FedSeg: Class-Heterogeneous Federated Learning for Semantic Segmentation

Jiaxu Miao, Zongxin Yang, Leilei Fan, Yi Yang[†]
ReLER, CCAI, Zhejiang University
{jiaxumiao, yangzongxin, 22151297, yangyics}@zju.edu.cn

Abstract

Federated Learning (FL) is a distributed learning paradigm that collaboratively learns a global model by multiple clients with data privacy-preserving. Although many FL algorithms have been proposed for classification tasks, few works focus on more challenging semantic segmentation tasks, especially in the class-heterogeneous FL situation. Compared with classification, the issues from heterogeneous FL for semantic segmentation are more severe: (1) Due to the non-IID distribution, different clients may contain inconsistent foreground-background classes, resulting in divergent local updates. (2) Class-heterogeneity for complex dense prediction tasks makes the local optimum of clients farther from the global optimum. In this work, we propose FedSeg, a basic federated learning approach for class-heterogeneous semantic segmentation. We first propose a simple but strong modified cross-entropy loss to correct the local optimization and address the foreground-background inconsistency problem. Based on it, we introduce pixel-level contrastive learning to enforce local pixel embeddings belonging to the global semantic space. Extensive experiments on four semantic segmentation benchmarks (Cityscapes, CamVID, PascalVOC and ADE20k) demonstrate the effectiveness of our FedSeg. We hope this work will attract more attention from the FL community to the challenging semantic segmentation federated learning.

1. Introduction

Semantic segmentation is the task of assigning a unique semantic label to every pixel in a given image, which is a fundamental research topic in computer vision and has many potential applications, such as autonomous driving, image editing and robotics [30]. Training a semantic segmentation model usually needs vast of data with pixel-level annotations, which is extremely hard to acquire. Collaborative training on multiple clients is a feasible way to solve the problem. However, collaborative training has the risk of leaking sensitive information. For example, for the autonomous driving task, the training images may include private information such as where the user arrived, where the user lives and what the user's house looks like. Thus, a privacy-preserving collaborative training method is requisite for semantic segmentation.

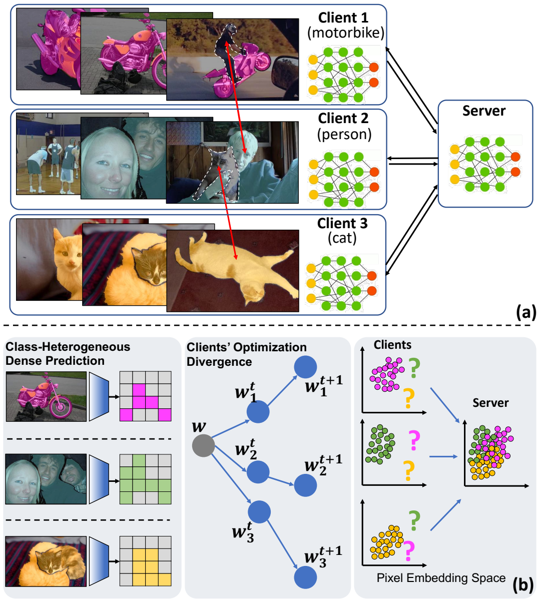

Figure 1. (a) The foreground-background inconsistency for class-heterogeneous semantic segmentation. (b) Local optimization divergence problem for the heterogeneous dense prediction task.

Federated Learning (FL) [31] is an emerging distributed machine learning paradigm that jointly trains a shared global model by multiple clients without exchanging their raw data. FedAvg [31] is a basic FL algorithm that learns local models with raw data on clients separately while aggregating weights to a global model on a server. One key problem of FL is the statistical heterogeneity of data distribution among different clients. Many recent FL algorithms [1, 21, 22, 26, 32] are proposed to tackle the problem. However, most of them evaluate their methods on classification, while few works focus on more challenging semantic segmentation. Although some federated learn-

[†] Corresponding author.

ing approaches [17, 29, 46, 52] for medical image segmentation have been proposed, they mainly address the simple foreground-background segmentation and cannot solve the class-heterogeneous problem for semantic segmentation with a variety of object classes. A recent FL approach, FedDrive [14], evaluates FL methods on an autonomous driving semantic segmentation dataset, Cityscapes [9]. However, FedDrive [14] focuses on domain heterogeneity (images from different cities) while ignoring the more challenging class-heterogeneous problem.

In this paper, we focus on class-heterogeneous federated learning for semantic segmentation, which has specific and more severe issues compared with classification. First, images for semantic segmentation are more complex, and pixel-level annotation is extremely time-consuming. Clients usually annotate the objects of frequent classes and ignore the rare ones. Due to the non-IID (non-Independent Identically Distribution) data distribution of different clients, classes ignored by one client may be foreground classes in another client. For example, in Fig. 1 (a), the ignored class "person" in Client 1 is annotated in Client 2. The foreground-background inconsistency across clients leads to divisive optimization directions and degrades the capability of the aggregated global model. Second, as shown in Fig. 1 (b), even if there is no foreground-background inconsistency, for non-IID distribution, complex dense prediction makes the local optimization direction diverging farther to the global optimum compared with classification tasks, resulting in poor convergence. From the perspective of the pixel embedding space, the local update in each client cannot learn the relative positions of different semantic classes in the pixel embedding space, leading to the confounded embedding space after global aggregation.

In this paper, we propose a new federated learning method for semantic segmentation, FedSeg, to address the above issues. A standard objective function for semantic segmentation is the cross-entropy (CE) loss which takes effect on foreground pixels and ignores the background pixels. For FL with non-IID data distribution, it makes the learned local optimum away from the global optimum. Thus, we propose a simple but strong baseline, a modified cross-entropy loss, by aggregating the probabilities of background classes. The modified loss corrects "client drift" in local updates and alleviates the foreground-background inconsistency problem. Then we further introduce a local-to-global pixel-level contrastive learning loss to enforce the local pixel embedding space close to the global semantic space, improving the convergence of the global model.

Extensive experiments on four semantic segmentation datasets (Cityscapes [9], CamVID [3], PascalVOC [13] and ADE20k [63]) are conducted to evaluate the effectiveness of our FedSeg. Experimental results show that the simple modified cross-entropy loss significantly improves the segmentation quality. Based on it, our proposed local-to-global pixel contrastive learning consistently improves the segmentation performance compared with previous FL algorithms [1, 22, 26, 31].

To summarize, the contributions of this paper are as follows:

- We systematically investigate federated learning for the semantic segmentation task with a variety of classes, particularly the class-heterogeneous problem.
- We propose a strong baseline with a simple modified CE loss and a local-to-global metrics learning method to alleviate the class distribution drift problem across clients.
- We provide benchmarks on four semantic segmentation datasets to evaluate our FedSeg for the semantic segmentation FL problem. We hope this work will motivate the FL community to further study the federated learning problem for challenging semantic segmentation tasks.

