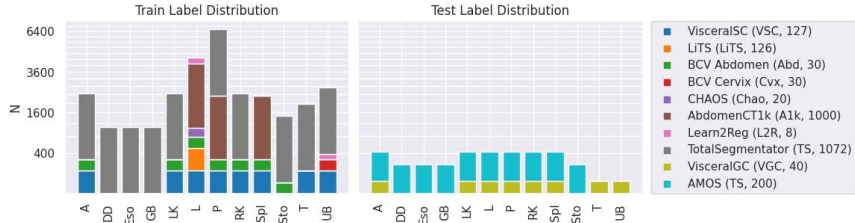
Fig. 3: Datasets used for training and testing and their respective label distribution. They differ in general quantity of training samples as well as number of annotated labels present. We used the following open data: Liver Tumor Segmentation (LiTS) [1], Beyond the Cranial Vault (BCV, Cervix and Abdomen) [15], Combined Healthy Abdominal Organ (CHAOS) [11], Learn2Reg [27], AbdomenCT-1k [19], Abdominal Multi-Organ Benchmark (AMOS) [9], and TotalSegmentator [25], and two in-house datasets termed VisceralGC and SC.

tures [21,28]. In another approach, the network is equipped with different encoders for each dataset in the learning procedure [26]. But this method falls short to account for different combinations of the various organs across all datasets present in the model training. Similar to equipping the model with different encoders multiple heads that share a custom backbone can be used [4]. In the approach most similar to ours called MultiTalent the network is equipped with different segmentation heads for each datasets in the learning procedure [24]. It is argued that this is largely needed due to different annotation protocols across the different datasets. However, this fails at leveraging unannotated organs across different datasets that enhances predictive quality due to larger amount of data samples. Further, we believe that varying annotation protocols represent some form of data dependent i.e. aleatoric uncertainty that can be captured with Bayesian techniques [5,12,13]. MultiTalent does not exploit ways of obtaining a final prediction by e.g. averaging of the trained segmentation heads [24].

*Bayesian Federated Learning.* In contrast to the point estimate used in frequentist deep learning, the Bayesian approach models the parameters with a distribution. Mathematically a prior distribution $p(\boldsymbol{\theta}|\alpha)$ is placed over the weights $\boldsymbol{\theta}$ of a neural network, governed by a hyperparameter $\alpha$. Our interest lies in the posterior after observing some data $\mathcal{D}$ $p(\boldsymbol{\theta}|\mathcal{D}, \alpha) = p(\mathcal{D}|\boldsymbol{\theta}, \alpha)p(\boldsymbol{\theta}|\alpha)/p(\mathcal{D})$. Unfortunately, this distribution is not tractable, but can be approximated with Bayesian inference techniques such as variational inference (VI).

In federated learning, numerous clients collaboratively train a unified global model, while adhering to data privacy as no data is shared with a central server managing the training process called federated averaging (FedAvg) [20]. For a given set of clients, let $\mathcal{D}_i = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}$, represent the data of client i and $\overline{\boldsymbol{\theta}}$ the weights of the global model. At the beginning of each training round the model

weights $\boldsymbol{\theta_i}$ of every client are initialized with the latest global model weights. Subsequently, each client performs stochastic gradient descent. After receiving the trained models from each client, FedAvg updates the global model by averaging the weights. This averaging is performed in proportion to the amount of data each client contributes, ensuring that clients with more data have a correspondingly greater influence on the final model $\overline{\boldsymbol{\theta}} \leftarrow \sum_i \frac{|n_i|}{n} \boldsymbol{\theta_i}$. The final model results from the iterative application of the two previously described equations.

The standard approach of federated averaging has been refined for frequentist model training in mutiple works (e.g., [10,17]). A Bayesian model capable of predicting uncertainty can be obtained by training in the frequentist manner on each client and subsequently by treating each client model as a sample from the posterior [2,18,23]. As FL includes a set of candidate client models it naturally lends itself towards model ensembling which is used in FedBE [3]. While especially model ensembles have received significant attention in the community the usage of full Bayesian neural networks with variational inference in the federated setting has been rarely explored.

