FedIOD: Federated Multi-Organ Segmentation from Partial Labels by Exploring Inter-Organ Dependency

Qin Wan, Zengqiang Yan, *Member, IEEE*, and Li Yu, *Senior Member, IEEE*

***Abstract*— Multi-organ segmentation is a fundamental task and existing approaches usually rely on large-scale fully-labeled images for training. However, data privacy and incomplete/partial labels make those approaches struggle in practice. Federated learning is an emerging tool to address data privacy but federated learning with partial labels is under-explored. In this work, we explore generating full supervision by building and aggregating inter-organ dependency based on partial labels and propose a single-encoder-multi-decoder framework named FedIOD. To simulate the annotation process where each organ is labeled by referring to other closely-related organs, a transformer module is introduced and the learned self-attention matrices modeling pairwise inter-organ dependency are used to build pseudo full labels. By using those pseudo-full labels for regularization in each client, the shared encoder is trained to extract rich and complete organ-related features rather than being biased toward certain organs. Then, each decoder in FedIOD projects the shared organ-related features into a specific space trained by the corresponding partial labels. Experimental results based on five widely-used datasets, including LiTS, KiTS, MSD, BCTV, and ACDC, demonstrate the effectiveness of FedIOD, outperforming the state-of-the-art approaches under in-federation evaluation and achieving the second-best performance under out-of-federation evaluation for multi-organ segmentation from partial labels. The source code is publicly available at https://github.com/vagabond-healer/FedIOD.**

***Index Terms*— Transformer, Partial Labeling, Federated Learning, Self-Attention, Organ Segmentation**

I. Introduction

ORGAN segmentation is a fundamental task in medical image analysis for disease diagnosis [1], radiomics analysis [2], treatment response assessment [3], and surgical planning and navigation [4], and automatic multi-organ segmentation is of great demand. Existing work on multi-organ segmentation relies on a collection of fully-labeled data for training either convolutional neural networks (CNN) [5]–[7] or transformers [8]–[10]. However, in clinical practice, training data can be decentralized and forbidden to share due to privacy, and organ labels are incomplete due to specific tasks, making those fully supervised approaches infeasible.

Federated learning [11], [12] addresses the former issue by sharing model parameters/gradients rather than data across clients for privacy preservation. Unfortunately, when each client only contains partially-labeled data for training, client-wise model parameters/updates would diverge significantly which in turn degrade the convergence performance of federated learning. Till now, federated learning with partial labels for multi-organ segmentation is under-explored and existing few works mainly focus on architecture design [13], [14] and loss function construction [15] to minimize the negative impact of mislabeled organs.

In partially supervised learning, one typical setting is multi-label classification. Specifically, Durand *et al.* [16] introduced a new classification loss to exploit label proportion information and predict missing labels. Dong *et al.* [17] proposed to argument partial labels by vicinal risk minimization and further extended it to federated learning [18]. Yu *et al.* [19] integrated self-supervision and self-distillation into an innovative multi-task framework to tackle label conflict in partial labels. Compared to multi-label classification, multi-organ segmentation additionally encounters greater variations in object scales, making it more challenging. Furthermore, in federated learning, the local model of each client will be more biased to its own partial labels, resulting in poor convergence of the federated model and degrading the overall segmentation performance.

In this paper, we propose to develop pseudo full-organ labels (*i.e*, containing all organs) from partially-labeled data by exploring inter-organ dependency and build a single-encoder-multi-decoder framework named FedIOD. In clinical practice, professionals annotate one certain organ by referring to other organs for localization and boundary determination. In other words, given a segmentation model for one specific organ, its encoder would also contain the features of those closely-related organs. Inspired by this, a transformer module is introduced to model inter-organ dependency through self-attention calculation. By aggregating the self-attention matrices of clients, it is promising to build a more precise and complete inter-organ dependency map (*i.e.*, indicating the positions of all organs) to regularize the training of the shared encoder to extract complete organ-related features. Then the partial label learning problem is formulated and solved by projecting

This work was supported in part by the National Natural Science Foundation of China under Grants 62271220 and 62202179 and in part by the Natural Science Foundation of Hubei Province of China under Grant 2022CFB585. *(Qin Wan and Zengqiang Yan are co-first authors contributed equally to this work.) (Corresponding author: Li Yu.)*

Qin Wan, Zengqiang Yan, and Li Yu are with the School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan 430074, China (e-mail: qwan@hust.edu.cn; z_yan@hust.edu.cn; hustlyu@hust.edu.cn).

the complete organ-related feature into various organ-specific spaces according to partial labels. The main contributions are summarized as follows:

- The first attempt to build pseudo full-organ labels by exploring inter-organ dependency from partial labels for multi-organ segmentation in federated learning.
- A single-encoder-multi-decoder architecture to formulate the partial label learning problem as feature projection from a common/shared space to any organ-specific space.
- A dynamic weighting mechanism for federated transformer aggregation through organ-specific contribution quantification.
- Superior performance on both abdominal and cardiac multi-organ segmentation under various evaluation settings against the state-of-the-art centralized and federated partial label learning approaches.

The rest of this paper is organized as follows. Section II summarizes related works on segmentation from partial labels under centralized learning and federated learning, and Section III describes FedIOD in detail. We present a thorough evaluation of FedIOD against the state-of-the-art methods in Section IV and ablation studies in Section V. Section VI concludes this paper.

II. Related Work

Segmentation from partial labels is to segment all objects based on training samples with only partially-labeled objects, which can be categorized into centralized learning and federated learning.

A. Centralized Learning with Partial Labels

In centralized learning, as all training samples with conflicting partial labels [20] are directly used for training, the main challenge is how to address label uncertainty in partial labels and recover real objects from mislabeled backgrounds. Existing works typically are based on loss functions, conditions/rules, or multi-path networks.

1) Based on Loss Functions: Zhou *et al.* [21] proposed to use the prior knowledge on organ sizes estimated based on a fully-labeled dataset to guide the training from partial labels. Both the loss function and the training process is complicated, and requiring a fully-labeled dataset makes it less attractive. Roulet *et al.* [20] constructed an adaptive cross-entropy loss to decompose the multi-class segmentation problem into several binary classification problems, and Fang *et al.* [22] designed a similar loss function TAL by directly regarding mislabeled objects as background in partial labels. It should be noted that the approaches in [20] and [22] can be regarded as a simplified form of noisy label learning [23], being effective if the ratio of mislabeled foreground objects is much smaller compared to the background. However, in medical image segmentation, especially in multi-organ segmentation, foreground organs would cover most regions in slices, making them struggle. Fidon *et al.* [24] proposed label-set loss functions to extend the Dice loss from fully-labeled to partially-labeled settings. Shi *et al.* [25] designed both marginal loss and exclusion loss for partially supervised multi-organ segmentation. Though both designs are extendable to existing loss functions, it may make the training process unstable, penalizing the model to focus more on the large-size organs while somewhat ignoring others.

2) Based on Conditions/Rules: Dmitriev *et al.* [26] introduced additional class-aware information as constraints to classical encoder-decoder architectures to dynamically determine the usage of specific features according to the corresponding partial labels, but projecting low-dimension features into a high dimension brings additional computational complexity. Zhang *et al.* [27] proposed DoDNet by using both a class-specific code and learning deep features as constraints and adaptively splitting the decoder into different groups through a dynamic filter network to control the decoder's outputs depending on partial labels. Compared to [26], DoDNet is more efficient and flexible for multi-organ and multi-lesion segmentation. Zhang *et al.* [28] proposed a conditional segmentation strategy to propagate labels from multiple partially-annotated images to each target image and a dual learning strategy to provide substantial supervision for unlabeled structures via CycleGAN [29]. Dong *et al.* [30] proposed Vicinal Labels Under Uncertainty (VLUU) to transform the partially supervised problem into a fully supervised problem based on human structure similarity, which is applicable to different network architectures. Both [28] and [30] focused on exploring prior knowledge of human organs like locations and relative sizes to assist those mislabeled organs in partially-labeled data.

3) Based on Multi-Path Networks: Huang *et al.* [31] and Zhang *et al.* [32] adopted multi-model and multi-stage frameworks to convert partial supervision into full supervision, achieving excellent performance on imbalanced multi-organ segmentation but suffering from higher training and model complexity. Chen *et al.* [33] deployed a multi-path decoder to deal with various partial labels and a shared encoder for organ-related representation learning. Petit *et al.* [34] proposed an iterative confidence self-training approach to relabel missing pixel labels. Liu *et al.* [35] applied incremental learning into partially supervised learning and utilized a light memory module and contrastive learning based loss functions to address catastrophic forgetting. Similarly, Zhou *et al.* [36] adopted incremental learning and designed a background label alignment strategy and an uncertainty-aware guidance strategy respectively to guide knowledge transfer. Despite the efficiency of multi-path networks, such approaches would encounter high training complexity in practice.

B. Federated Learning with Partial Labels

Compared to centralized learning where partial labels covering all organs are accessible for training, in federated learning, each client only owns partial labels of certain organs, and different clients are not allowed to share partially-labeled data. As each client in federated learning trains the global model only based on its partial labels, there exist severe model parameter variations across clients, making it more challenging for federated model convergence.

Dong *et al.* [18] investigated federated learning with partial labels for classification. Shen *et al.* [14] explored training a multi-task segmentation model based on several independent

datasets with different annotations of organs and tumors in federated learning but suffered from knowledge conflict. Xu *et al.* [13] proposed Fed-MENU for multi-organ segmentation where a multi-encoder-single-decoder UNet-like network, named MENU-Net was designed for feature extraction of each organ by an individual encoder. In addition, to learn informative and distinctive organ-specific features, an auxiliary generic decoder (AGD) was introduced to regularize the training of MENU-Net. Liu *et al.* [15] first trained a segmentation model based on a fully-labeled dataset and then adapted the model to each client based on partially-labeled data. Relying on fully-labeled data makes it less attractive in clinical practice. Till now, federated learning with partial labels is under-explored.