2. Related Work

2.1. Federated Learning

Federated Learning (FL) has attracted more and more attention in recent years, which provides a decentralized machine learning paradigm with data privacy-preserving [1, 10, 11, 15, 18, 19, 21, 26, 31, 32, 36, 38, 44, 60, 61]. FedAvg [31] is the earliest federated learning approach, which optimizes the local model with its individual data and simply conducts weighted averaging to aggregate weights in the server. Recent works [21, 26] show that the client drifts during clients' updates caused by non-IID data damage convergence in heterogeneous settings. To solve the statistic heterogeneity problem, FedProx [22] proposes to add a proximal regularization term on the local model, which restricts the updated local parameter close to the global model and prevents gradient divergence. Scaffold [21] introduces a variance reduction strategy to correct the drifted local update. FedDyn [1] proposes a dynamic regularizer for each device to align global and local objectives. MOON [26] utilizes a local model contrastive loss to push the current local representation closer to the global representation while pushing the current local representation away from the local representation of the previous round. However, all these methods [1, 21, 22, 26] evaluate their models only on the classification task while ignoring the more challenging semantic segmentation task. FedProx [22] and MOON [26] restrict the local model close to the global model from the entire model perspective, which cannot address the dense prediction problem. In this paper, we propose FedSeg to regularize the local update in a more fine-grained way.

Recently, some FL algorithms for the medical image segmentation task are proposed [6, 12, 17, 20, 22, 27, 29, 37, 39, 41, 46, 48, 49, 52, 57, 65]. For instance, FedDG [29] pro-

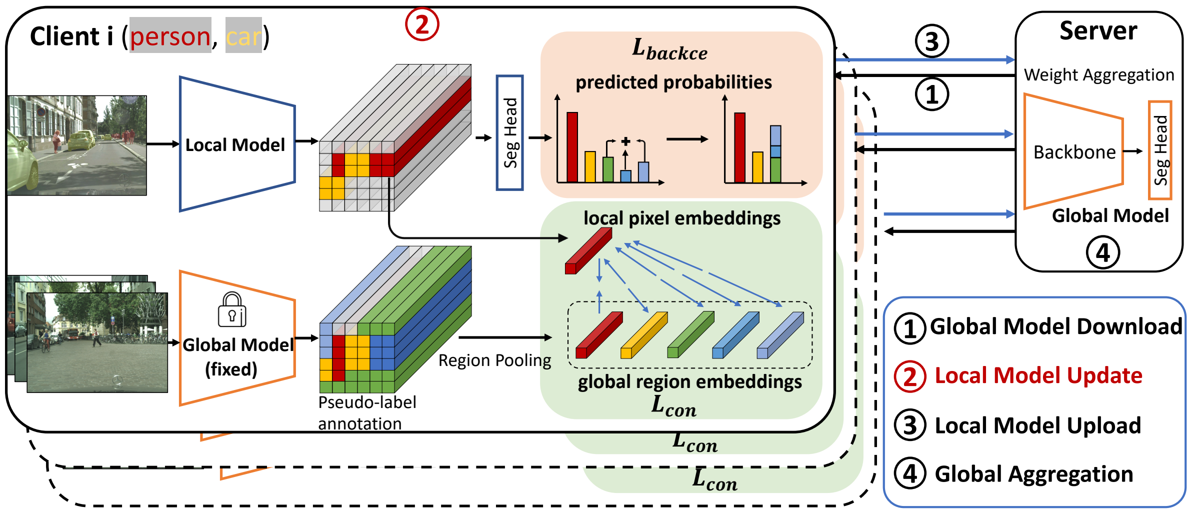

Figure 2. A standard federated learning framework includes four processes per round, *i.e.*, (1) Global Model Download, (2) Local Model Update, (3) Local Model Upload and (4) Global Aggregation. We modify the Local Model Update process by proposing two objective functions, $\mathcal{L}_{backce}$ and $\mathcal{L}_{con}$, without extra information exchanging.

poses to exchange amplitude spectrum after Fourier transform across clients for domain generalization. FedSM [52] adopts a model selector to decide the closest model/data distribution for any test data. However, the medical image segmentation task is usually foreground-background segmentation without a large variety of semantics. This paper focuses on more generic semantic segmentation tasks. Some recent FL approaches for semantic segmentation [4, 14, 16, 40, 56] evaluate FL methods on semantic segmentation datasets [9, 13], *e.g.*, FedDrive [14]. However, they [4, 14, 16, 40, 56] focus on domain heterogeneity while ignoring more challenging class heterogeneity for semantic segmentation.

2.2. Semantic Segmentation

Starting from the fully convolutional networks [30], many subsequent FCN-based models have greatly advanced semantic segmentation [7, 23–25, 28, 33–35, 53–55, 62, 64], mainly including Deeplab [7], PSPNet [62], HRNet [43], OCRNet [59], *etc*. Recently, Transformer-based approaches [8, 50] are proposed for the semantic segmentation task. A basic objective function for FCN-based semantic segmentation models is the pixel-wise cross-entropy loss. There are some other objective functions for semantic segmentation, *e.g.*, DiceLoss [42], IoULoss [2], BCELoss [51] *etc*. All these losses are proposed for a centralized training paradigm, and in this paper, we propose a simple modified cross-entropy loss to tackle the class-heterogeneity under the decentralized training process. Although a similar objective function has been proposed in [5], the perspectives are totally different. [5] focuses on the incremental learning problem and tackles the issue that the background may contain the old classes under the centralized training paradigm. Differently, our FedSeg modifies the cross-entropy loss to correct the local update direction for the decentralized training.

Another relative method explores cross-image pixel contrast for semantic segmentation [47]. The purpose of [47] is to learn a robust pixel embedding space by contrastive learning. Differently, our FedSeg proposes local-to-global contrastive learning, which aims to enforce the local pixel embedding space close to the global embedding space.

3. Method

In this section, we first present the preliminary and problem formulation. Then we introduce a simple but strong baseline, which modifies the cross-entropy loss to correct the local optimization. Finally, we illustrate the details of our local-to-global pixel contrastive learning. The pipeline of FedSeg is shown in Fig. 2.

3.1. Preliminary

Suppose there are N clients and each client has a local semantic segmentation dataset $\mathcal{D}^i$. For the class-heterogeneous situation, $P_{c\sim D_i} \not\sim P_{c\sim D_j} (i \neq j)$ where c denotes class. Our goal is to train a semantic segmentation model w over the dataset $\mathcal{D} := \cup_{i\in[N]}\mathcal{D}^i$ without exchanging the raw data of clients. The objective is as follows:

$$\arg\min_{w} \mathcal{L}(w) = \sum_{i=1}^{N} \frac{|\mathcal{D}^i|}{\mathcal{D}} \mathcal{L}_i(w), \tag{1}$$

where $\mathcal{L}_i(w) = \mathbb{E}_{(\mathbf{x},\mathbf{y})\in\mathcal{D}^i}[l_i(\mathbf{x},\mathbf{y};w)]$ represents the local objective in Client i.