## 3   Methods

*Uncertainty.* To reduce computational load we opt for the practical Bayesian technique of Monte Carlo (MC) dropout [5]. Using dropout during training and testing has been shown to approximate the true posterior, while not introducing more computational load to capture uncertainty. Uncertainty is divided into aleatoric (data dependent) and epistemic (model dependent) uncertainty. We follow [14] and define uncertainty as

$$\mathcal{U} = \underbrace{\frac{1}{T} \left( \sum_{t=1}^{T} \operatorname{diag}(\hat{\mathbf{p}}_t) - \hat{\mathbf{p}}^{\otimes 2} \right)}_{\text{aleatoric}} + \underbrace{\frac{1}{T} \sum_{t=1}^{T} (\hat{\mathbf{p}}_t - \overline{\mathbf{p}})^{\otimes 2}}_{\text{epistemic}} , \tag{1}$$

where T is the number of MC sampling steps, $\otimes$ denotes the outer product, $f$ the network with parameters $\hat{\theta}_t$ in step $t$ for input $x^*$, $\overline{\mathbf{p}} = 1/T \sum_{t=1}^{T} \hat{\mathbf{p}}_t$ and $\hat{\mathbf{p}}_t = \mathbf{p}(\hat{\boldsymbol{\theta}}_t) = \operatorname{Softmax}\{f^{\hat{\boldsymbol{\theta}}_t}(\mathbf{x}^*)\}$. The first part captures the inherent variability in the data that cannot be reduced even if more data were to be collected, while the second term captures the variance in the model's output that can potentially be reduced with more data.

Neural networks tend to produce overconfident predictions, meaning the predictive uncertainty is smaller than the predictive error, this is also known as miscalibration [16] and can be expressed with

$$\mathbb{E}_{\hat{p}} \left[ |\mathbb{P} \left( \hat{y} = y | \hat{p} = q \right) - q| \right] , \quad \forall q \in [0, 1] , \tag{2}$$

which quantifys the expectation of the difference between predicted softmax likelihood $\hat{p}$ and accuracy and can be approximated by the Expected Calibration

Error (ECE) [6]. The predicted probabilities from a neural network are partitioned into $M$ bins and a weighted average of the difference between accuracy and confidence (i.e. softmax likelihood) is taken:

$$\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} |\text{acc}(B_m) - \text{conf}(B_m)| \, , \tag{3}$$

with total number of inputs $n$ and set of indices $B_m$ whose confidences fall into that particular bin. Intuitively for all pixels that are predicted with a softmax likelihood of 0.2 the expected accuracy should equal 20 percent [16].

*Uncertainty Weighted Averaging.* For training a model across partially labelled clients each client's model gains a separate segmentation head i.e. the last layer. The backbone of the model until the last layer is trained in a federated fashion. This enables us to learn each client's labels while simultaneously perform feature extraction across all tasks similar to [24]. To obtain the final prediction we average the predictions from all classifiers per channel and divide each channel by the number of clients possessing annotations of that particular label $\overline{\mathbf{p}} = \frac{1}{\mathbf{k}} \sum_{k=1}^{K} \hat{\mathbf{p}}_k$ with $\mathbf{k} = [k_1, k_2, ...] \in \mathbb{R}^K$ the number of clients that particular label was present and $\hat{\mathbf{p}}_k \in \mathbb{R}^{k_i \times H \times W}$ the predicted per-channel softmax probability of each segmentation head.