C. Similarity and Dissimilarity with Existing Works

Addressing label conflicts is crucial in partially supervised multi-organ segmentation in both centralized learning and federated learning. Existing works focus on developing appropriate loss functions, adopting conditioned networks, and constructing multi-path architectures. FedIOD can be categorized as a multi-path architecture. Different from existing works, it adopts a single-encoder-multi-decoder architecture. Partial label learning is a sub-task of noisy label learning [30]. It has been identified that for DNN-based models, deep layers (*i.e.*, close to output) are more sensitive to label noise [37]. In other words, shallow layers are more robust to reduce the noise caused by label conflicts in partial labels while deep layers are more closely-related to segmentation supervised by partial labels, forming a single-encoder-multi-decoder architecture.

In addition to balancing organ-specific partial labels like existing approaches, FedIOD is the first trial to generate pseudo full-organ labels to explicitly regularize the shared encoder for complete organ-relevant feature extraction. It is realized by introducing a transformer module and globally aggregating the learned self-attention matrices to build and fuse inter-organ dependency.

III. Method

The main idea behind FedIOD is to construct pseudo full-organ supervision by exploring inter-organ dependency in partial labels. Benefiting from self-attention, it is possible to build pair-wise dependency across the image space. Unfortunately, supervised by partial labels, transformer tends to focus more on specific organs instead of building global dependency for all organs. Considering the stable structure of organs in medical imaging, even given organ-specific partial labels, self-attention may activate other organ regions reflecting inter-organ dependency. Therefore, aggregating self-attention matrices is likely to construct pseudo full-organ labels for regularization. In the following sections, we explain in detail how to build pseudo full labels of all organs by self-attention aggregation and supervise the training of the shared encoder to extract complete organ-related features.

A. Preliminaries

Given K clients in federated learning and the corresponding data $D = \{D_1, ..., D_k, ..., D_K\}$ where $D_k = \{X_k, Y_k\}$ represents the data and labels of client k, both data and labels across clients are different and non-overlapping, namely $X_i \cap X_j = \emptyset$ and $Y_i \cap Y_j = \emptyset$ for any client pair i and j. Following common settings and simulating the most challenging scenarios, Y_k contains only one class in binary and each client aims to segment only one organ type. In other words, the total number of classes C equals K, *i.e.* $C = K$. In the following sections, we still use C to represent the total number of classes.

For federated training, the federated model consists of one shared encoder Enc and C decoders $\{Dec_1, ...Dec_c, ...Dec_C\}$ where D_c is the decoder for class c. After training on all K clients, the federated model is updated based on the aggregated model parameters/gradients on the server.

B. Overall Architecture

The overall architecture of FedIOD follows a shared-encoder-multi-decoder design. As the primary goal is to segment all organs based on partially-labeled local data, the encoder is expected to model complete organ-related features, and thus all clients share the same encoder during training. On the other hand, different clients vary in annotations (*i.e.*, including only partial organs), which is formulated as projecting the organ-related features into different annotation spaces. Thus, each client would share a category-specific decoder according to its labels. It is also based on the observation that, given partial labels, some foreground classes would be labeled as background. Training a shared decoder based on such "noisy" labels may be counter-productive.

During local training, each client k with partial labels of class c would jointly update both Enc and Dec_c based on $D_k = \{X_k, Y_k\}$ and keep other decoders frozen. Both the Dice loss L_{dice} and the Cross-Entropy loss L_{ce} are used for the training of each client. During back-propagation, the loss of each client k with partial labels of class c is backpropagated from the decoder Dec_c to the shared encoder Enc and finally to the input end. During federated updating, Enc is updated by all clients while Dec_c is only updated by those labels containing c. In this way, the shared encoder Enc is trained by labels of all organs to extract complete organ-relevant features. To avoid being dominated by large-scale organs, we further build pseudo full-organ labels to regularize Enc by aggregating cross-client inter-organ dependency which is described in the following.

C. Building Client-Wise Inter-Organ Dependency

In clinical practice, organs are not annotated independently. Instead, inter-organ relationships are used for annotation implicitly. For instance, when annotating the pancreas, it is natural to first localize the liver with a much larger size and then search for the pancreas with a smaller size. Inspired by this, given a certain organ/class, if we can quantitatively evaluate the importance of other organs for its segmentation, it is promising to capture the inter-organ relationships and localize other organs. By building inter-organ relationships among all organs, it is possible to produce a complete annotation containing all organs from partial labels.

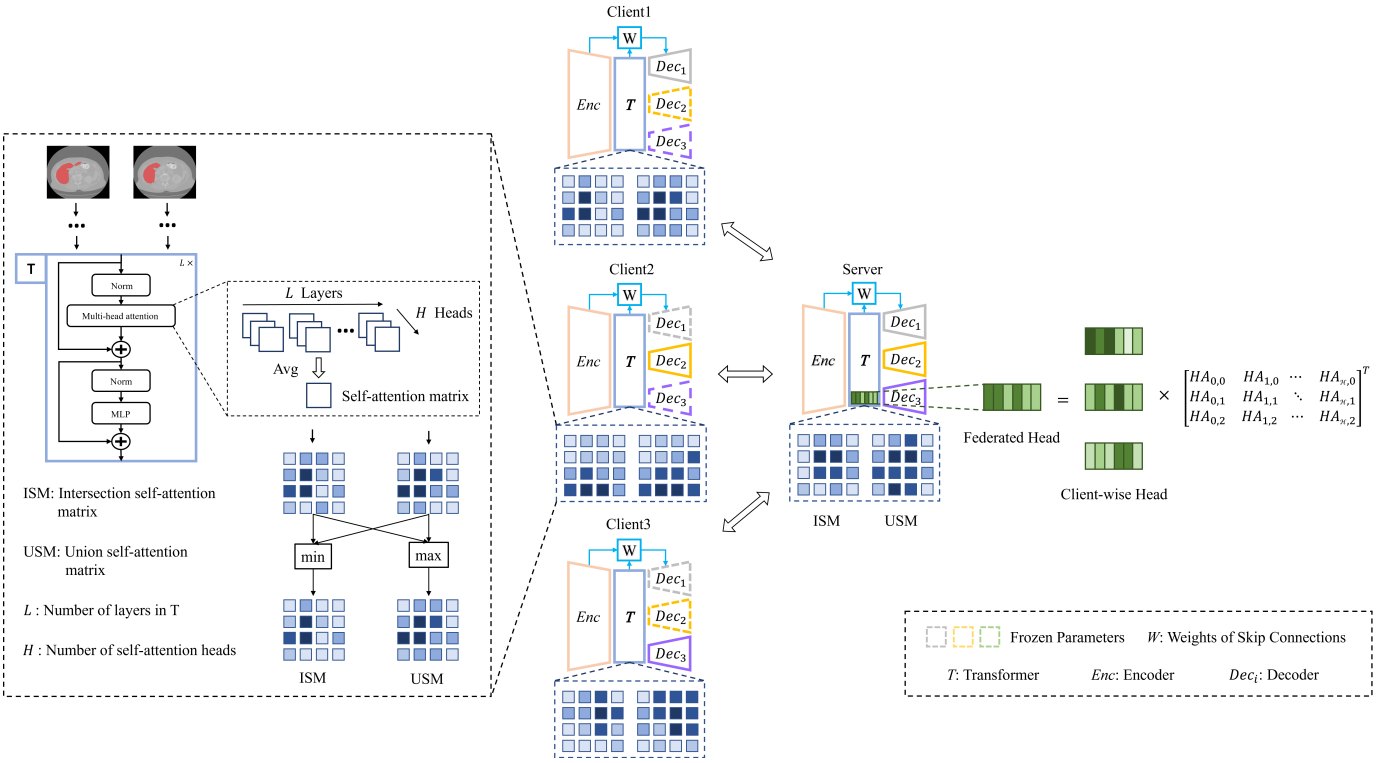

Fig. 1: Overview of FedIOD. During local updating, each client jointly trains the shared encoder and a class-specific decoder according to its available partial labels and then sends the intersection and union self-attention matrices calculated based on learned self-attention matrices to the server for federation. During federated updating, Enc and Dec_i are aggregated by FedAvg while the self-attention heads in T are updated by dynamically re-weighting client-wise self-attention heads.

Transformer [38], [39] is adopted to build inter-organ dependency locally in each client, due to its superior performance in capturing long-range dependency. In addition, considering convolution's relatively limited receptive fields, it is more likely to focus on the target organ/class trained by partial labels. Comparatively, transformers build pairwise dependency for each token/patch, it is more likely to include "redundancy", indicating the importance of other organs. Therefore, a transformer module T is attached to the shared encoder Enc as illustrated in Fig. 1. Specifically, T is to build inter-organ dependency through self-attention calculation and re-balance the features corresponding to different organs in Enc.