We use FedAvg [31] as the base learning framework. For each training round, all clients optimize their local models on the local datasets. Then the server takes the expectation of the local model parameters to update the global model as follows:

$$w = \sum_{i=1}^{N} \frac{|\mathcal{D}^i|}{\mathcal{D}} w_i, \tag{2}$$

where w_i is the local parameters of Client i. Our FedSeg modifies the local optimization process by redesigning objectives, without extra information exchanging.

3.2. Optimization Correction via Background Classes Aggregation

For the semantic segmentation task under the centralized learning paradigm, a standard objective function is the cross-entropy (CE) loss

$$\mathcal{L}_{ce}(x, y) = -\frac{1}{|\mathcal{P}|} \sum_{j \in \mathcal{P}} \log q_x(j, y_j), \tag{3}$$

where y_j is the ground truth label of pixel j and $\mathcal{P}$ is the set of annotated pixels. $q_x(j, y_j)$ is the predicted probability of pixel j. The background pixels without annotation are ignored. However, for the decentralized FL, the local client with only part of the semantic classes makes the local optimization diverge across clients, resulting in the poor convergence of the global model. For example, the optimization of a local client with only the "cat" annotation is highly different from another client with only the "person" annotation, because the gradient direction of the local model is towards the local optimum to recognize the "cat" while ignoring the direction to other classes. Thus, we propose to modify the CE loss by aggregating the probabilities of other classes (classes not in this client), which corrects the local optimization by providing the gradient direction to other classes.

Formally, the modified cross-entropy loss for the semantic segmentation is

$$\mathcal{L}_{backce}^{i}(x, y) = -\frac{1}{|\mathcal{P}|} \sum_{j \in \mathcal{P}} \log \hat{q}_x(j, y_j), \tag{4}$$

where $\mathcal{P}$ is the set of all pixels, y_j is the ground truth label of pixel j and

$$\hat{q}_x(j, c) = \begin{cases} q_x(j, c) & \text{if } c \in \mathcal{C}_i \\ \sum_{k \in \mathcal{C} \setminus \mathcal{C}_i}^{K} q_x(j, k) & \text{if } c \notin \mathcal{C}_i. \end{cases} \tag{5}$$

The $q_x(j, c)$ denotes the predicted probability of the class c for the pixel j. $\mathcal{C}$ is the set of overall semantic classes and $\mathcal{C}_i$ is the set of annotated classes in Client i. $K := |\mathcal{C}|$ is the total number of classes. Note that there are some other usually used objective functions, *e.g.*, BCELoss, DiceLoss [58] and Lov*á*szLoss [2] for the semantic segmentation. We will compare $\mathcal{L}_{backce}$ with them in Experiments (Sec.4.3).

Discussion. From the optimization perspective, the purpose of the proposed $\mathcal{L}_{backce}$ is correcting the optimization to make it similar to centralized learning. Suppose the predicted logit of class c for pixel j is z_c^j, the gradient of $\mathcal{L}_{ce}$ with respect to z_c^j is

$$\frac{\partial \mathcal{L}_{ce}}{\partial z_c^j} = \begin{cases} p_c^j - 1 < 0 & \text{if } y_j = c \\ p_c^j > 0 & \text{if } y_j \neq c, \end{cases} \tag{6}$$

where $p_c^j = \frac{e^{z_c}}{\sum_{k=1}^{K} e^{z_k}}$ is the predicted probability of class c for pixel j. To simplify the notations of the formula, we use z_c instead of z_c^j. For the centralized semantic segmentation, since all classes are in the dataset, the optimization with respect to z_c contains both the negative and positive direction considering the pixel label is c or not.

However, for the decentralized FL, suppose the annotated data of Client i only contains class l. For class $c \notin \mathcal{C}_i$, the optimization with respect to z_c of standard CE is only the positive direction, *i.e.*, $\frac{\partial \mathcal{L}_{ce}}{\partial z_c} = p_c > 0$, which is different from the centralized learning and away from the global optimum. Thus, we correct the optimization direction by $\mathcal{L}_{backce}$. For the background pixels where $y_j \neq l$, the gradient of $\mathcal{L}_{backce}$ with respect to z_c is

$$\begin{aligned} \frac{\partial \mathcal{L}_{backce}}{\partial z_c} &= -\frac{e^{z_c}}{\sum_{k=1}^{K} e^{z_k}} \cdot \frac{e^{z_l}}{\sum_{k \neq l}^{K} e^{z_k}} \\ &= -p_c \cdot \frac{e^{z_l}}{\sum_{k \neq l}^{K} e^{z_k}} \approx -p_c \cdot p_l, \end{aligned} \tag{7}$$

where p_c and p_l denote the predicted probabilities of class c and l, respectively. More details are shown in Appendix.

Equation 7 shows that for the background pixels where $y_j \neq l$ in local Client i, $\mathcal{L}_{backce}$ provides a negative direction of optimization for the logit z_c of class c, which is related to two terms, p_c and p_l. Since the local model is started from the aggregated global model, which contains information of all classes, the predicted probabilities can provide pseudo-label information. If p_c is larger, *i.e.*, the predicted probability of the pixel label y_j to be c is high, the gradient $\frac{\partial \mathcal{L}_{backce}}{\partial z_c}$ is larger towards class c. If p_l is larger, since the label is not l, the gradient provides a larger number to make the direction towards class c and away from class l.

From the embedding space perspective, the local model cannot learn the relative position of different classes in the pixel embedding space, if it never accessed the data of other classes. Thus the global model cannot distinguish different semantics, as shown in Fig. 1 (b). $\mathcal{L}_{backce}$ learns a relative position between the local classes and other classes. Then the aggregated model obtains a discriminative embedding space by the background alignment, as shown in Fig. 3.

Although [5] proposed a similar objective function, the perspectives are totally different: [5] addresses incremental learning under the centralized learning paradigm, and aims

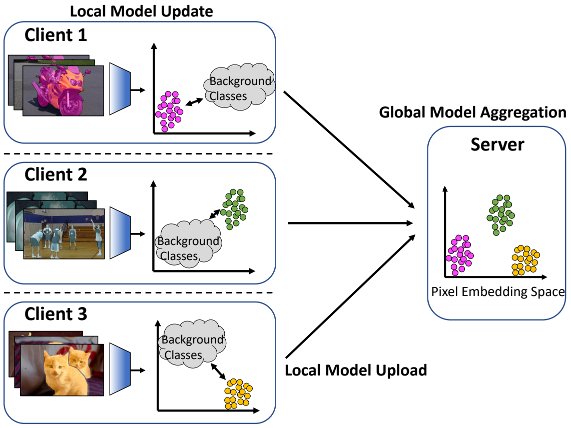

Figure 3. A pixel embedding perspective to understand $\mathcal{L}_{backce}$.

to solve the problem that the background may contain old classes. We proposed $\mathcal{L}_{backce}$ to correct the optimization under the decentralized learning paradigm, which aims to solve the optimization divergence problem across clients. Experiments on PascalVOC [13] (Sec. 4.2) show that although the background does not contain any other classes, our $\mathcal{L}_{backce}$ still improves the segmentation performance significantly under the decentralized FL setting.

