# Supplement to FedSeg: Class-Heterogeneous Federated Learning for Semantic Segmentation

## A. Implementation Details

We use BiSeNetv2 [7] as the segmentation model and a 2-layer MLP as the projection head to extract pixel embeddings. The SGD optimizer with an initial learning rate of $0.05$ was used. SGD weight decay was set to 5e-4, and batch size was set to 8. The training images are augmented by random scaling, random flipping and random cropping. The random scaling factor is [0.5, 1.5] and the cropping size is $1024 \times 512$, $512 \times 512$, $480 \times 480$ for Cityscapes, CamVID, PascalVOC/ADE20k, respectively. The temperature $\tau$ of the contrastive loss is $0.07$. $\lambda$ in Equation 9 is set to $1$ for Cityscapes, CamVID and ADE20k, while $0.1$ for PascalVOC. For each subset of the four datasets, we further split it into several clients. The total number of clients for Cityscapes, CamVID, PascalVOC and ADE20k is $152$, $22$, $60$ and $450$, respectively. In each communication round $5$ clients are randomly selected. The model is trained for 1,500, 1,200, 800 communication rounds for Cityscapes, PascalVOC and CamVID/ADE20k, respectively, with 2 local epochs in each round. All the comparable methods (FedAvg [6], FedProx [4], FedDyn [1] and MOON [5]) use the same training protocals for fairness.

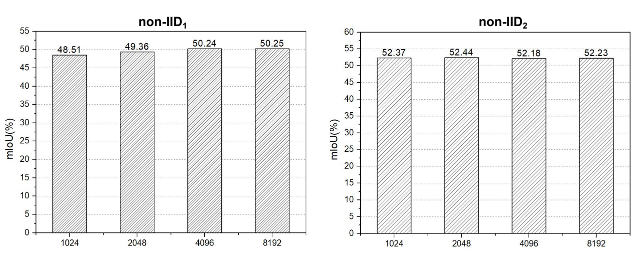


Figure 1. Effect of the number of negative pairs on Cityscapes.

Table 1. Effect of the projection head on mIoU score.

| Method | Cityscapes | | CamVID | |
|---|---|---|---|---|
| | non-IID$_1$ | non-IID$_2$ | non-IID$_1$ | non-IID$_2$ |
| w/ proj head | 45.74% | 46.12% | 59.31% | 60.24% |
| w/o proj head | 50.24% | 52.18% | 63.50% | 64.67% |

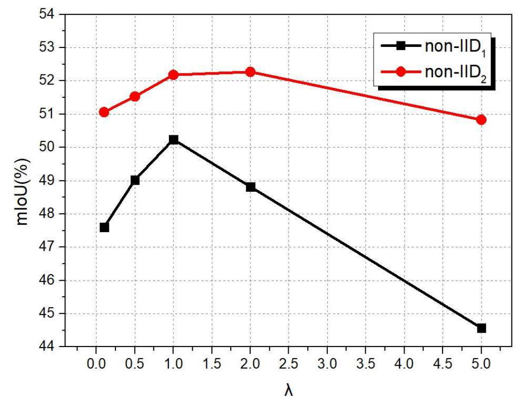


Figure 2. Effect of $\lambda$.

## B. Effect of the Number of Negative Pairs

For pixel-level contrastive learning, the negative samples are pixel embeddings of other classes. Since the pixel embeddings of the same class in one image contain similar information, we randomly sample $N$ pixel embeddings for the local-to-global pixel contrastive learning. Fig. 1 shows the mIoU performance on Cityscapes [3] non-IID$_1$ and non-IID$_2$. $N$ is set to 1,024, 2,048, 4,096 and 8,192. For higher heterogeneous data (non-IID$_1$), more sampled pixels as negative pairs achieve better performance. On Cityscapes [3] (non-IID$_2$), the number of negative pairs is not critical to affect the mIoU performance.

## C. Effect of the Projection Head

We use a 2-layer projection head to map the pixel representations. Here we compare our model with and without the extra project head to show the effect of the projection head. We conduct experiments on Cityscapes [3] (non-IID$_1$ and non-IID$_2$) and CamVID [2] (non-IID$_1$ and non-IID$_2$). Table. 1 shows that adding the projection head improves the segmentation performance by about +4%.