Given the ℓ-th transformer layer consisting of $\mathscr{H}$ self-attention heads in T and the corresponding self-attention matrices $A_1^\ell, A_2^\ell, ..., A_{\mathscr{H}}^\ell \in \mathbb{R}^{(w\times h)\times(w\times h)}$. According to [40], the first few self-attention matrices would focus more on the low-level similarity (*e.g.*, texture) and deeper transformer layers emphasize more high-level semantic similarity for dependency establishment. In the meantime, self-attention matrices of deep transformer layers can be messier due to complicated semantic information. Therefore, an average self-attention matrix is calculated by

$$\bar{A} = \frac{1}{\mathscr{H}\times\mathscr{L}}\sum_{\ell=1}^{\mathscr{L}}\sum_{\hbar=1}^{\mathscr{H}} A_\hbar^\ell \in \mathbb{R}^{(w\times h)\times(w\times h)}, \tag{1}$$

where $\mathscr{L}$ denotes the total number of transformer layers in T. In this way, $\bar{A}$ builds long-range dependency based on global features while being robust to noise. Based on $\bar{A}$, the importance of each token/patch is calculated by column-wise summation, *i.e.*,

$$W = \textbf{Softmax}(\sum_{i=1}^{w\times h}\bar{A}_i) \in \mathbb{R}^{(w\times h)\times 1}, \tag{2}$$

where $\bar{A}_i$ represents the i-th column of $\bar{A}$. After reshaping W into $W \in \mathbb{R}^{w\times h}$, each value $W_{i,j}$ represents the dependency between the token/position (i,j) with all other tokens/positions. In other words, the higher $W_{i,j}$ is, the more important the token (i,j) is. As the segmentation of any certain organ is more related to other organs (*i.e.*, foreground) rather than the background, higher $W_{i,j}$ is more likely to be organs. By using W to re-weight skip connections from Enc, features corresponding to different organs are re-balanced in Enc. Through T for re-weighting, even without complete supervision, the shared encoder Enc would focus not only on a specific organ according to partial labels but also on other organs.

D. Aggregating Cross-Client Inter-Organ Dependency

As described above, through the transformer module T on each client, even based on partial labels, the learned self-attention matrix would indicate the locations of some other organs. However, without proper supervision, the self-attention matrix from each client would mainly contain the specific

organ regions and those highly-dependent organs while ignoring others. Fortunately, thanks to the stable structures and locations of organs, if we combine the self-attention matrices from all clients, it is possible to build a complete self-attention matrix containing all organs. By using the complete self-attention matrix to supervise the learning of each client's self-attention matrix, the shared encoder Enc is trained for feature extraction of all organs with equal importance. The key is how to build a high-quality complete self-attention matrix from class-specific self-attention matrices from clients.

Despite the stable structures of organs, there exist variations across scans/patients, making direct fusion sub-optimal. Thus, given each client's self-attention matrices, we separately calculate an Intersection Self-attention Matrix (ISM) to highlight the common inter-organ dependency and a Union Self-attention Matrix (USM) to preserve cross-scan variations. Given any client k and its data D_k containing M_k 3D scans, let $D_{k,m}$ denote the number of 2D slices of the m-th 3D scan in D_k. To pursue high-quality self-attention matrices, only the top $\tau\%$ 3D scans with the lowest Dice losses are included for self-attention fusion, denoted as $\mathbb{M}_k$. After local training, given any slice $X_{m,d}$ from a 3D scan $X_m \in \mathbb{M}_k$, its averaged self-attention matrix is extracted as $\bar{A}_k^{m,d} \in \mathbb{R}^{(w\times h)\times(w\times h)}$. Then, by stacking the self-attention matrices of all the slices from X_m, we can obtain the 3D self-attention matrix $\bar{A}_k^m \in \mathbb{R}^{D_{k,m}\times(w\times h)\times(w\times h)}$ by

$$\bar{A}_k^m = \textbf{Stack}(\bar{A}_k^{m,d}; d = 1, ..., D_{k,m}). \tag{3}$$

As scans vary significantly in depth (*i.e.* $D_{k,m}$), we project all 3D self-attention matrices across clients into the same depth by interpolation

$$\tilde{A}_k^m = \textbf{Interpolate}(\bar{A}_k^m) \in \mathbb{R}^{\bar{D}\times(w\times h)\times(w\times h)}, \tag{4}$$

where $\bar{D}$ is the median depth of all scans across clients. After reshaping the 3D self-attention matrix $\tilde{A}_k^m$ of each scan across clients to the same depth, the intersection self-attention matrix of client k is defined as

$$I_k = \min_{m\in\mathbb{M}_k} \tilde{A}_k^m \in \mathbb{R}^{\bar{D}\times(w\times h)\times(w\times h)}, \tag{5}$$

and the union self-attention matrix of client k is defined as

$$U_k = \max_{m\in\mathbb{M}_k} \tilde{A}_k^m \in \mathbb{R}^{\bar{D}\times(w\times h)\times(w\times h)}. \tag{6}$$

Both I_k and U_k of any client k are sent to the server to calculate the global intersection self-attention $\hat{I}$ and the global union self-attention matrix $\hat{U}$ according to

$$\begin{aligned} \hat{I} &= \frac{1}{K}\sum_{k=1}^{K} \phi_k I_k, \\ \hat{U} &= \frac{1}{K}\sum_{k=1}^{K} \rho_k U_k, \end{aligned} \tag{7}$$

where ϕ_k is defined as

$$\phi_k = \frac{e^{-(I_k - \frac{1}{K}\sum_{i=1}^{K} I_i)}}{\sum_{j=1}^{K} e^{-(I_j - \frac{1}{K}\sum_{i=1}^{K} I_i)}} \tag{8}$$

and ρ_k is defined as

$$\rho_k = \frac{e^{-(U_k - \frac{1}{K}\sum_{i=1}^{K} U_i)}}{\sum_{j=1}^{K} e^{-(U_j - \frac{1}{K}\sum_{i=1}^{K} U_i)}} \tag{9}$$

to down-weight those clients distant from the average. In this way, $\hat{I}$ and $\hat{U}$ are expected to contain the complete information of all organs, based on which to supervise the training of I_k and U_k would penalize the shared encoder Enc to better explore inter-organ dependency.

E. Regularizing Client-Wise Inter-Organ Dependency

As discussed above, both $\hat{I}$ and $\hat{U}$ are utilized to regularize the training of each client. Given any 2D slice $X_{m,d}$ from a scan X_m with depth $D_{k,m}$, the corresponding intersection and union self-attention matrices $\hat{I}_k^{m,d} \in \mathbb{R}^{(w\times h)\times(w\times h)}$ and $\hat{U}_k^{m,d} \in \mathbb{R}^{(w\times h)\times(w\times h)}$ are obtained from $\hat{I}$ and $\hat{U}$ by reverse interpolation. Given the self-attention matrix $\bar{A}_k^{m,d}$ of $X_{m,d}$, the goal is to ensure that $\bar{A}_k^{m,d}$ is a superset of $\hat{I}_k^{m,d}$ (*i.e.*, $\hat{I}_k^{m,d} \subseteq \bar{A}_k^{m,d}$) to preserve organ-related features and $\bar{A}_k^{m,d}$ is a subset of $\hat{U}_k^{m,d}$ (*i.e.*, $\bar{A}_k^{m,d} \subseteq \hat{U}_k^{m,d}$) to avoid redundant dependency. It is implemented by

$$\mathscr{L}_{I,k}^{m,d} = \frac{\sum_{i=1}^{w\times h}\sum_{j=1}^{w\times h} \hat{I}_k^{m,d}(i,j) \times \Gamma(\bar{A}_k^{m,d}(i,j); \alpha_I, \beta_I)}{\sum_{i=1}^{w\times h}\sum_{j=1}^{w\times h} \hat{I}_k^{m,d}(i,j)}, \tag{10}$$

and

$$\mathscr{L}_{U,k}^{m,d} = \frac{\sum_{i=1}^{w\times h}\sum_{j=1}^{w\times h} \bar{A}_k^{m,d}(i,j) \times \Gamma(\hat{U}_k^{m,d}(i,j); \alpha_U, \beta_U)}{\sum_{i=1}^{w\times h}\sum_{j=1}^{w\times h} \bar{A}_k^{m,d}(i,j)}, \tag{11}$$

where $\Gamma(A; \alpha, \beta)$ is a gating function defined in [41] defined as

$$\Gamma(A; \alpha, \beta) = \frac{1}{1 + e^{(-\alpha(A-\beta))}}, \tag{12}$$

where β is the gating threshold and α is a scaling factor. Given $A \geq \beta$, $\Gamma(A; \alpha, \beta)$ approaches 1. Otherwise, $\Gamma(A; \alpha, \beta)$ is close to 0. Then, the overall regularization losses of client k are rewritten as

$$\mathscr{L}_{I,k} = \sum_{m=1}^{M_k}\sum_{d=1}^{D_{k,m}} \mathscr{L}_{I,k}^{m,d}, \tag{13}$$

and

$$\mathscr{L}_{U,k} = \sum_{m=1}^{M_k}\sum_{d=1}^{D_{k,m}} \mathscr{L}_{U,k}^{m,d}. \tag{14}$$

F. Federated Model Parameter Aggregation

FedAvg [11] is the most commonly-used aggregation strategy. In FedAvg, the parameters of the federated model at round $r+1$ are updated by

$$\theta^{r+1} = \sum_{k=1}^{K} w_k \theta_k^r, \tag{15}$$

where θ_k^r denotes the model parameters updated by client k at round r and w_k is defined as

$$w_k = \frac{|D_k|}{\sum_{k=1}^{K} |D_k|}. \tag{16}$$

In federated learning with partial labels, clients vary dramatically in the availability of labels. As each client only has single-target/-organ labels, there exists a great difference in client-wise learning difficulty, resulting in various optimization degrees of locally-trained models. Therefore, simply aggregating locally-trained models based on data quantity is inappropriate, especially for the transformer module. As discussed in [42], [43], different self-attention heads vary in the regions of interest. In this work, as the shared transformer module T is updated at each client with different partial labels, the contributions of self-attention heads could be more different across organs. Therefore, it is necessary to dynamically adjust the weights of self-attention heads across clients based on their contributions to specific organs.

