A New Perspective to Boost Performance Fairness For Medical Federated Learning

Yunlu Yan[1], Lei Zhu[1,6] (✉), Yuexiang Li[2], Xinxing Xu[3], Rick Siow Mong Goh[3], Yong Liu[3], Salman Khan[4,5], Chun-Mei Feng[3]

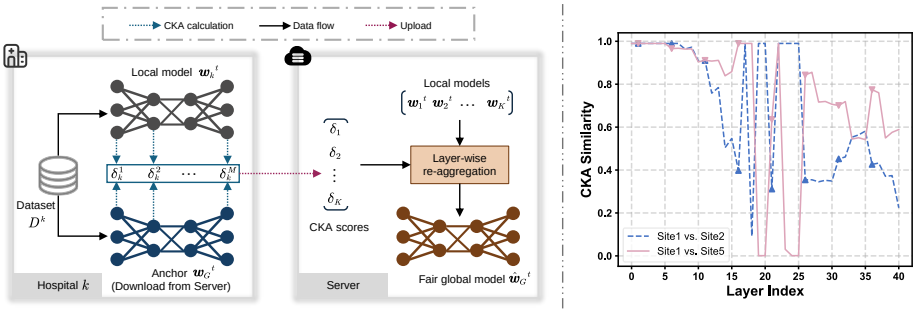
[1] The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China
[2] Guangxi Medical University, Nanning, China
[3] Institute of High Performance Computing (IHPC), Agency for Science, Technology and Research (A*STAR), Singapore
[4] Mohamed bin Zayed University of Artificial Intelligence (MBZUAI), UAE
[5] Australian National University, Canberra ACT, Australia
[6] The Hong Kong University of Science and Technology, Hongkong, China
leizhu@ust.hk
https://github.com/IAMJackYan/Fed-LWR

Abstract. Improving the fairness of federated learning (FL) benefits healthy and sustainable collaboration, especially for medical applications. However, existing fair FL methods ignore the specific characteristics of medical FL applications, *i.e.*, domain shift among the datasets from different hospitals. In this work, we propose `Fed-LWR` to improve performance fairness from the perspective of feature shift, a key issue influencing the performance of medical FL systems caused by domain shift. Specifically, we dynamically perceive the bias of the global model across all hospitals by estimating the layer-wise difference in feature representations between local and global models. To minimize global divergence, we assign higher weights to hospitals with larger differences. The estimated client weights help us to re-aggregate the local models per layer to obtain a fairer global model. We evaluate our method on two widely used federated medical image segmentation benchmarks. The results demonstrate that our method achieves better and fairer performance compared with several state-of-the-art fair FL methods.

Keywords: Federated Learning · Fairness · Medical Image Analysis.

1 Introduction

Federated learning [15,21] (FL) has emerged as a hot research topic in healthcare [5], offering a framework for effectively leveraging diverse datasets to learn a better global model without compromising privacy, significantly facilitating communication and collaboration among medical institutions. However, most of the existing works [9,3,10,27,19] primarily focus on improving the global performance (*e.g.*, average accuracy), incurring disproportionately advantages or



(a) (b)

Fig. 1: (a) **Overview of** `Fed-LWR`. During the parameter aggregation stage of t-th round, `Fed-LWR` calculate layer-wise CKA similarity $\delta_k = \{\delta_k^1, \delta_k^2 \ldots, \delta_k^M\}$ between the local model $\boldsymbol{w}_k^t$ and anchor $\boldsymbol{w}_G^t$ averaged by server on hospital k. The CKA scores will be used to re-aggregate the local models to obtain fair global model $\hat{\boldsymbol{w}}_G^t$. (b) Variations in CKA similarity versus the layers of local models from two pairs of clients, which are randomly selected. This reveals that the differences between models vary with different layers.

disadvantages on some institutions. Fairness has always been an important evaluation criterion for machine learning models, especially for FL applications. An unfair FL system can hurt the incentives of users to participate.

Existing fairness-related FL studies primarily focus on two aspects: *collaboration fairness* [20,31] and *performance fairness* [16,14]. The former pays more attention to resource allocation during the learning process, while the latter concerns the balance of model outcomes. Considering that most real-world users are consequentialists, they are more concerned about the intuitive performance of the learned model on their private dataset, rather than the intricate details of the learning process. Therefore, the main goal of this work is to improve the performance fairness of medical FL system. Performance fairness is often an overlooked concern when striving for higher global performance. For example, although the average accuracy may be high, the accuracy at some institutions could be significantly lower. This skew in performance outcomes necessitates a more nuanced approach to ensure equitable improvements across all participating entities in FL applications. However, overemphasizing fairness in performance is also nonsensical, as the goal of fair FL is not to achieve identical accuracy for every user. The key challenge lies in striking a trade-off between 'high' and 'fair'.