3.3. Local-to-Global Pixel Contrastive Learning

Previous FL methods (*e.g.*, FedProx [22], MOON [26]) usually correct the local training by a coarse constraint between the entire local and global model to tackle the non-IID issue. For example, FedProx [22] proposed a proximal term to constrain the local and global weight. MOON [26] constrains the local and global model representations. However, for the dense prediction task of semantic segmentation, a fine-grained method is needed to precisely restrict the local model similar to the global model. Thus, our FedSeg proposes a local-to-global pixel contrastive learning method. Particularly, during the local update process, we extract the pixel representations of the local model and the global model. Then a local pixel representation is pulled close to the global representation of the same semantic class while pushed away from the global representation of different semantic classes, as shown in Fig. 2. FedSeg aims to enforce the learned local pixel embedding space close to the global embedding space in a fine-grained way.

Formally, for a pixel representation $\mathbf{v}_l$ extracted by the local model with its semantic label c, the positive samples are pixel representations $\mathbf{v}_g^+$ extracted by the global model belonging to the same label c, while the negatives are the pixel representations $\mathbf{v}_g^-$ extracted by the global model belonging to the other classes $\mathcal{C} \setminus c$. The local-to-global pixel contrastive loss is as follows:

$$\mathcal{L}_{con}^{j} = \frac{1}{|\mathcal{P}_j|} \sum_{\mathbf{v}_g^+ \in \mathcal{P}_j} -\log \frac{\exp(\mathbf{v}_l \cdot \mathbf{v}_g^+/\tau)}{\exp(\mathbf{v}_l \cdot \mathbf{v}_g^+/\tau) + \sum_{\mathbf{v}_g^- \in \mathcal{N}_j} \exp(\mathbf{v}_l \cdot \mathbf{v}_g^-/\tau)}, \tag{8}$$

where τ is a temperature hyper-parameter. $\mathcal{P}_j$ and $\mathcal{N}_j$ denote the sets of the positive and negative global pixel representations, respectively, for pixel j.

For the class-heterogeneous problem in FL training, a local client contains partial semantic labels and the negative samples are insufficient. The background pixels may contain other semantic classes. Thus, we propose to employ the global model to provide pseudo labels for the background pixels, because the aggregated global model is capable of predicting all semantic classes. We set a threshold θ and the predicted probability of a pixel greater than θ is annotated as the pseudo label of the corresponding class.

Pixel-to-Region Contrast. In practice, the above pixel-to-pixel contrastive learning is computationally inefficient because there are vast numbers of pixel samples in the dense prediction setting. Most of them are redundant especially pixels from the same objects. Therefore, we choose to maintain the representation for each class per image, since pixels belonging to the same class in one image contain similar information. We adopt the averaging operation to aggregate the global pixel representations of the same class to a region representation. Specifically, for a local dataset of Client i with N images, we extract region representations $\mathbf{V}_g^r \in \mathbb{R}^{N \times K \times D}$ by the global model, where K denotes the total semantic classes, and D denotes the dimension of pixel embeddings. The positive and negative samples are provided by $\mathbf{V}_g^r$. The pixel-to-region contrastive learning significantly increases the efficiency.

The overall local objective function is

$$\mathcal{L} = \mathcal{L}_{backce} + \lambda \mathcal{L}_{con} \tag{9}$$

where λ is a hyper-parameter to control the weight of the pixel contrastive loss.

4. Experiments

4.1. Data Settings

We conduct experiments on four semantic segmentation datasets, *i.e.*, Cityscapes [9], CamVID [3], PascalVOC [13] and ADE20k [63]:

Cityscapes [9] and **CamVID** [3] are two semantic segmentation datasets of street view with 19 and 11 semantic classes, respectively. Unlike classification, an image from semantic segmentation datasets contains objects of many classes that are hard to split. To generate the class-heterogeneous data partition among clients, we split Cityscapes and CamVID into K subsets. Each subset maintains one or two semantic classes and sets other classes as background. K is set to 19 and 11 for Cityscapes and CamVID, respectively. In this setting, there exists an inconsistent foreground-background problem for different clients.

Table 1. Main results and comparison of FedSeg. Non-IID$_1$ and non-IID$_2$ denote a subset with one and two classes, respectively.

(a) Results of FedSeg(%) to show the effectiveness of $\mathcal{L}_{backce}$ and $\mathcal{L}_{con}$.

Method	Cityscapes				CamVID				VOC		ADE20k	
	non-IID$_1$		non-IID$_2$		non-IID$_1$		non-IID$_2$					
	mIoU	Acc	mIoU	Acc	mIoU	Acc	mIoU	Acc	mIoU	Acc	mIoU	Acc
FedAvg [31]	10.40	31.90	28.60	73.76	19.06	51.71	32.12	69.55	8.56	34.44	6.91	59.25
FedAvg+$\mathcal{L}_{backce}$	45.08	87.98	47.67	89.48	58.38	88.51	62.13	90.00	32.28	54.83	8.31	61.60
FedAvg+$\mathcal{L}_{backce}$+$\mathcal{L}_{con}$	**50.24**	**90.06**	**52.18**	**91.38**	**63.50**	**90.68**	**64.67**	**91.25**	32.20	54.50	**8.64**	**62.10**

(b) Comparison with other FL methods(%). *All of them use $\mathcal{L}_{backce}$ as baseline.

	mIoU	Acc	mIoU	Acc	mIoU	Acc	mIoU	Acc	mIoU	Acc	mIoU	Acc
FedAvg* [31]	45.08	87.98	47.67	89.48	58.38	88.51	62.13	90.00	32.28	54.83	8.31	61.60
FedProx* [22]	44.85	87.50	47.17	89.81	58.29	87.28	62.04	90.61	32.17	55.19	8.25	61.01
FedDyn* [1]	45.19	88.26	47.69	90.38	59.44	89.32	62.18	90.20	32.20	54.59	-	-
MOON* [26]	45.84	88.58	47.87	89.59	58.90	87.96	62.77	90.98	30.92	53.91	-	-
FedSeg	**50.24**	**90.06**	**52.18**	**91.38**	**63.50**	**90.68**	**64.67**	**91.25**	32.20	54.50	**8.64**	**62.10**

PascalVOC [13] is an image semantic segmentation dataset with $2,913$ images, 20 foreground classes and the background class. Images from PascalVOC are simpler, and most of them contain one or two foreground classes. We split it into 20 subsets corresponding to 20 foreground classes to generate the non-IID data partitions. The background of a subset does not contain other semantic classes. **ADE20k [63]** is a large-scale semantic segmentation dataset with $20,210$ images and 150 semantic classes. The distribution of semantic classes is long-tailed. To generate the non-IID data partitions, we gradually split the tail class into a subset and finally generate 150 subsets. The subset of a more frequent class does not contain the tail classes. The background of a subset does not contain other classes.

We adopt two commonly used metrics for semantic segmentation: **mIoU** indicates the intersection-over-union between the predicted and ground truth pixels, averaged over all the classes. **Pixel Accuracy** indicates the proportion of correctly classified pixels. Please refer to Appendix for the implement details.