As discussed in [44], the importance of filters in convolutional neural networks can be measured based on their produced features. Similarly, the importance of each self-attention head in T could be measured based on the contribution of its produced features toward certain organs. Inspired by [45], [46], the self-attention matrix of each head, containing inter-organ dependency, is used for importance evaluation. Denoting the locally-updated model at client k as $F_k(\cdot)$ containing $\mathscr{H}$ self-attention heads, the corresponding self-attention matrices are denoted as $A_{1,k}, A_{2,k}, ..., A_{\mathscr{H},k}$. Then, top $\mu\%$ samples/scans with higher Dice scores from D_k are selected, forming a subset $\mathbb{U}_k$, and the importance of any self-attention matrix $A_{i,k}$ is measured by

$$\begin{aligned} &AA_{i,k}(A_{i,k}; X_u) \\ &\quad = A_{i,k} \times \int_{\varepsilon=0}^{1} \frac{\partial F([\varepsilon A_{1,k}, ..., \varepsilon A_{H,k}]; X_u)}{\partial A_{i,k}} \mathrm{d}\varepsilon \\ &\quad \approx \frac{1}{S} A_{i,k} \times \sum_{s=1}^{S} \frac{\partial F([\frac{s}{S} A_{1,k}, ..., \frac{s}{S} A_{H,k}]; X_u)}{\partial A_{i,k}}, \end{aligned} \tag{17}$$

where S is the step size. The importance of the corresponding self-attention head $Head_{i,k}$ is calculated by

$$HA_{i,k} = \max_{X_u \in \mathbb{U}_k} (AA_{i,k}(A_{i,k}; X_u)). \tag{18}$$

For federated parameter updating of T, the weight of the i-th self-attention head $Head_{i,k}$ of client k is defined as

$$\lambda_{i,k} = \frac{HA_{i,k}}{\sum_{j}^{K} HA_{i,J}}. \tag{19}$$

Then, the i-th self-attention head $Head_i$ in the federated model is aggregated and updated by

$$Head_i = \sum_{k=1}^{K} \lambda_{i,k} Head_{i,k}. \tag{20}$$

It should be noted that, except for the shared transformer module T, all other components are updated through FedAvg.

IV. Evaluation

A. Datasets

1) Abdominal: Following [13], four public abdominal CT image datasets, including 1) the liver tumor segmentation challenge (LiTS) dataset [47], denoted as **LIVER**, 2) the kidney tumor segmentation challenge (KiTS) dataset [48], [49], denoted as **KIDNEY**, and 3) the medical segmentation decathlon dataset (Task #7) [50], denoted as **PANCREAS**, and 4) the multi-atlas labeling beyond the cranial vault challenge (**BTCV**) dataset [51].

LIVER contains 131 images with pixel-wise annotations of the liver and tumors. For convenience, both liver and tumor regions are combined and regarded as the liver.

KIDNEY contains 210 images with pixel-wisely labeled kidney and tumor. The kidney tumor regions are treated as parts of the kidney for convenience.

PANCREAS contains 281 images with pixel-wise annotations of the pancreas and tumors. The pancreas tumor regions are merged as parts of the pancreas in our experiments.

BTCV contains 30 images with 13 pixel-wisely annotated organs from which the liver, kidney, and pancreas are selected.

2) Cardiac: The ACDC dataset [52] consists of short-axis cine-MRI from 150 patients acquired at the University Hospital of Dijon. Structures of interest, including left ventricle (LV), right ventricle (RV), and myocardium (Myo) were annotated manually by experienced experts on end-diastolic (ED) and end-systolic (ES) phase instants.

B. Evaluation Strategy

Following [13], we conduct both in-federation and out-of-federation evaluations. For in-federation evaluation, the test sets of all clients are "combined" for evaluation. In this way, the test set follows a similar distribution to the training sets of clients. For out-of-federation evaluation, a separate dataset is used for evaluation. In this case, the test set differs in the data distribution compared to the training sets, validating the generalizability of approaches. More specifically, **BTCV** is used as a separate test set for out-of-federation evaluation.

In terms of the evaluation metric, the Dice similarity coefficient (DSC) and the average symmetric surface distance (ASD) are utilized for evaluation.

C. Implementation Details

All 2D methods were implemented in PyTorch and trained using an SGD optimizer with an initial learning rate of 1e-2, a momentum of 0.9, and a batch size of 8 on one NVIDIA Geforce RTX 3090 GPU for 500 epochs. For training, each slice was randomly cropped and resized to 256×256 pixels and augmented by random rotation, intensity change, and gamma augmentation.

All 3D approaches were also implemented in PyTorch and trained using an SGD optimizer. The learning rate is initialized to be 0.01 and decayed throughout the training following a poly learning rate policy with a momentum factor of 0.9. Constrained by GPU memories, we set the batch size as 2. In training, 3D images were randomly cropped to 224×224×32

TABLE I: Quantitative comparison of different methods on the in-federation abdominal CT image datasets (*i.e.*, **LIVER**, **KIDNEY**, and **PANCREAS**). The best and second-best results are marked in bold and underlined. All comparison results are reproduced according to publicly-available source codes.

Method	Type	In-federation DSC (Mean(SD) %)				In-federation ASD (Mean(SD) mm)			
		Avg.	**LIVER**	**KIDNEY**	**PANCREAS**	Avg.	**LIVER**	**KIDNEY**	**PANCREAS**
MELoss [15]	3D	84.58	87.12(21.21)	91.79(14.89)	74.84(15.31)	3.98	5.85(11.51)	3.12(14.72)	2.97(4.62)
DoDNet [27]		83.54	81.92(18.11)	91.89(13.06)	76.80(9.69)	7.42	12.72(8.80)	6.09(34.52)	3.46(3.44)
Fed-MENU [13]		87.18	87.98(20.91)	92.34(13.89)	**81.22(11.04)**	4.45	7.10(11.13)	4.03(11.72)	**2.23(3.18)**
PIPO [22]	2D	86.13	94.15(3.46)	92.64(7.47)	71.59(10.53)	**1.68**	1.71(1.73)	**0.72(0.91)**	2.62(1.77)
Med* [33]		85.75	93.62(4.61)	92.97(6.92)	70.67(10.43)	2.14	**1.59(1.81)**	2.10(11.35)	2.72(1.99)
C2FNAS* [14]		81.68	85.55(7.43)	90.78(8.86)	68.73(14.16)	2.80	3.35(2.05)	1.10(3.37)	3.95(7.69)
DoDNet* [27]		87.50	94.08(3.53)	94.08(6.77)	74.33(10.78)	2.93	3.82(5.72)	1.53(6.70)	3.44(7.54)
Fed-MENU* [13]		86.45	89.55(4.81)	93.82(4.01)	75.99(9.68)	2.38	3.70(2.78)	1.02(1.26)	2.42(2.59)
FedIOD		**88.10**	**94.96(3.46)**	**94.20(5.91)**	75.15(9.57)	2.04	1.93(4.08)	1.96(10.98)	2.24(1.55)

$*$ represents approaches originally proposed for 3D medical imaging and re-implemented to 2D for additional evaluation.

pixels and augmented by random rotation, intensity change, and gamma augmentation. During inference, any unseen image is divided into a series of patches of $224\times224\times32$ pixels, and fed to the trained segmentation network. Patch-wise segmentation maps are assembled as final segmentation.

For in-federation evaluation on abdominal datasets, **LIVER**, **KIDNEY**, and **PANCREAS** were randomly split into training/validation/testing sets with a fixed ratio of 60%:10%:30%. For in-federation evaluation on cardiac datasets, ACDC was randomly and patient-wisely divided into three subsets (*i.e.*, clients), and each subset was randomly split into training/validation/testing sets following the same ratio as 60%:10%:30%.

FedIOD is compared against the state-of-the-art methods on partial label learning, including (1) one centralized method with specific loss functions (*i.e.*, PIPO [22]), (2) one centralized method with conditioned manipulation (*i.e.*, DoDNet [27]), (3) one centralized method with multi-path networks (*i.e.*, Med [33]), and (4) three federated methods (i.e., C2FNAS [14], MEloss [15], and Fed-MENU [13]).

D. In-Federation Evaluation on Abdominal Datasets

Quantitative comparison results of different approaches on in-federation evaluation are summarized in Table I. Despite the effectiveness of DoDNet [27] in centralized learning, it is sub-optimal compared to both MELoss [15] and C2FNAS [14] under 3D federated partial label learning. In contrast, 3D Fed-MENU outperforms other comparison methods on **PANCREAS** with at least an average increase of 6.07% in DSC, mainly benefiting from additional cross-slice information in 3D which is helpful to segment small-size organs like pancreas. On the contrary, its segmentation performance on large-size organs like liver and kidney is less competitive. Comparatively, FedIOD achieves the best results on both **LIVER** and **KIDNEY** and the second best results on **PANCREAS** among 2D approaches. On average, FedIOD outperforms other 2D approaches with at least an average increase of 0.6% in DSC and the SOTA 3D approach Fed-MENU [13] with an average increase of 0.92% in DSC.

In terms of shape preservation measured by ASD, all 2D approaches outperform 3D approaches, mainly due to a better balance across slices and organs. Specifically, PIPO [22] achieves the best performance, leading to better segmentation of large-size organs like LIVER and KIDNEY. Comparatively, FedIOD is the second-best approach slightly worse than PIPO but with better performance on small-size organs like PANCREAS. It is mainly due to the regularization through inter-organ dependency, which somewhat re-balances organs of various sizes. After visualizing learned self-attention matrices, we find pancreas regions are more frequently activated when building inter-organ dependency, leading to greater weights during regularization.

Qualitative results are illustrated in Fig. 2. In general, across different organs, Fed-MENU (3D) is likely to encounter the under-segmentation problem where some regions are mislabeled as background. For other comparison approaches, they may suffer from either under-segmentation or over-segmentation depending on the target slices, indicating their unstable segmentation performance. Comparatively, FedIOD achieves the most stable results across organs and slices, being more consistent with ground truth.