Surprisingly, we noticed that although all classifiers are not aware of at least two labels present only at other clients, these unannotated structures in their local dataset were visible in the classifier's predictive uncertainty (see Fig. 5). The segmentation heads seem to have learned that there might be a structure but did not have the ability to assign a specific label to it due to predefined channels. We interpret the uncertainty as the probability of "something" being present, while the background represents the probability for "nothing". Consequently, we reduce the likelihood for a pixel being background by the probability of the presence of a structure encoded in the uncertainty. In areas where $\mathbf{u}$ is large, the probability for a structure being present is high. We thus multiply the background channel (the first channel 0) with the inverse of the uncertainty $\overline{\mathbf{p}}_0 = (1 - \mathbf{u})\overline{\mathbf{p}}_0$. Fig. 2 shows a graphical explanation of the proposed averaging procedure for one specific pixel and for the whole image. Our approach represents a self-supervised method for reweighting the predictive probability by the self-supervised learned un-annotated structures for each segmentation head, while still obtaining an unambiguous prediction due to the preservation of probabilities. To ensure the probabilities for each pixel sum to one we must adjust the MultiTalent pipeline to be trained with cross entropy followed by a softmax. This enables us for each pixel to have the softmax probability of "something" compared to the probability to "nothing" that sums to one. We control for calibration of our probabilities in terms of ECE.

*Data and Model Architecture.* In total we used 8 datasets for training comprising a total of 2413 3D images across 12 unique classes and 2 datasets for testing with 240 samples. The training datasets are further split 80/20 % for training

Table 1: Comparison of DICE scores of trained models on unseen 20% test splits and completely unseen sites. For comparison we train a model on one dataset only and evaluated on the same and all others respectively (row 1 and 2). Further we compare our federated trained version to the centralized case (Fed Avg and Cen Avg). For both we also use uncertainty weighted averaging (FUNAvg and CUNAvg). **M** denotes the row-wise mean. We applied a Wilcoxon signed-rank test for each dataset between vanilla and uncertainty weighted averaging.

| Method | VSC | LiTS | Abd | Cvx | Chao | A1k | L2R | TS | VGC | Amo | M |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Same | **76.92** | **95.21** | 85.83 | 85.87 | **97.68** | **90.02** | 74.36 | 82.29 | 78.84 | **80.86** | **84.79** |
| Others | 63.76 | 41.63 | 64.10 | 42.54 | 37.11 | 37.06 | 27.81 | 67.14 | 57.09 | 58.88 | 49.71 |
| Cen Avg | 77.52 | 91.24 | 84.56 | **85.05** | 95.41 | 82.98 | **87.95** | 69.65 | **87.44** | 55.57 | 81.74 |
| CUNAvg | 74.51 | 87.95 | 82.39 | 80.4 | 92.86 | 80.16 | 85.91 | 69.07 | 83.18 | 57.60 | 79.40 |
|  | $< 0.01$ | $< 0.01$ | 0.02 | 0.07 | 0.03 | 0.01 | 0.41 | 0.35 | $< 0.01$ | 0.63 | $< 0.01$ |
| Fed Avg | **77.67** | 92.51 | 84.22 | 78.25 | 91.92 | 87.73 | 85.11 | 77.94 | 85.12 | 65.94 | 82.64 |
| FUNAvg | 77.26 | **93.06** | **86.87** | 78.88 | **94.63** | **86.63** | 86.35 | **85.19** | 85.31 | **73.49** | **84.77** |
|  | 0.04 | 0.45 | 0.01 | 0.89 | 0.19 | 0.54 | 0.78 | 0.03 | 0.8 | $< 0.01$ | 0.07 |

and testing. Each dataset represents one client in the federation. We chose our test set such that no client is in possession of all labels and that each label is at least present once in both test- and training sets. A notable characteristic of these datasets is their diversity in field of view, contrast and noise (see Fig. 6 in the appendix). This variation not only mirrors the real-world scenarios encountered in medical imaging but also challenges and validates the robustness of our model in handling diverse imaging conditions and patient-specific variations. In Fig. 3 we show all labels present in the datasets and their respective distribution across the federation. Preprocessing and training was done using the nnUNet framework [8] as in [24] with the above mentioned adaptations.