4.2. Main Results

Results of FedSeg. Table 1 (a) shows the main results of our FedSeg, *i.e.*, the effectiveness of $\mathcal{L}_{backce}$ and $\mathcal{L}_{con}$. We use FedAvg [31] as the base FL framework. FedAvg in Table 1 (a) means the FedAvg method with the standard cross-entropy loss. FedAvg+$\mathcal{L}_{backce}$ denotes we use our modified CE loss instead of standard CE loss while FedAvg+$\mathcal{L}_{backce}$+$\mathcal{L}_{con}$ means both $\mathcal{L}_{backce}$ and $\mathcal{L}_{con}$ are used. We try three times and report the average number.

Effectiveness of $\mathcal{L}_{backce}$. Table 1 (a) shows that the background aggregation CE loss significantly improves the segmentation performance compared with the standard CE loss, illustrating that the optimization correction by background aggregation is critical to class-heterogeneous FL for semantic segmentation. For Cityscapes, CamVID and PascalVOC, $\mathcal{L}_{backce}$ improves more than +20% for mIoU. For the difficult large-scale ADE20k, $\mathcal{L}_{backce}$ also has +1.4% mIoU improvement.

The comparison between non-IID$_1$ and non-IID$_2$ of Cityscapes and CamVID shows that higher heterogeneous distribution (a client only contains one class) significantly reduces the segmentation performance for the standard CE loss. For instance, the mIoU score of non-IID$_2$ for Cityscapes is +18.2% larger than non-IID$_1$. Thus the improvement of $\mathcal{L}_{backce}$ is larger when heterogeneity is higher.

Note that for PascalVOC, although the background of a client does not contain other classes from other clients, $\mathcal{L}_{backce}$ still improves mIoU by +23%, illustrating the gain of $\mathcal{L}_{backce}$ is from the optimization correction under the decentralized learning instead of only learning the background classes. This is different from a similar method [5].

Effectiveness of $\mathcal{L}_{con}$. Table 1 (a) shows that adding $\mathcal{L}_{con}$ improves mIoU by +2.5% $\sim$ 5.2% on Cityscapes and CamVID, demonstrates the effectiveness of $\mathcal{L}_{con}$. For PascalVOC, the effectiveness of $\mathcal{L}_{con}$ is limited. This is because contrastive loss needs enough negative samples for good performance. However, PascalVOC is a simple dataset that each client contains one or two classes, *i.e.*, negative classes are quite limited and $\mathcal{L}_{con}$ doesn't take effects. For the difficult large-scale ADE20k, the mIoU is slightly improved by +0.3%.

Comparison with FL Methods. We compare FedSeg with other FL methods including FedAvg [31], FedProx [22], FedDyn [1] and MOON [26] in Table 1 (b). All of them use $\mathcal{L}_{backce}$ during the local update for fair comparison since $\mathcal{L}_{backce}$ can be seen as a strong baseline. Table 1 (b) shows that previous FL methods (FedProx [22], FedDyn [1], MOON [26]) for non-IID problem achieve similar or even lower segmentation performance than FedAvg [31]. This indicates that a coarse regularization of entire model weights or representations cannot perform well for the dense prediction task. Our FedSeg employs fine-grained pixel-wise contrastive learning and improves mIoU

Table 2. Results of FedSeg(%) on IID setting.

Method	Cityscapes		CamVID	
	mIoU	Acc	mIoU	Acc
FedAvg [31]	54.12	92.82	64.54	91.35
FedAvg+$\mathcal{L}_{backce}$	53.62	92.60	65.66	91.70
FedAvg+$\mathcal{L}_{backce}$+$\mathcal{L}_{con}$	59.08	93.12	71.05	92.77

Table 3. Comparison with other semantic segmentation loss.

Method	Cityscapes		CamVID		VOC	
	mIoU	Acc	mIoU	Acc	mIoU	Acc
CELoss	10.40	31.90	19.06	51.71	8.56	34.44
BCELoss	32.53	78.81	40.48	73.82	27.19	51.68
Lov*á*szLoss [2]	24.25	64.26	39.29	71.05	25.78	51.27
DiceLoss [58]	29.34	69.44	42.06	74.9	32.20	54.80
BackCELoss	45.08	87.98	58.38	88.51	32.28	54.83

score effectively.

Evaluation on IID Setting. We report the performances of FedSeg on the IID distribution setting, which randomly splits images into different clients in Table 2. Results show that using $\mathcal{L}_{backce}$ achieves similar performance to the standard CE loss. This is because $\mathcal{L}_{backce}$ is designed to correct the optimization and tackles non-IID problem. Since every client contains all classes for IID, the gradient updating is similar between $\mathcal{L}_{backce}$ and $\mathcal{L}_{ce}$. Adding $\mathcal{L}_{con}$ achieves better performance because it learns a better pixel embedding space.

4.3. Empirical Analysis

Comparison with other Semantic Segmentation Loss. We further compare $\mathcal{L}_{backce}$ with commonly used semantic segmentation losses, including BCELoss, DiceLoss [58] and Lov*á*szLoss [2]. These semantic segmentation losses conduct gradients only on the logits of foreground classes, which alleviates the optimization divergence problem for the FL segmentation. As shown in Table 3, the segmentation performances of these structure-aware optimization criteria surpass the standard CE Loss. Our proposed $\mathcal{L}_{backce}$ further improves the segmentation performance compared with these losses, illustrating that $\mathcal{L}_{backce}$ corrects the optimization direction more similar to the centralized training. Specifically, $\mathcal{L}_{backce}$ surpasses the best competition DiceLoss [58] by +15.7% and +16.3% on Cityscapes (non-IID$_1$) and CamVID (non-IID$_1$), respectively. On PascalVOC, the improvement of $\mathcal{L}_{backce}$ is not as significant as other datasets. This is because $\mathcal{L}_{backce}$ addresses background inconsistency problem and is more suitable to tackle the scenario that a client's background contains other clients' classes. However, for PascalVOC, each client doesn't contain others' classes.

Comparison between Local-to-Global and Local-to-Local Contrastive Loss. We propose the local-to-global pixel contrastive loss $\mathcal{L}_{con}$ to enforce the pixel embedding space of the local model close to the global model. A recent work [47] shows that pixel contrastive learning itself can improve the segmentation performance by learning a better embedding space. Thus, we compare our proposed local-to-global pixel contrastive loss with the pixel contrastive loss only on the local model. Table 4 illustrates that the local-to-global pixel contrastive loss outperforms the local-to-local contrastive loss, indicating that the reason of performance gain is correcting the optimization for decentralized learning instead of only the contrastive learning itself.

Table 4. Comparison between local-to-global and local-to-local contrastive loss on mIoU score.

Method	Cityscapes		CamVID	
	non-IID$_1$	non-IID$_2$	non-IID$_1$	non-IID$_2$
Local2Local [47]	47.91%	50.10%	62.00%	63.05%
Local2Global	50.24%	52.18%	63.50%	64.67%

Table 5. Comparison between pixel-to-pixel and pixel-to-region contrastive loss on mIoU score.