E. Out-of-Federation Evaluation on Abdominal Datasets

Quantitative comparison results on out-of-federation evaluation are summarized in Table II. When testing on a separate dataset with a different data distribution compared to the training datasets, all approaches encounter performance degradation. In terms of different organs, Kidney suffers the most performance degradation, indicating greater cross-domain variations. Among comparison approaches, 3D Fed-MENU [13] encounters the least performance drop with an average decrease of 2.69% benefiting from 3D cross-slice information. Among 2D approaches, FedIOD achieves the best overall performance, outperforming C2FNAS [14] and Fed-MENU [13] with an average increase of 2.11% and 1.78% in DSC respectively.

Similar to in-federation evaluation, in terms of shape preservation measured by ASD, most 2D approaches achieve better performance but still suffer from performance degradation given an unseen dataset. It is noticed that FedIOD outperforms all other approaches including PIPO [22] and Med [33].

Qualitative comparison results on out-of-federation evaluation are illustrated in Fig. 3. For liver segmentation, due to its relatively larger size, most approaches effectively segment

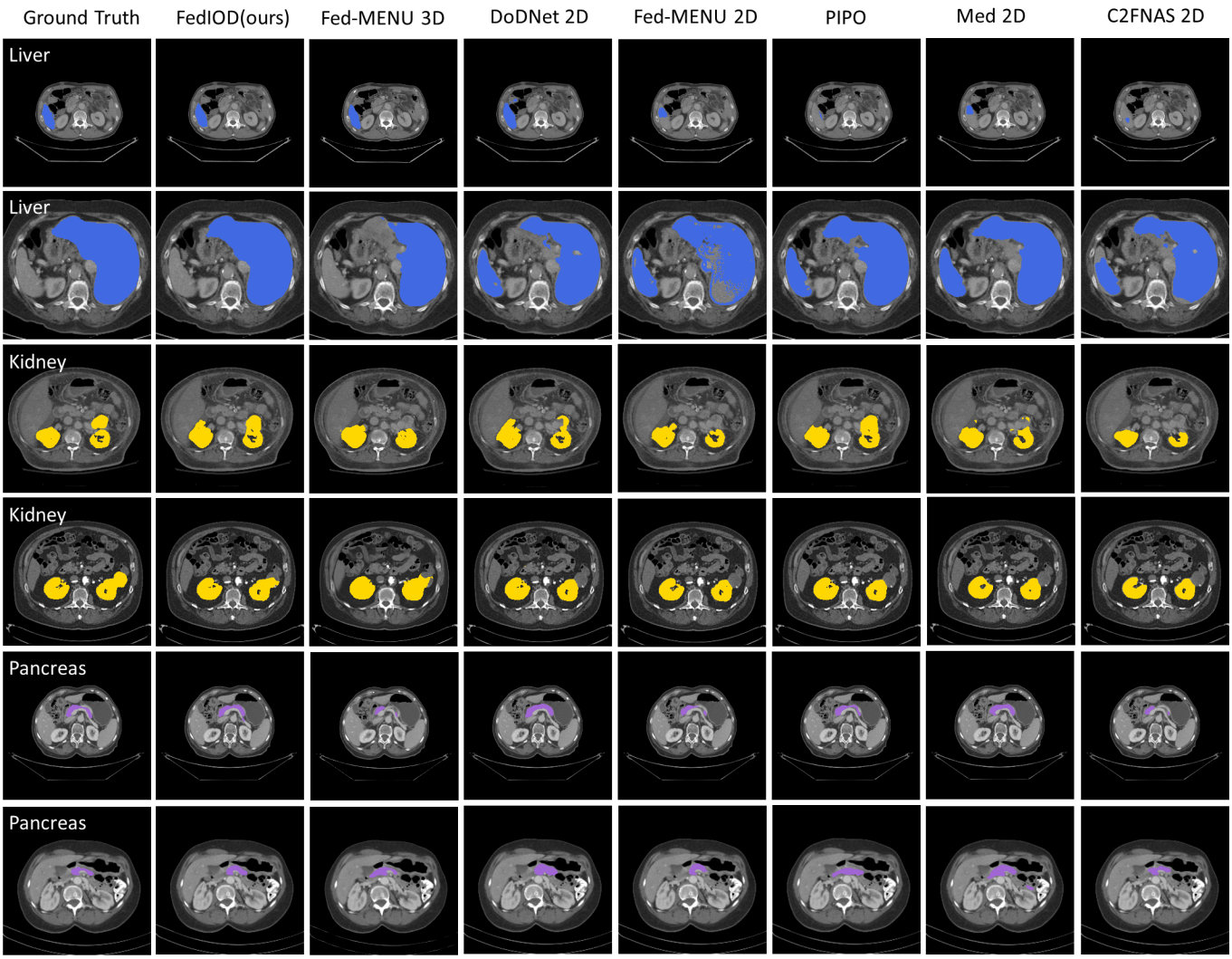

Fig. 2: Qualitative comparison results of different methods on the in-federation abdominal CT image datasets, including Fed-MENU (3D) [13], DoDNet (2D) [27], Fed-MENU (2D) [13], PIPO [22], Med (2D) [33], C2FNAS (2D) [14], and FedIOD.

TABLE II: Quantitative comparison of different methods on the out-of-federation abdominal CT image dataset (*i.e.*, BTCV). The best and second-best results are marked in bold and underlined. All comparison results are reproduced according to publicly-available source codes.

Method	Type	Out-of-federation DSC (Mean(SD) %)				Out-of-federation ASD (Mean(SD) mm)			
		Avg.	Liver	Kidney	Pancreas	Avg.	Liver	Kidney	Pancreas
MELoss [15]	3D	79.22	92.21(6.21)	84.18(19.75)	61.28(19.14)	4.82	**2.81(4.86)**	7.13(24.27)	4.53(5.57)
DoDNet [27]		78.34	86.82(9.61)	77.96(24.34)	70.24(11.61)	7.09	7.69(5.77)	8.29(19.11)	5.28(10.48)
Fed-MENU [13]		**84.49**	**93.38(7.42)**	82.09(20.06)	**78.00(7.03)**	4.40	3.87(6.88)	7.46(17.83)	**1.86(1.06)**
PIPO [22]	2D	79.25	91.29(7.38)	81.67(17.85)	64.79(12.39)	3.63	3.48(5.64)	**3.46(5.59)**	3.94(3.28)
Med* [33]		80.55	91.49(7.44)	84.43(15.15)	65.75(10.68)	3.63	3.13(5.32)	4.02(6.55)	3.72(3.26)
C2FNAS* [14]		81.53	88.18(7.48)	84.53(16.04)	71.87(10.19)	5.28	7.13(5.95)	4.92(10.15)	3.79(4.14)
DoDNet* [27]		80.24	92.79(5.20)	84.3(16.97)	63.63(16.52)	5.22	5.10(6.81)	4.70(10.07)	5.86(6.20)
Fed-MENU* [13]		81.86	88.06(6.33)	**85.22(16.93)**	72.31(10.12)	4.98	5.34(7.06)	5.91(15.73)	3.69(3.88)
FedIOD		83.64	92.86(7.29)	85.09(15.18)	72.98(9.37)	**3.23**	3.32(5.15)	3.83(6.98)	2.54(1.64)

∗ represents approaches originally proposed for 3D medical imaging and re-implemented to 2D for additional evaluation.

the liver regions and some of them suffer from the under-segmentation problem. Comparatively, different approaches vary significantly in kidney segmentation, which is consistent with the quantitative results in Table I. Specifically, DoDNet completely misses the kidney regions while other comparison methods produce more false negatives (*i.e.*, under-segmentation). In terms of pancreas segmentation, except for Fed-MENU (3D), all other comparison approaches encounter the under-segmentation problem. In general, FedIOD achieves the best segmentation performance, leading to more complete

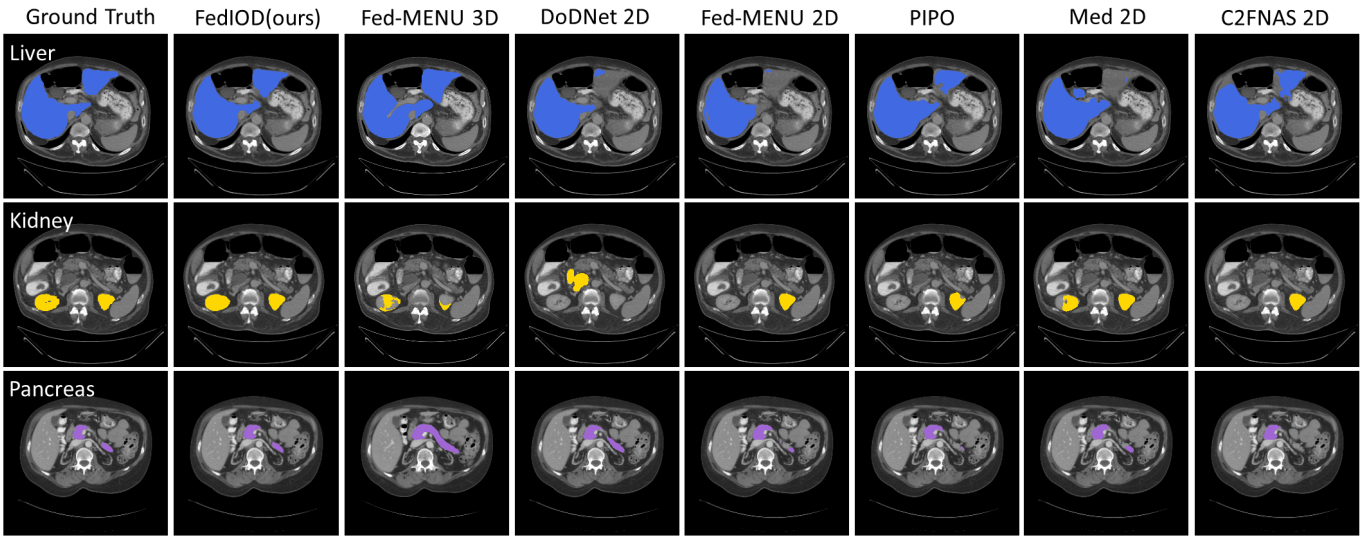

Fig. 3: Qualitative comparison results of different methods on the out-of-federation abdominal CT image datasets, including Fed-MENU (3D) [13], DoDNet (2D) [27], Fed-MENU (2D) [13], PIPO [22], Med (2D) [33], C2FNAS (2D) [14], and FedIOD.