## 4   Results

We performed an 80-20% train-test split at each client. For a baseline, we trained models on single clients (no FL) and evaluated them in the intra-client scenario on the same client (row 1 in Tab. 1) and inter-client scenario on the test split of all other clients where the target label was available (row 2 in Tab. 1). We then performed training similar to MultiTalent [24], where each dataset obtained a separate segmentation head. We performed vanilla averaging and uncertainty weighted averaging for the centralized (Cen Avg and CUNAvg, rows 3 and 4) and federated setting (Fed Avg and FUNAvg, rows 5 and 6).

Despite the challenges of different quantities of annotated labels, CT scans, and fields of view, federated training was successful for all structures present at any client. Our method (DICE=84.77) is on par with the models trained and evaluated on the same dataset (84.79) and outperforms the centralized setting (81.74) when uncertainty is utilized to enhance predictions. This improvement is especially high in scenarios with limited data, such as Learn2Reg [27],

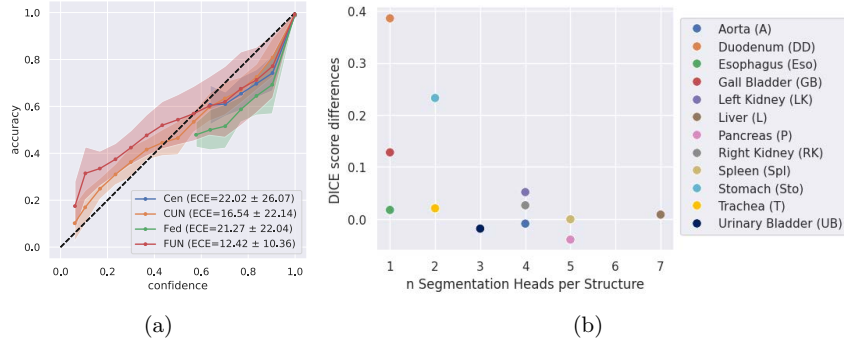(a)                                    (b)

Fig. 4: Calibration of the different methods in terms of Expected Calibration Error (a) and performance gain of FUNAvg in comparison to vanilla averaging the different logits from the federated training segmentation heads (b).
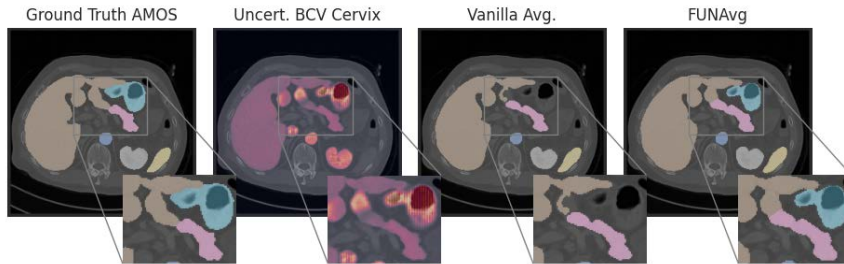


Fig. 5: Qualitative results on the AMOS test set. The classifier trained on the BCV Cervix dataset does not predict any of the labels present in the ground truth. Still, it learns their presence in the uncertainty.

where overfitting occurs in the non-federated case (86.35 vs. 74.36). The models trained on one dataset generally performed poorly when applied to other datasets (49.71). In the centralized case, the predictions could not be enhanced by performing uncertainty weighted averaging (81.74 vs. 79.40). This could be explained by a worse calibration of the central trained model in terms of ECE (12.42 vs. 16.54, see Fig. 4).

## 5    Discussion

The information for segmenting all structures across the datasets is encoded in the (federated trained) backbone. Unable to assign a previously unseen label in its fixed number of channels, the segmentation head only learns the presence of some structures encoded in the uncertainty, without being explicitly trained for it. For example, Fig. 5 illustrates how a network with a head trained on the BCV Cervix[15] dataset can segment organs well in the uncertainty when applied to

the AMOS [9] dataset, which contains none of the labels used for training. In federated learning, the segmentation heads with non-overlapping labels may use the same feature maps from the backbone, whereas in the centralized case, the heads tend to use different feature maps. The improvement of our proposed FUNAvg is larger for underrepresented labels (see Fig. 4 and Tab. 2 in the Appendix). We believe that predictions for these structures are encoded in the uncertainties of the other clients, whereas for labels present at many sites, predictions are already encoded in the output. For these labels, our proposed method may lead to slightly worse results. Future work could explore whether uncertainty can be used more targeted during averaging in such cases.