Method	Cityscapes		CamVID	
	non-IID$_1$	non-IID$_2$	non-IID$_1$	non-IID$_2$
Pixel2Pixel	49.95%	51.79%	63.30%	63.95%
Pixel2Region	50.24%	52.18%	63.50%	64.67%

Communication Efficiency. Fig. 4 shows the mIoU score in each round during training. FedAvg* [31], FedProx* [22], FedDyn* [1] and MOON* [26] use $\mathcal{L}_{backce}$ as the objective function of local update for a fair comparison. The speed of mIoU improvement of these comparable FL methods is similar. Our FedSeg adds the local-to-global contrastive loss for fine-grained local updates correction, which consistently increases the segmentation performance and communication efficiency. As shown in Fig. 4, the speed of mIoU improvement for FedSeg is significantly faster than FedAvg* [31] at the beginning on both Cityscapes (non-IID$_1$ and non-IID$_2$) and CamVID (non-IID$_1$ and non-IID$_2$) datasets.

Visualization. We use t-SNE [45] to visualize the pixel embeddings of semantic classes from the Cityscapes validation dataset, as shown in Fig. 5. We compare FedAvg [31] (with standard CE loss), FedAvg+$\mathcal{L}_{backce}$ and FedAvg+$\mathcal{L}_{backce}$+$\mathcal{L}_{con}$. Fig. 5 shows that the model trained with the standard CE loss learns poor embeddings and pixel embeddings of different semantic classes are even mixed. Using the background aggregation CE loss learns better pixel embeddings. Adding the pixel contrastive learning further improves the divergence of different classes in the embedding space.

Comparison between Pixel-to-Pixel (P2P) and Pixel-to-Region (P2R) Contrastive Loss. We compare the segmentation performance of $\mathcal{L}_{con}$ between different sampling

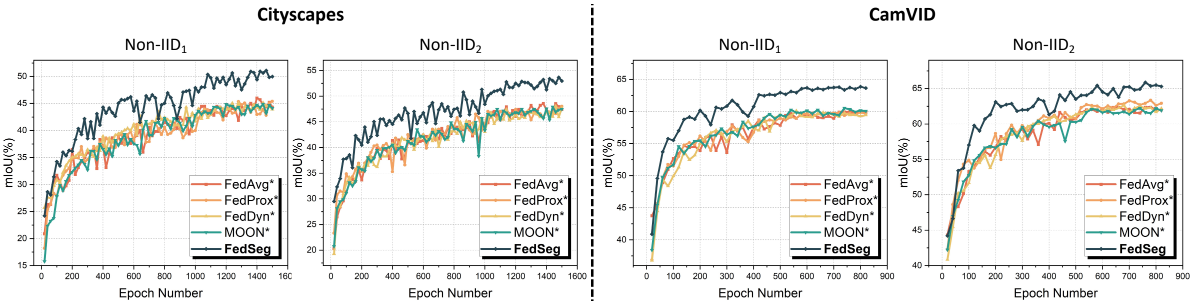

Figure 4. Comparison of communication efficiency on Cityscapes and CamVID.

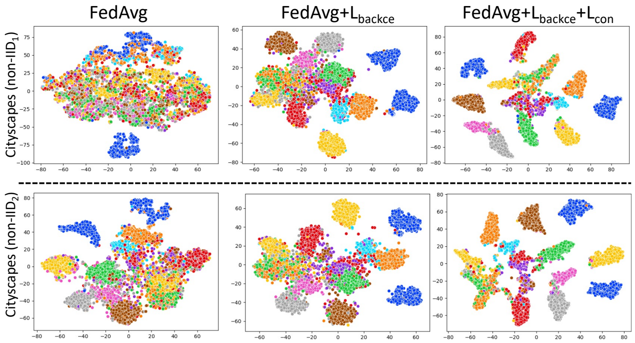

Figure 5. Visualization of the pixel embeddings for differenct semantic classes.

strategies in Table 5. P2P and P2R mean sampling the positive and negative embeddings at pixel level and region level, respectively. Results show that the two sampling strategies achieve similar mIoU score. Then we compare the training efficiency between the two strategies using Cityscapes on the same device (V100 GPU), and the training speeds are 36 v.s. 27 steps (P2R *vs*. P2P). Formally, give the image feature size ($H \times W$), the class number (C), the complexity is $\mathcal{O}(HWC)$ v.s. $\mathcal{O}((HW)^2)$ (P2R v.s. P2P). Since $C \ll HW$, P2R is faster than P2P. Thus, we choose pixel-to-region for $\mathcal{L}_{con}$.

Number of Participant Clients in Each Round. We evaluate our FedSeg on different numbers of participant clients. Specifically, we randomly select $5, 10, 15, 20$ clients to participate in federated learning in each round. As shown in Fig. 6, with the number of clients growing, the mIoU performance of FedAvg [31] with the standard CE loss improves significantly since more participant clients per round provide more data to alleviate the client shift problem. The performance of FedSeg slightly improves with the number of clients growing in Fig. 6.

Number of Local Epochs. We study the effect of the number of local epochs on non-IID$_1$ setting of Cityscapes and CamVID. The total epoch number $= G \times L$ where G and L denotes the number of global and local epochs, respectively. By keeping the total epoch number unchanged, the segmentation performances are shown in Fig. 7 with the number of local epochs growing. We found that a larger number of local epochs makes the segmentation performance degrade, since more local epochs tend to the local optimum under the non-IID setting. We choose the best number of local epochs in the paper.

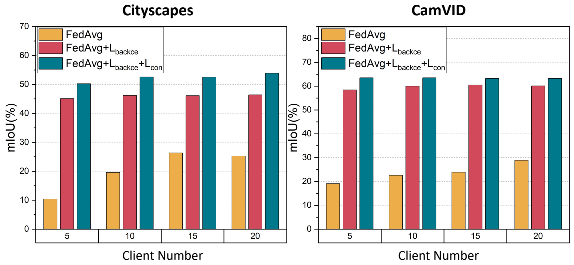

Figure 6. Effect of the participant client number per round. Non-IID$_1$ setting is used.

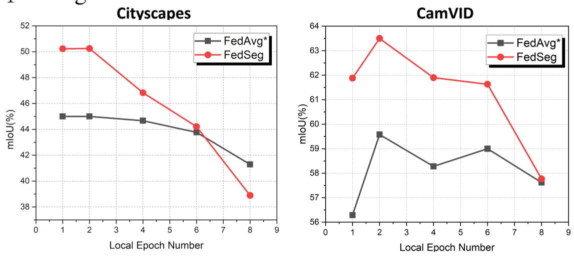

Figure 7. Effect of the local epoch number per round. Non-IID$_1$ setting is used.