TABLE III: Quantitative comparison of different methods on the in-federation cardiac MRI image dataset (*i.e.*, ACDC). The best and second-best results are marked in bold and underlined. All comparison results are reproduced according to publicly-available source codes.

Method	Type	In-federation DSC (Mean(SD) %)				In-federation ASD (Mean(SD) mm)			
		RV	MYO	LV	Avg.	RV	MYO	LV	Avg.
Fed-MENU [13]	3D	79.94(10.41)	80.03(6.34)	89.21(4.69)	83.06	4.54(4.84)	1.44(1.03)	1.20(0.60)	2.39
PIPO [22]	2D	76.61(17.47)	79.47(5.01)	92.19(5.10)	82.76	2.93(2.24)	1.35(0.88)	1.17(0.82)	1.82
Med* [33]		78.17(11.12)	68.59(8.00)	89.19(5.88)	78.65	2.57(1.23)	2.15(1.37)	3.11(2.79)	2.61
C2FNAS* [14]		85.31(10.65)	84.39(2.85)	93.42(4.76)	87.71	**1.50(1.11)**	**0.71(0.39)**	0.69(0.51)	**0.97**
MELoss* [15]		57.87(23.25)	60.54(11.47)	89.73(6.09)	69.38	5.83(4.89)	2.51(1.53)	1.26(0.86)	3.20
DoDNet* [27]		83.98(10.95)	**85.36(3.14)**	93.50(5.28)	87.61	3.05(4.18)	1.62(1.19)	1.40(2.18)	2.02
FedMENU* [13]		86.48(9.72)	83.15(6.05)	91.28(5.30)	86.97	1.59(1.19)	0.97(0.75)	0.69(0.33)	1.08
FedIOD		**88.07(7.70)**	83.41(4.01)	**93.96(3.47)**	**88.48**	**1.50(1.11)**	1.25(0.80)	**0.62(0.50)**	1.13

∗ represents approaches originally proposed for 3D medical imaging and re-implemented to 2D for additional evaluation.

organ segmentation with better shape preservation.

F. In-Federation Evaluation on Cardiac Datasets

In addition to abdominal multi-organ segmentation, we further conduct cardiac multi-organ segmentation under in-federation evaluation. Quantitative comparison results are summarized in Table III. One interesting observation is that 3D Fed-MENU under-performs 2D approaches. It is because ACDC is anisotropic containing less cross-slice information which in turn affects the convergence of 3D Fed-MENU. As a result, 2D Fed-MENU consistently outperforms 3D Fed-MENU across all organs with large margins. Comparatively, though FedIOD under-performs DoDNet and C2FNAS on MYO segmentation, it achieves better overall segmentation performance with an average increase of 0.87% and 0.77% in DSC compared to DoDNet and C2FNAS respectively.

In terms of ASD performance, 2D approaches achieve better performance compared to 3D approaches, being consistent with quantitative results in Table I. Specifically, C2FNAS* [14] achieves the best overall performance while FedIOD achieves the best performance on the segmentation of RV and LV approaching to the overall performance of C2FNAS* and FedMENU [13].

Qualitative comparison results on cardiac multi-organ segmentation are illustrated in Fig. 4. In general, cardiac organs are much smaller than those abdominal organs as shown in Figs. 2 and 3. Such small-size organs are more challenging for segmentation, resulting in more performance variations across different approaches, especially for PIPO and Med2D. In terms of DoDNet and Fed-MENU (2D), both of them encounter under-segmentation, especially for RV and LV, which are consistent with the quantitative results in Table III. Comparatively, FedIOD achieves the best segmentation performance, significantly outperforming other approaches on the segmentation of RV and LV.

The above comparison results on both abdominal and cardiac multi-organ segmentation validate the effectiveness of FedIOD in learning from partial labels and building complete inter-organ dependency for pseudo full-organ supervision.

G. Evaluation on the Effects of Backbone

As both Med [33] and PIPO [22] adopt multi-decoder designs like FedIOD, we modify Med [33] and PIPO [22]

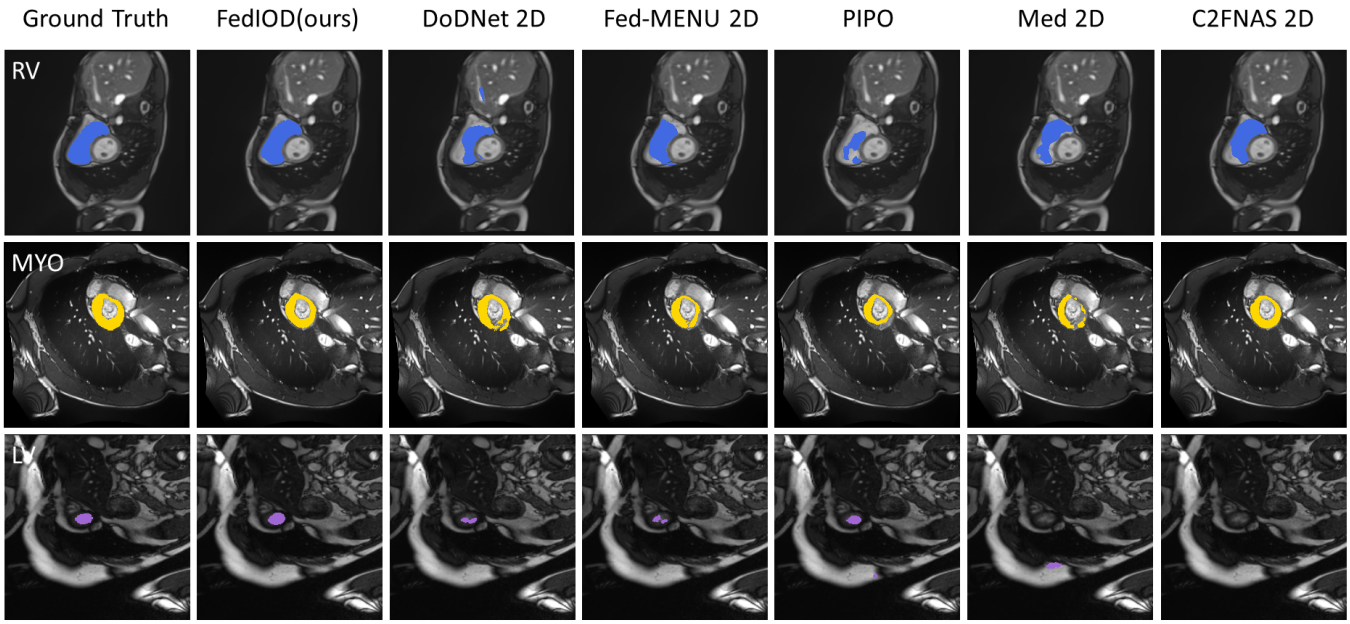

Fig. 4: Qualitative comparison results of different methods on the in-federation cardiac MRI image datasets, including DoDNet [27], Fed-MENU (2D) [13], PIPO [22], Med (2D) [33], C2FNAS (2D) [14], and FedIOD.

TABLE IV: Quantitative comparison of centralized methods (*i.e.*, PIPO [22] and Med [33]) with the same backbone as FedIOD on the in-federation (*i.e.*, **LIVER**, **KIDNEY**, and **PANCREAS**) and out-of-federation (*i.e.*, BTCV) abdominal CT image datasets. The best and second-best results are marked in bold and underlined. $+T$ means introducing the same transformer module as FedIOD.

Method	Type	In-federation DSC (Mean %)				Out-of-federation DSC (Mean %)			
		LIVER	**KIDNEY**	**PANCREAS**	Avg.	Liver	Kidney	Pancreas	Avg.
Med	2D	93.62	92.97	70.67	85.75	91.49	84.43	65.75	<u>80.55</u>
Med + T		93.76	92.59	68.97	85.10	<u>91.83</u>	83.05	64.98	79.95
PIPO		<u>94.15</u>	92.64	71.59	86.13	91.26	81.67	64.79	79.25
PIPO + T		91.82	<u>93.63</u>	**75.69**	<u>87.05</u>	84.67	**85.76**	<u>70.97</u>	80.47
FedIOD		**94.96**	**94.20**	<u>75.15</u>	**88.10**	**92.86**	<u>85.09</u>	**72.98**	**83.64**

TABLE V: Quantitative comparison of FedIOD and PIPO [22] based on single-layer organ-specific decoders on the in-federation (*i.e.*, **LIVER**, **KIDNEY**, and **PANCREAS**) and out-of-federation (*i.e.*, BTCV) abdominal CT image datasets. The best and the second-best results are marked in bold and underlined. *single-layer* means changing the organ-specific decoders of FedIOD to a single-layer architecture similar to PIPO [22].