## D. Effect of $\lambda$

We evaluate FedSeg with different $\lambda$ and the mIoU scores are shown in Fig. 2 on Cityscapes [3] (non-IID$_1$

Table 2. The effectiveness of our method on different semantic segmentation models.

| | Dataset | FedAvg | $+L_b$ | $+L_b+L_c$ | | FedAvg | $+L_b$ | $+L_b+L_c$ |
|---|---|---|---|---|---|---|---|---|
| PSPNet [8] | Cityscapes$_1$ | 32.3 | 57.4 | **60.2** | Cityscapes$_2$ | 49.5 | 59.5 | **61.0** |
| | CamVid$_1$ | 41.0 | 61.4 | **63.5** | CamVid$_2$ | 54.6 | 65.0 | **66.7** |
| BiseNetv2 [7] | Cityscapes$_1$ | 10.4 | 45.0 | **50.2** | Cityscapes$_2$ | 28.6 | 47.6 | **52.1** |
| | CamVid$_1$ | 19.0 | 58.3 | **63.5** | CamVid$_2$ | 32.1 | 62.1 | **64.6** |

Table 3. Effectiveness on different federated learning methods.

| | FedProx | FedProx+Ours | FedDyn | FedDyn+Ours | MOON | MOON+Ours |
|---|---|---|---|---|---|---|
| Cityscapes$_1$ | 44.85 | **50.18** | 45.19 | **49.98** | 45.84 | **49.55** |
| CamVid$_1$ | 58.29 | **62.56** | 59.44 | **61.22** | 58.90 | **61.38** |

and non-IID$_2$). $\lambda$ is set to 0.1, 0.5, 1, 2, 5. When $\lambda$ is small ($\lambda = 0.1$) the performance of FedSeg is similar to FedAvg [6] since the impact of the pixel contrastive learning is small. Too large $\lambda$ also drops the segmentation performance. $\lambda = 1$ is a reasonable choice, where FedSeg achieves at least 2.5% higher mIoU than FedAvg.

## E. Effectiveness on Different Semantic Segmentation Models

To show the generalization of our method, we applied our proposed losses on another semantic segmentation model, PSPNet [8]. Results in Table. 2 show that using different segmentation models, BiseNet [7] and PSPNet [8], adding our losses ($\mathcal{L}_{backce}$ , $\mathcal{L}_{con}$) consistently improves mIoU performance, illustrating the generalization of our method. $L_b$ and $L_c$ in Table. 2 indicate $\mathcal{L}_{backce}$ and $\mathcal{L}_{con}$, respectively.

## F. Effectiveness on Different Federated Learning Methods

We added more experiments to apply our proposed losses to FedProx [4], FedDyn [1] and MOON [5], as shown in Table. 3. Results show that adding our FedSeg to these federated learning methods consistently improves the performance, illustrating the generalization of our method.

## G. Details of the Gradient Analysis for $\mathcal{L}_{backce}$

The purpose of $\mathcal{L}_{backce}$ is correcting the gradients for decentralized non-IID FL to make it similar to the centralized training. For centralized training the gradient directions of the logit $z_c$ for class $c$ contain positives and negatives corresponding to the label $y_j$ is $c$ or not. For 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.*, $\frac{\partial \mathcal{L}_{ce}}{\partial z_c} = p_c > 0$. Thus we correct the optimization direction by $\mathcal{L}_{backce}$. For the background pixels where $y_j \neq l$,

$$\mathcal{L}^j_{backce} = -\log \frac{\sum_{k\neq l}^K e^{z_k}}{\sum_{k=1}^K e^{z_k}}, \tag{1}$$

and the gradient of $\mathcal{L}_{backce}$ with respect to $z_c$ is

$$\begin{aligned}
\frac{\partial \mathcal{L}_{backce}}{\partial z_c} &= -\frac{\sum_{k=1}^K e^{z_k}}{\sum_{k\neq l}^K e^{z_k}} \cdot \frac{e^{z_c} \cdot \sum_{k=1}^K e^{z_k} - e^{z_c} \cdot \sum_{k\neq l}^K e^{z_k}}{\left(\sum_{k=1}^K e^{z_k}\right)^2} \\
&= -\frac{e^{z_c} \cdot \left(\sum_{k=1}^K e^{z_k} - \sum_{k\neq l}^K e^{z_k}\right)}{\sum_{k=1}^K e^{z_k} \cdot \sum_{k\neq l}^K e^{z_k}} \\
&= -\frac{e^{z_c}}{\sum_{k=1}^K e^{z_k}} \cdot \frac{e^{z_l}}{\sum_{k\neq l}^K e^{z_k}} \\
&= -p_c \cdot \frac{e^{z_l}}{\sum_{k\neq l}^K e^{z_k}} \approx -p_c \cdot p_l,
\end{aligned} \tag{2}$$

where $p_c$ and $p_l$ denote the predicted probabilities of class $c$ and $l$, respectively. $\mathcal{L}_{backce}$ provides a negative gradient direction with respect to $z_c$ where $y_j \neq l$. The gradient is larger if $p_c$ is larger, which means if the predicted probability of class $c$ is large, the gradient tends to make the model recognize the pixel as class $c$. Since the local model is started from the global model which can predict all classes, $p_c$ can be seen as a pseudo label.

## 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

[2] 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. 1

[3] 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. 1

[4] 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] 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

[6] 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

[7] Changqian Yu, Changxin Gao, Jingbo Wang, Gang Yu, Chunhua Shen, and Nong Sang. Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation. *International Journal of Computer Vision*, 129(11):3051–3068, 2021. 1, 2

[8] 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. 2