5. Conclusion

We investigate class-heterogeneous federated learning for semantic segmentation. To address the foreground-background inconsistency and the client drifts during local updates, we propose a baseline method FedSeg with a modified CE loss and a local-to-global pixel contrastive loss. Extensive experiments are conducted on four semantic segmentation datasets to show the effectiveness of FedSeg. We hope the baseline and benchmarks can help class-heterogeneous FL for semantic segmentation be extensively studied in the future.

Acknowledgments. This work is supported by National Key R&D Program of China under Grant 2021ZD0112801 and the Fundamental Research Funds for the Central Universities (No. 226-2022-00051).

References

[1] Durmus Alp Emre Acar, Yue Zhao, Ramon Matas Navarro, Matthew Mattina, Paul N Whatmough, and Venkatesh Saligrama. Federated learning based on dynamic regularization. *arXiv preprint arXiv:2111.04263*, 2021. 1, 2, 6, 7

[2] Maxim Berman, Amal Rannen Triki, and Matthew B Blaschko. The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4413–4421, 2018. 3, 4, 7

[3] Gabriel J Brostow, Jamie Shotton, Julien Fauqueur, and Roberto Cipolla. Segmentation and recognition using structure from motion point clouds. In *European conference on computer vision*, pages 44–57. Springer, 2008. 2, 5

[4] Debora Caldarola, Barbara Caputo, and Marco Ciccone. Improving generalization in federated learning by seeking flat minima. In *European Conference on Computer Vision*, pages 654–672. Springer, 2022. 3

[5] Fabio Cermelli, Massimiliano Mancini, Samuel Rota Bulo, Elisa Ricci, and Barbara Caputo. Modeling the background for incremental learning in semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9233–9242, 2020. 3, 4, 6

[6] Qi Chang, Hui Qu, Yikai Zhang, Mert Sabuncu, Chao Chen, Tong Zhang, and Dimitris N Metaxas. Synthetic learning: Learn from distributed asynchronized discriminator gan without sharing medical image data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13856–13866, 2020. 2

[7] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. 3

[8] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Advances in Neural Information Processing Systems*, 34:17864–17875, 2021. 3

[9] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 2, 3, 5

[10] Ittai Dayan, Holger R Roth, Aoxiao Zhong, Ahmed Harouni, Amilcare Gentili, Anas Z Abidin, Andrew Liu, Anthony Beardsworth Costa, Bradford J Wood, Chien-Sung Tsai, et al. Federated learning for predicting clinical outcomes in patients with covid-19. *Nature medicine*, 27(10):1735–1743, 2021. 2

[11] Jiahua Dong, Lixu Wang, Zhen Fang, Gan Sun, Shichao Xu, Xiao Wang, and Qi Zhu. Federated class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10164–10173, 2022. 2

[12] Nanqing Dong and Irina Voiculescu. Federated contrastive learning for decentralized unlabeled medical images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 378–387. Springer, 2021. 2

[13] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. 2, 3, 5, 6

[14] Lidia Fantauzzo, Eros Fani, Debora Caldarola, Antonio Tavera, Fabio Cermelli, Marco Ciccone, and Barbara Caputo. Feddrive: Generalizing federated learning to semantic segmentation in autonomous driving. *arXiv preprint arXiv:2202.13670*, 2022. 2, 3

[15] Liang Gao, Huazhu Fu, Li Li, Yingwen Chen, Ming Xu, and Cheng-Zhong Xu. Feddc: Federated learning with non-iid data via local drift decoupling and correction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10112–10121, 2022. 2

[16] Xuan Gong, Abhishek Sharma, Srikrishna Karanam, Ziyan Wu, Terrence Chen, David Doermann, and Arun Innanje. Ensemble attention distillation for privacy-preserving federated learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15076–15086, 2021. 3

[17] Pengfei Guo, Dong Yang, Ali Hatamizadeh, An Xu, Ziyue Xu, Wenqi Li, Can Zhao, Daguang Xu, Stephanie Harmon, Evrim Turkbey, et al. Auto-fedrl: Federated hyperparameter optimization for multi-institutional medical image segmentation. In *ECCV*, 2022. 2

[18] Sungwon Han, Sungwon Park, Fangzhao Wu, Sundong Kim, Chuhan Wu, Xing Xie, and Meeyoung Cha. Fedx: Unsupervised federated learning with cross knowledge distillation. *arXiv preprint arXiv:2207.09158*, 2022. 2

[19] Wenke Huang, Mang Ye, and Bo Du. Learn from others and be yourself in heterogeneous federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10143–10153, 2022. 2

[20] Georgios A Kaissis, Marcus R Makowski, Daniel Rückert, and Rickmer F Braren. Secure, privacy-preserving and federated machine learning in medical imaging. *Nature Machine Intelligence*, 2(6):305–311, 2020. 2

[21] Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning*, pages 5132–5143. PMLR, 2020. 1, 2

[22] Daiqing Li, Amlan Kar, Nishant Ravikumar, Alejandro F Frangi, and Sanja Fidler. Federated simulation for medical imaging. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 159–168. Springer, 2020. 1, 2, 5, 6, 7

[23] Liulei Li, Tianfei Zhou, Wenguan Wang, Jianwu Li, and Yi Yang. Deep hierarchical semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1246–1257, 2022. 3

[24] Mengze Li, Tianbao Wang, Haoyu Zhang, Shengyu Zhang, Zhou Zhao, Jiaxu Miao, Wenqiao Zhang, Wenming Tan, Jin

Wang, Peng Wang, et al. End-to-end modeling via information tree for one-shot natural language spatial video grounding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8707–8717, 2022. 3

[25] Mengze Li, Tianbao Wang, Haoyu Zhang, Shengyu Zhang, Zhou Zhao, Wenqiao Zhang, Jiaxu Miao, Shiliang Pu, and Fei Wu. Hero: Hierarchical spatio-temporal reasoning with contrastive action correspondence for end-to-end video object grounding. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3801–3810, 2022. 3

[26] Qinbin Li, Bingsheng He, and Dawn Song. Model-contrastive federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10713–10722, 2021. 1, 2, 5, 6, 7

[27] Wenqi Li, Fausto Milletarì, Daguang Xu, Nicola Rieke, Jonny Hancox, Wentao Zhu, Maximilian Baust, Yan Cheng, Sébastien Ourselin, M Jorge Cardoso, et al. Privacy-preserving federated brain tumour segmentation. In *International workshop on machine learning in medical imaging*, pages 133–141. Springer, 2019. 2

[28] Chen Liang, Wenguan Wang, Jiaxu Miao, and Yi Yang. Gmmseg: Gaussian mixture based generative semantic segmentation models. In *Advances in Neural Information Processing Systems*, 2022. 3

[29] Quande Liu, Cheng Chen, Jing Qin, Qi Dou, and Pheng-Ann Heng. Feddg: Federated domain generalization on medical image segmentation via episodic learning in continuous frequency space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1013–1023, 2021. 2

[30] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 1, 3

[31] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017. 1, 2, 3, 6, 7, 8

[32] Matias Mendieta, Taojiannan Yang, Pu Wang, Minwoo Lee, Zhengming Ding, and Chen Chen. Local learning matters: Rethinking data heterogeneity in federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8397–8406, 2022. 1, 2