Method	Type	In-federation DSC (Mean %)				Out-of-federation DSC (Mean %)			
		LIVER	**KIDNEY**	**PANCREAS**	Avg.	Liver	Kidney	Pancreas	Avg.
PIPO [22]	2D	<u>94.15</u>	<u>92.64</u>	71.59	86.13	<u>91.26</u>	81.67	64.79	79.25
single-layer FedIOD		93.77	92.25	<u>73.85</u>	<u>86.62</u>	90.40	<u>84.97</u>	<u>69.18</u>	<u>81.52</u>
FedIOD		**94.96**	**94.20**	**75.15**	**88.10**	**92.86**	**85.09**	**72.98**	**83.64**

by using the same backbone as FedIOD for comparison as summarized in Table IV. For Med, simply adding a transformer is useless mainly due to its less powerful encoder. In contrast, introducing a transformer to PIPO brings an average increase of 1.08% in Dice. It should be noted that FedIOD consistently outperforms Med and PIPO with the same backbone, demonstrating the effectiveness of FedIOD's designs.

H. Evaluation on Lightweight Decoder in FedIOD

According to FedIOD, given more organs, more organ-specific decoders are needed, resulting in higher model complexity. To validate the effectiveness of FedIOD, we further reduce the organ-specific decoder into a single-layer architecture similar to the decoder part of PIPO [22], and make a comparison with PIPO [22] under the same settings as summarized in Table V. Though adopting a lightweight decoder architecture unavoidably encounters performance degradation, it still outperforms PIPO [22] without additional tuning, demonstrating the effectiveness of FedIOD. In future work, we will explore to address federated partial label learning through a unified segmentation backbone.

I. In-Federation Evaluation on Tumor Segmentation

As summarized in Table VI, when treating organ tumors as separate labels, there exists an average of 5% decrease in

TABLE VI: Quantitative performance evaluation of FedIOD with and without treating organ-specific tumors as separate labels. The best results are marked in bold.

Setting	In-federation DSC (Mean %)			
	Avg.	**LIVER**	**KIDNEY**	**PANCREAS**
organ	**88.10**	**94.96**	**94.20**	**75.15**
organ + tumor	83.63	91.49	90.50	68.90

TABLE VII: Quantitative efficiency comparison of different methods in terms of the average Dice (*i.e.*, DSC), the parameter amount (*i.e.*, *Param*) measured in millions, and GFLOPs on the in-federation abdominal CT image datasets. The best and second-best results are marked in bold and underlined. All comparison results are reproduced according to publicly-available source codes. GFLOPs is the abbreviation of floating-point operations per second measured in billions.

Method	Type	DSC (Mean %)	*Param* (M)	GFLOPs
MELoss [15]	3D	84.58	23.53	441.90
DoDNet [27]		83.54	17.31	348.51
Fed-MENU [13]		87.18	29.51	854.88
PIPO [22]	2D	86.13	27.91	44.44
Med* [33]		85.75	24.32	28.29
C2FNAS* [14]		81.68	<u>7.88</u>	18.22
DoDNet* [27]		<u>87.50</u>	**5.94**	<u>11.03</u>
Fed-MENU* [13]		86.45	20.68	40.55
FedIOD		**88.10**	11.15	**8.14**

$*$ represents approaches originally proposed for 3D medical imaging and re-implemented to 2D for additional evaluation.

Dice, as it will worsen cross-client variations, especially in target scales. Consequently, it is more challenging to balance different clients for federated aggregation. In addition, as FedIOD adopts a shared encoder for both organs and tumors, features of different targets may distract each other, which in turn results in sub-optimal feature learning. For joint organ and tumor segmentation, more specific designs shall be introduced to FedIO as future work.

J. Model Complexity Analysis

Quantitative results of model complexity of different methods on the in-federation abdominal CT image datasets are summarized in Table VII. Specifically, DoDNet* [27] has the fewest model parameters, and C2FNAS* [14] is the second-most lightweight model. It is noticed that, though FedIOD has a transformer module introducing more parameters, it still outperforms most CNN-based models. More importantly, FedIOD noticeably outperforms both DoDNet* and C2FNAS* on segmentation while requiring the least computational resources, making it more deployable in clinical applications.

As summarized in Table VII, for 3D approaches, model complexity seems proportional to model performance. This is because volumetric data provides more information to explore. Comparatively, increasing model complexity would not necessarily bring performance improvement for 2D approaches. For instance, PIPO [22] is much larger than FedIOD but its performance is inferior. It indicates that increasing model complexity may result in severe redundancy which in turn affects model performance.

V. Discussion

A. Component-Wise Ablation Study

To validate the component-wise effectiveness of FedIOD, each component is progressively introduced to the baseline (*i.e.*, $Enc + Dec$) for evaluation and comparison on both in-federation and out-of-federation evaluation as summarized in Table VIII. Compared to the quantitative results in Table I, even based on the baseline architecture, FedIOD outperforms most approaches. Specifically, it outperforms other 2D comparison approaches except for DoDNet on in-federation evaluation while surpassing all 2D approaches on out-of-federation evaluation, validating the superiority of the single-encoder-multi-decoder architecture. When introducing the transformer module T to re-weight skip connections, the shared encoder Enc would focus more on organ-related features and in turn somewhat alleviate the negative impact of partial labels on being biased to certain organs. As a result, consistent performance improvements are achieved across organs on both in-federation and out-of-federation evaluation. Introducing regularization to supervise the training of self-attention matrices in each client benefits the most. Through refining the self-attention matrices in each client, the weights of skip connections, together with the features of Enc, are further balanced to guide Enc for complete organ-related representation learning. As a result, the segmentation performance of all organs across different evaluation settings is effectively improved with large margins. One interesting observation is that such results are even better than the complete framework (*i.e.*, FedIOD) on in-federation evaluation. One possible reason is that, under in-federation evaluation, the federated model's performance is more likely to be affected by certain clients/datasets, which in turn degrade its generalizability. It explains why the complete framework achieves much segmentation results on out-of-federation evaluation as stated in the last row of Table VIII.

B. Self-Attention Visualization

To validate the effectiveness of self-attention aggregation for regularization, we introduce the transformer module T to the baseline model (*i.e.*, *Enc* + *Dec*) denoted as *Enc* + *Dec* + T but exclude self-attention aggregation for comparison with FedIOD. As illustrated in Fig. 5, whether aggregating self-attention or not, each local model not only focuses on labeled organs but also activates unlabeled organs. For instance, the local model trained with pancreas labels also pays attention to the adjacent kidney. It is consistent with our analysis that T helps capture inter-organ dependency even with only partial labels. Comparing the self-attention matrices of global models learned by *Enc* + *Dec* + T and FedIOD, we find that inter-organ dependency learned by FedIOD is more complete, indicating the value of self-attention aggregation and the following regularization for complete full-organ feature extraction.

It is noticed that not all attention matrices clearly capture inter-organ dependency. For instance, for LIVER, attention matrices mainly focus on the liver regions where the attention scores of other regions are much lower. This is because LIVER is relatively easy to segment and less relies on global information. Comparatively, for KIDNEY and PANCREAS, more

TABLE VIII: Component-wise ablation study of FedIOD on the in-federation and out-of-federation abdominal CT image datasets. Here, Reg denotes the regularization losses in Section III.E and DA is short for the dynamic aggregation in Section III.F. The best results are marked in bold.

Components					In-federation DSC (Mean %)				Out-of-federation DSC (Mean %)			
Enc	*Dec*	*T*	*Reg*	*DA*	**LIVER**	**KIDNEY**	**PANCREAS**	Avg.	Liver	Kidney	Pancreas	Avg.
✓	✓				93.47	93.76	74.23	87.16	92.19	84.32	69.25	81.92
✓	✓	✓			93.98	94.24	74.29	87.50	91.90	84.64	71.58	82.71
✓	✓	✓	✓		94.76	**94.30**	**75.28**	**88.11**	92.76	84.90	72.27	83.31
✓	✓	✓	✓	✓	**94.96**	94.20	75.15	88.10	**92.86**	**85.09**	**72.98**	**83.64**

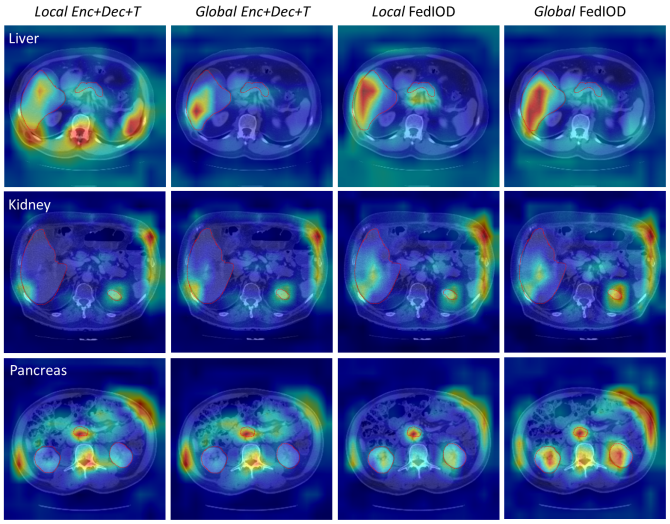

Fig. 5: Visualized self-attention matrices of *Enc* + *Dec* + *T* and FedIOD on the in-federation abdominal CT image datasets, including local and global models of both. Here, *Enc* + *Dec* + *T* denotes the baseline model (*i.e. Enc* + *Dec*) with the transformer module *T* but without self-attention aggregation for regularization.

regions from other organs are activated in attention matrices, as the locations of other organs can provide additional guidance for localization and segmentation. In other words, challenging organs are more likely to build inter-organ dependency in attention matrices. In addition, it is observed that even in the attention matrices of LIVER certain organ regions are assigned with greater weights compared to background. This is because background varies more significantly than organs in medical volumetric data, making it more likely to build stable dependency on other organs rather than the background. After aggregating attention matrices from various slices and clients, dependency on non-organ regions would be further suppressed, making inter-organ dependency more evident for regularization.