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

# References

1. Bilic, P., Christ, P., Li, H.B., Menze, B., et al.: The liver tumor segmentation benchmark (lits). Medical Image Analysis **84**, 102680 (2023). `https://doi.org/https://doi.org/10.1016/j.media.2022.102680`

2. Boughorbel, S., Jarray, F., Venugopal, N., Moosa, S., Elhadi, H., Makhlouf, M.: Federated Uncertainty-Aware Learning for Distributed Hospital EHR Data. In: Machine Learning for Health (ML4H) at NeurIPS (2019)

3. Chen, H.Y., Chao, W.L.: FedBE: Making Bayesian model ensemble applicable to federated learning. In: ICLR (2021). `https://doi.org/10.48550/arXiv.2009.01974`

4. Fuchs, M., González, C., Mukhopadhyay, A.: Practical uncertainty quantification for brain tumor segmentation. In: MIDL. vol. 172, pp. 407–422 (2022)

5. Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: ICML. vol. 48, pp. 1050–1059 (2016). `https://doi.org/10.5555/3045390.3045502`

6. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On Calibration of Modern Neural Networks. In: ICML. vol. 70, pp. 1321–1330 (2017). `https://doi.org/10.5555/3305381.3305518`

7. Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H.R., Xu, D.: Swin UNETR: Swin transformers for semantic segmentation of brain tumors in mri images. In: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries at MICCAI. pp. 272–284 (2022). `https://doi.org/10.1007/978-3-031-08999-2_22`

8. Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J., Maier-Hein, K.H.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods **18**(2), 203–211 (2021). https://doi.org/10.1038/s41592-020-01008-z

9. Ji, Y., Bai, H., GE, C., Yang, J., Zhu, Y., Zhang, R., Li, Z., Zhanng, L., Ma, W., Wan, X., Luo, P.: AMOS: A Large-Scale Abdominal Multi-Organ Benchmark for Versatile Medical Image Segmentation. In: NeuRIPS. vol. 35, pp. 36722–36732 (2022). https://doi.org/10.5281/zenodo.7262581

10. Karimireddy, S.P., Kale, S., Mohri, M., Reddi, S., Stich, S., Suresh, A.T.: SCAF-FOLD: Stochastic controlled averaging for federated learning. In: ICML. vol. 119, pp. 5132–5143 (2020)

11. Kavur, A.E., Gezer, N.S., Barış, M., , Isensee, F., Maier-Hein, K.H., Selver, M.A., et al.: CHAOS Challenge - combined (CT-MR) healthy abdominal organ segmentation. Medical Image Analysis **69**, 101950 (2021). https://doi.org/10.1016/j.media.2020.101950

12. Kendall, A., Gal, Y.: What uncertainties do we need in bayesian deep learning for computer vision? In: NeuRIPS. vol. 30 (2017)

13. Kohl, S., Romera-Paredes, B., Meyer, C., De Fauw, J., Ledsam, J.R., Maier-Hein, K., Eslami, S.M.A., Jimenez Rezende, D., Ronneberger, O.: A Probabilistic U-Net for Segmentation of Ambiguous Images. In: NeuRIPS. vol. 31 (2018)

14. Kwon, Y., Won, J.H., Kim, B.J., Paik, M.C.: Uncertainty quantification using Bayesian neural networks in classification: Application to ischemic stroke lesion segmentation. In: MIDL (2018), https://openreview.net/forum?id=Sk_P2Q9sG

15. Landman, B., Xu, Z., Igelsias, J., Styner, M., Langerak, T., Klein, A.: MICCAI multi-atlas labeling beyond the cranial vault–workshop and challenge. In: MICCAI Multi-Atlas Labeling Beyond Cranial Vault-Workshop Challenge (2015). https://doi.org/10.7303/SYN3193805