[33] Jiaxu Miao, Xiaohan Wang, Yu Wu, Wei Li, Xu Zhang, Yunchao Wei, and Yi Yang. Large-scale video panoptic segmentation in the wild: A benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21033–21043, 2022. 3

[34] Jiaxu Miao, Yunchao Wei, Yu Wu, Chen Liang, Guangrui Li, and Yi Yang. Vspw: A large-scale dataset for video scene parsing in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4133–4143, 2021. 3

[35] Jiaxu Miao, Yunchao Wei, and Yi Yang. Memory aggregation networks for efficient interactive video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10366–10375, 2020. 3

[36] Umberto Michieli and Mete Ozay. Prototype guided federated learning of visual feature representations. *arXiv preprint arXiv:2105.08982*, 2021. 2

[37] Ahmad Naeem, Tayyaba Anees, Rizwan Ali Naqvi, and Woong-Kee Loh. A comprehensive analysis of recent deep and federated-learning-based methodologies for brain tumor diagnosis. *Journal of Personalized Medicine*, 12(2):275, 2022. 2

[38] Sashank Reddi, Zachary Charles, Manzil Zaheer, Zachary Garrett, Keith Rush, Jakub Konečnỳ, Sanjiv Kumar, and H Brendan McMahan. Adaptive federated optimization. *arXiv preprint arXiv:2003.00295*, 2020. 2

[39] Micah J Sheller, G Anthony Reina, Brandon Edwards, Jason Martin, and Spyridon Bakas. Multi-institutional deep learning modeling without sharing patient data: A feasibility study on brain tumor segmentation. In *International MICCAI Brainlesion Workshop*, pages 92–104. Springer, 2018. 2

[40] Donald Shenaj, Eros Fanì, Marco Toldo, Debora Caldarola, Antonio Tavera, Umberto Michieli, Marco Ciccone, Pietro Zanuttigh, and Barbara Caputo. ”learning across domains and devices: Style-driven source-free domain adaptation in clustered federated learning. In *Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2023. 3

[41] Santiago Silva, Boris A Gutman, Eduardo Romero, Paul M Thompson, Andre Altmann, and Marco Lorenzi. Federated learning in distributed medical databases: Meta-analysis of large-scale subcortical brain data. In *2019 IEEE 16th international symposium on biomedical imaging (ISBI 2019)*, pages 270–274. IEEE, 2019. 2

[42] Carole H Sudre, Wenqi Li, Tom Vercauteren, Sebastien Ourselin, and M Jorge Cardoso. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *Deep learning in medical image analysis and multimodal learning for clinical decision support*, pages 240–248. Springer, 2017. 3

[43] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5693–5703, 2019. 3

[44] Yue Tan, Guodong Long, Lu Liu, Tianyi Zhou, Qinghua Lu, Jing Jiang, and Chengqi Zhang. Fedproto: Federated prototype learning across heterogeneous clients. In *AAAI Conference on Artificial Intelligence*, volume 1, page 3, 2022. 2

[45] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 7

[46] Jiacheng Wang, Yueming Jin, and Liansheng Wang. Personalizing federated medical image segmentation via local calibration. In *ECCV*, 2022. 2

[47] Wenguan Wang, Tianfei Zhou, Fisher Yu, Jifeng Dai, Ender Konukoglu, and Luc Van Gool. Exploring cross-image pixel contrast for semantic segmentation. In *Proceedings*

of the IEEE/CVF International Conference on Computer Vision, pages 7303–7313, 2021. 3, 7

[48] Jeffry Wicaksana, Zengqiang Yan, Dong Zhang, Xijie Huang, Huimin Wu, Xin Yang, and Kwang-Ting Cheng. Fedmix: Mixed supervised federated learning for medical image segmentation. *arXiv preprint arXiv:2205.01840*, 2022. 2

[49] Yawen Wu, Dewen Zeng, Zhepeng Wang, Yiyu Shi, and Jingtong Hu. Federated contrastive learning for volumetric medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 367–377. Springer, 2021. 2

[50] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021. 3

[51] Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In *Proceedings of the IEEE international conference on computer vision*, pages 1395–1403, 2015. 3

[52] An Xu, Wenqi Li, Pengfei Guo, Dong Yang, Holger R Roth, Ali Hatamizadeh, Can Zhao, Daguang Xu, Heng Huang, and Ziyue Xu. Closing the generalization gap of cross-silo federated medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20866–20875, 2022. 2, 3

[53] Yi Yang, Yueting Zhuang, and Yunhe Pan. Multiple knowledge representation for big data artificial intelligence: framework, applications, and case studies. *Frontiers of Information Technology & Electronic Engineering*, 22(12):1551–1558, 2021. 3

[54] Zongxin Yang, Yunchao Wei, and Yi Yang. Associating objects with transformers for video object segmentation. In *Advances in Neural Information Processing Systems*, volume 34, pages 2491–2502, 2021. 3

[55] Zongxin Yang, Yunchao Wei, and Yi Yang. Collaborative video object segmentation by multi-scale foreground-background integration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):4701–4712, 2021. 3

[56] Chun-Han Yao, Boqing Gong, Hang Qi, Yin Cui, Yukun Zhu, and Ming-Hsuan Yang. Federated multi-target domain adaptation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1424–1433, 2022. 3

[57] Liping Yi, Jinsong Zhang, Rui Zhang, Jiaqi Shi, Gang Wang, and Xiaoguang Liu. Su-net: an efficient encoder-decoder model of federated learning for brain tumor segmentation. In *International Conference on Artificial Neural Networks*, pages 761–773. Springer, 2020. 2

[58] Jiahui Yu, Yuning Jiang, Zhangyang Wang, Zhimin Cao, and Thomas Huang. Unitbox: An advanced object detection network. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 516–520, 2016. 4, 7

[59] Yuhui Yuan, Xilin Chen, and Jingdong Wang. Object-contextual representations for semantic segmentation. In *European conference on computer vision*, pages 173–190. Springer, 2020. 3

[60] Lin Zhang, Li Shen, Liang Ding, Dacheng Tao, and Ling-Yu Duan. Fine-tuning global model via data-free knowledge distillation for non-iid federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10174–10183, 2022. 2

[61] Michael Zhang, Karan Sapra, Sanja Fidler, Serena Yeung, and Jose M Alvarez. Personalized federated learning with first order model optimization. *arXiv preprint arXiv:2012.08565*, 2020. 2

[62] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017. 3

[63] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017. 2, 5, 6

[64] Tianfei Zhou, Wenguan Wang, Ender Konukoglu, and Luc Van Gool. Rethinking semantic segmentation: A prototype view. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2582–2593, 2022. 3

[65] Alexander Ziller, Dmitrii Usynin, Nicolas Remerscheid, Moritz Knolle, Marcus Makowski, Rickmer Braren, Daniel Rueckert, and Georgios Kaissis. Differentially private federated deep learning for multi-site medical image segmentation. *arXiv preprint arXiv:2107.02586*, 2021. 2