VI. Conclusion

In this paper, we present a novel single-encoder-multi-decoder framework named FedIOD for multi-organ segmentation from partial labels in federated learning. The key idea is to build pseudo full-organ labels by exploring inter-organ dependency. It is based on the observation that, in clinical practice, segmenting each organ relies on the prior knowledge/information of other organs, which can be effectively captured by self-attention matrices learned by a transformer module. Through dynamically aggregating client-wise self-attention matrices and using them for local supervision, partial labels of clients are transformed into pseudo full-organ labels to train the shared encoder for complete organ-relevant feature extraction. Extensive experiments on widely-used datasets demonstrate the superiority of FedIOD against the state-of-the-art approaches under various federated learning settings.

References

[1] J. Starekova, D. Hernando, P. J. Pickhardt, and S. B. Reeder, "Quantification of liver fat content with CT and MRI: State of the art," *Radiology*, vol. 301, no. 2, pp. 250–262, 2021.

[2] A. A. Borhani, *et al.*, "Imaging evaluation of living liver donor candidates: Techniques, protocols, and anatomy," *RadioGraphics*, vol. 41, no. 6, pp. 1572–1591, 2021.

[3] H. P. Clark, W. F. Carson, P. V. Kavanagh, C. P. Ho, P. Shen, and R. J. Zagoria, "Staging and current treatment of hepatocellular carcinoma," *Radiographics*, vol. 25, pp. S3–S23, 2005.

[4] S. J. Westra, J. Hurteau, A. Galindo, M. F. McNitt-Gray, M. I. Boechat, and H. Laks, "Cardiac electron-beam CT in children undergoing surgical repair for pulmonary atresia," *Radiology*, vol. 213, no. 2, pp. 502–512, 1999.

[5] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. MICCAI*, 2015, pp. 234-241.

[6] C. Zotti, Z. Luo, A. Lalande, and P. Jodoin, "Convolutional neural network with shape prior applied to cardiac MRI segmentation," *IEEE J. Biomed. Health Inform.*, vol. 23, no. 3, pp. 1119–1128, 2018.

[7] F. Isensee *et al.*, "nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation," *Nat. Methods*, vol. 18, no. 2, pp. 203-211, 2021.

[8] S. Zheng *et al.*, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proc. CVPR*, 2021, pp. 6881-6890.

[9] X. Lin, L. Yu, K. -T. Cheng, and Z. Yan, "BATFormer: Towards boundary-aware lightweight transformer for efficient medical image segmentation," *IEEE J. Biomed. Health Inform.*, 2023.

[10] J. M. J. Valanarasu, P. Oza, I. Hacihaliloglu, and V. M. Patel, "Medical transformer: Gated axial-attention for medical image segmentation," 2021, *arXiv:2102.10662*.

[11] B. McMahan *et al.*, "Communication-efficient learning of deep networks from decentralized data," in *Proc. AISTAT*, 2017, pp. 1273-1282.

[12] Z. Yan, J. Wicaksana, Z. Wang, X. Yang, and K. -T. Cheng, "Variation-aware federated learning with multi-source decentralized medical data," *IEEE J. Biomed. Health Informat.*, vol. 25, no. 7, pp. 2615-2628, 2021.

[13] X. Xu and P. Yan, "Federated multi-organ segmentation with partially labeled data," 2022, *arXiv:2206.07156*.

[14] C. Shen *et al.*, "Joint multi organ and tumor segmentation from partial labels using federated learning," in *Proc. MICCAI FAIR*, 2022.

[15] P. Liu, M. Sun, and S. K. Zhou, "Multi-site organ segmentation with federated partial supervision and site adaptation," 2023, *arXiv:2302.03911*.

[16] T. Durand *et al.*, "Learning a deep convnet for multi-label classification with partial labels." in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019.

[17] D. Nanqing *et al.*, "Revisiting vicinal risk minimization for partially supervised multi-label classification under data scarcity." in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[18] D. Nanqing *et al.*, "Federated partially supervised learning with limited decentralized medical images." *IEEE Transactions on Medical Imaging*, 2022.
[19] Y. Xiaotong *et al.*, "Self-distillation and self-supervision for partial label learning." *Pattern Recognition*, 2024.
[20] N. Roulet, D. F. Slezak, and E. Ferrante, "Joint learning of brain lesion and anatomy segmentation from heterogeneous datasets," in *Proc. MIDL*, 2019.
[21] Y. Zhou *et al.*, "Prior-aware neural network for partially-supervised multi-organ segmentation," in *Proc. ICCV*, 2019.
[22] X. Fang and P. Yan, "Multi-organ segmentation over partially labeled datasets with multi-scale feature abstraction," *IEEE Trans. Med. Imag.*, vol. 39, no. 11, pp. 3619-3629, 2020.
[23] N. Natarajan *et al.*, "Learning with noisy labels," in *Proc. NeurIPS*, 2013.
[24] L. Fidon *et al.*, "Label-set loss functions for partial supervision: Application to fetal brain 3D MRI parcellation," in *Proc. MICCAI*, 2021, pp. 647-657.
[25] G. Shi *et al.*, "Marginal loss and exclusion loss for partially supervised multi-organ segmentation," *Med. Image Anal.*, vol. 70, p. 101979, 2021.
[26] K. Dmitriev and A. E. Kaufman, "Learning multi-class segmentations from single-class datasets," in *Proc. CVPR*, 2019, pp. 9501-9511.
[27] J. Zhang *et al.*, "DoDNet: Learning to segment multi-organ and tumors from multiple partially labeled datasets," in *Proc. CVPR*, 2021, pp. 1195-1204.
[28] K. Zhang and X. Zhuang, "Deep compatible learning for partially-supervised medical image segmentation," 2022, *arXiv:2206.09148*.
[29] J. -Y. Zhu *et al.*, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. ICCV*, 2017.
[30] N. Dong *et al.*, "Towards robust partially supervised multi-structure medical image segmentation on small-scale data," *Appl. Soft Comput.*, vol. 114, p. 108074, 2022.
[31] R. Huang *et al.*, "Multi-organ segmentation via co-training weight-averaged models from few-organ datasets," in *Proc. MICCAI*, 2020, pp. 146-155.
[32] L. Zhang *et al.*, "Unsupervised ensemble distillation for multi-organ segmentation," in *Proc. ISBI*, 2022, pp. 1-5.
[33] S. Chen, K. Ma, and Y. Zheng, "Med3d: Transfer learning for 3d medical image analysis," 2019, *arXiv:1904.00625*.
[34] O. Petit, N. Thome, and L. Soler, "Iterative confidence relabeling with deep ConvNets for organ segmentation with partial labels," *Comput. Med. Imaging Graph.*, vol. 91, p. 101938, 2021.
[49] N. Heller *et al.*, "The state of the art in kidney and kidney tumor segmentation in contrast-enhanced CT imaging: Results of the KiTS19 challenge," *Med. Image Anal.*, p. 101821, 2020.
[35] P. Liu *et al.*, "Learning incrementally to segment multiple organs in a CT image," 2022, *arXiv:2203.02100*.
[36] Y. Zhou *et al.*, "Uncertainty-aware incremental learning for multi-organ segmentation," 2021, *arXiv:2103.05227*.
[37] B. Yingbin *et al.*, "Understanding and improving early stopping for learning with noisy labels." in *Neural Information Processing Systems*, 2021.
[38] S. Zheng *et al.*, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proc. CVPR*, 2021, pp. 6881-6890.
[39] J. Chen *et al.*, "Transunet: Transformers make strong encoders for medical image segmentation," 2021, *arXiv:2102.04306*.
[40] R. Li *et al.*, "TransCAM: Transformer attention-based CAM refinement for Weakly supervised semantic segmentation," *J. Vis. Commun. Image Represent.*, p. 103800, 2023.
[41] K. Li *et al.*, "Tell me where to look: Guided attention inference network," in *Proc. CVPR*, 2018.
[42] H. Wu *et al.*, "Self-supervised models are good teaching assistants for vision transformers," in *Proc. ICML*, 2022.
[43] M. Caron *et al.*, "Emerging properties in self-supervised vision transformers," in *Proc. CVPR*, 2021.
[44] Y. Tang *et al.*, "Scop: Scientific control for reliable neural network pruning," in *Proc. NeurIPS*, 2020, pp. 10936-10947.
[45] Y. Hao *et al.*, "Self-attention attribution: Interpreting information interactions inside transformer," in *Proc. AAAI*, 2021.
[46] M. Sundararajan, T. Ankur, and Q. Yan, "Axiomatic attribution for deep networks," in *Proc. ICML*, 2017.
[47] P. Bilic *et al.*, "The liver tumor segmentation benchmark (lits)," 2019, *arXiv:1901.04056*.
[48] N. Heller *et al.*, "The KiTS19 challenge data: 300 kidney tumor cases with clinical context, CT semantic segmentations, and surgical outcomes," 2019, *arXiv:1904.00445*.
[50] M. Antonelli *et al.*, "The medical segmentation decathlon," 2021, *arXiv:2106.05735*.
[51] B. Landman, Z. Xu, J. Igelsias, M. Styner, T. Langerak, and A. Klein, "MICCAI multi-atlas labeling beyond the cranial vault–workshop and challenge," in *Proc. MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge*, vol. 5, 2015.
[52] O. Bernard *et al.*, "Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: Is the problem solved?" *IEEE Trans. Med. Imag.*, vol. 37, no. 11, pp. 2514-2525, 2018.