16. Laves, M.H., Tölle, M., Schlaefer, A., Engelhardt, S.: Posterior temperature optimized bayesian models for inverse problems in medical imaging. Medical Image Analysis **78**, 102382 (2022). https://doi.org/10.1016/j.media.2022.102382

17. Li, T., Sahu, A.K., Zaheer, M., Sanjabi, M., Talwalkar, A., Smith, V.: Federated optimization in heterogeneous networks. In: MLSys (2020)

18. Linsner, F., Adilova, L., Däubener, S., Kamp, M., Fischer, A.: Approaches to uncertainty quantification in federated deep learning. In: Workshop on Parallel, Distributed, and Federated Learning at ECML. vol. 2, pp. 128–145 (2021)

19. Ma, J., Zhang, Y., Gu, S., Zhu, C., Ge, C., Zhang, Y., An, X., Wang, C., Wang, Q., Liu, X., Cao, S., Zhang, Q., Liu, S., Wang, Y., Li, Y., He, J., Yang, X.: AbdomenCT-1K: Is Abdominal Organ Segmentation a Solved Problem? **44**(10), 6695–6714 (2022). https://doi.org/10.1109/TPAMI.2021.3100536

20. McMahan, H.B., Moore, E., Ramage, D., Hampson, S., y Arcas, B.A.: Communication-efficient learning of deep networks from decentralized data. In: AISTATS (2017)

21. Oktay, O., Ferrante, E., Kamnitsas, K., Heinrich, M., Bai, W., Caballero, J., Cook, S., Marvao, A., Dawes, T., O'Regan, D., Kainz, B., Glocker, B., Rueckert, D.: Anatomically constrained neural networks (ACNN): Application to cardiac image enhancement and segmentation **37**(2) (2018). https://doi.org/10.1109/TMI.2017.2743464

22. Rädsch, T., Reinke, A., Weru, V., Tizabi, M.D., Schreck, N., Kavur, A.E., Pekdemir, B., Roß, T., Kopp-Schneider, A., Maier-Hein, L.: Labelling instructions matter in biomedical image analysis. Nature Machine Intelligence **5**(3), 273–283 (2023). https://doi.org/10.1038/s42256-023-00625-5

23. Thorgeirsson, A.T., Gauterin, F.: Probabilistic Predictions with Federated Learning **23**(1), 41. https://doi.org/10.3390/e23010041
24. Ulrich, C., Isensee, F., Wald, T., Zenk, M., Baumgartner, M., Maier-Hein, K.H.: MultiTalent: A Multi-Dataset Approach to Medical Image Segmentation. In: MIC-CAI (2023). https://doi.org/10.1007/978-3-031-43898-1_62
25. Wasserthal, J., Breit, H.C., Meyer, M.T., Pradella, M., Hinck, D., Sauter, A.W., Heye, T., Boll, D., Cyriac, J., Yang, S., Bach, M., Segeroth, M.: TotalSegmentator: robust segmentation of 104 anatomical structures in CT images. Radiology: Artificial Intelligence **5**(5) (2023). https://doi.org/10.1148/ryai.230024
26. Xu, X., Deng, H.H., Gateno, J., Yan, P.: Federated multi-organ segmentation with inconsistent labels. Transactions on Medical Imaging **42**(10), 2948–2960 (2023). https://doi.org/10.1109/TMI.2023.3270140
27. Xu, Z., Lee, C.P., Heinrich, M.P., Modat, M., Rueckert, D., Ourselin, S., Abramson, R.G., Landman, B.A.: Evaluation of Six Registration Methods for the Human Abdomen on Clinically Acquired CT. Transactions on Biomedical Engineering **63**(8), 1563–1572 (2016). https://doi.org/10.1109/TBME.2016.2574816
28. Zhou, Y., Li, Z., Bai, S., Wang, C., Chen, X., Han, M., Fishman, E., Yuille, A.: Prior-aware neural network for partially-supervised multi-organ segmentation. In: ICCV (2019). https://doi.org/10.1109/ICCV.2019.01077