To address this, some studies [22,6,2] introduced additional objectives to constrain the optimization process. Similarly, Ditto [14] not only introduced additional objectives but also personalization in the loss function for different clients to improve fairness. In contrast, another promising solution is to re-weight clients by various metrics, such as empirical loss [16,13] or validation result [31], yielding a flexible fairness/accuracy trade-off. In the medical area, FedCE [8] also proposed a novel re-weighting strategy by estimating client contribution based

on gradients and validation loss. Re-weighting is based on a straightforward idea: assigning higher weights to hospitals with poorer performance, thereby ensuring a more uniform accuracy of the global model [16]. Thus, the key is how to dynamically estimate the differences between hospitals. Different from previous re-weighting methods, we try to estimate client weights from a novel perspective based on the characteristics of medical FL. ***First***, the private data from different medical institutions is typically collected from various devices, leading to domain shifts across different datasets, resulting in variations in the feature representations of different local models, namely ***feature shift*** [17,32,7]. Such variations are the primary factor of the disparate performance of the global model across different datasets. ***Second***, previous methods estimated an overall weight for each client, which is a coarse way. As shown in Fig. 1 (b), the degree of differences across local models vary with different layers. Therefore, directly assigning a single weight would ignore these differences in layers and lead to suboptimal aggregation.

In this work, we propose a novel **L**ayer-**W**ise **R**e-weighting method from the perspective of feature shift, namely `Fed-LWR`, to improve the performance fairness of medical FL. The main idea of `Fed-LWR` is to quantify the differences in feature representations among different local models. To achieve this, we employ an averaged aggregated global model as the anchor and then estimate the layer-wise centered kernel alignment (CKA) [1,12] similarity between the local models and the anchor. The estimated similarity reflects the performance differences of the global model across all hospitals. Therefore, we assign larger weights to hospitals with larger feature differences, thereby minimizing the overall performance discrepancies of the global model across different hospitals. The new client weights are used to re-aggregate local models per layer, thereby obtaining a fairer global model. We evaluated our method on two federated medical image segmentation benchmarks. The results indicate that `Fed-LWR` achieves better fairness/accuracy trade-off compared with several state-of-the-art fair FL methods.

2 Method

2.1 Preliminaries

Federated Learning. Assume that there are K hospitals participating in a federated learning system, communicating through a trusted central server. Each hospital $k \in [K]$ has a private dataset D_k with n_k training samples $\{\boldsymbol{X}_i, \boldsymbol{Y}_i\}_{i=1}^{n_k}$ and trains a neural network $f = h^M \circ \cdots \circ h^2 \circ h^1$ with M layers. FL typically minimizes empirical risk $\mathcal{L}_k$ to optimize local models and updates the global model through averaging aggregation after each round t:

$$\mathcal{L} = \frac{1}{K}\sum_{k=1}^{K}\mathcal{L}_k(\boldsymbol{w}_k^t), \qquad \text{and} \quad \boldsymbol{w}_G^t = \frac{1}{K}\sum_{k=1}^{K}\boldsymbol{w}_k^t, \tag{1}$$

where $\boldsymbol{w}_G^t$ and $\boldsymbol{w}_k^t$ are the parameters of global and local models. However, in medical FL applications, data from different hospitals are collected using diverse

devices, resulting in differences in data distribution. Such disparities can impair the fairness of the global model across different hospitals. To better understand our objective, we introduce the definition of performance fairness from [16].

Definition 1. *(Performance Fairness) For two trained models $\boldsymbol{w}$ and $\hat{\boldsymbol{w}}$ from FL, the model $\hat{\boldsymbol{w}}$ provides a more fair solution if the performance of $\hat{\boldsymbol{w}}$ on the K hospitals is more uniform than the performance of $\boldsymbol{w}$ on the K hospitals.*

Centered Kernel Alignment. CKA [1,12] is a reliable technique for quantifying the similarity between pairs of neural network representations, which is widely used in various applications [11,25,23]. Let z_1 and z_2 be the features from an arbitrary layer of two neural networks, $\boldsymbol{U} \in \mathbb{R}^{n \times z_1}$ and $\boldsymbol{V} \in \mathbb{R}^{n \times z_2}$ denote their feature matrices on the same dataset with n samples, $\boldsymbol{K} = \boldsymbol{U}\boldsymbol{U}^\top \in \mathbb{R}^{n \times n}$ and $\boldsymbol{L} = \boldsymbol{V}\boldsymbol{V}^\top \in \mathbb{R}^{n \times n}$ are the Gram matrices of $\boldsymbol{U}$ and $\boldsymbol{V}$. Based on Hilbert-Schmidt Independence Criterion (HISC) [4], the calculation process of CKA similarity score $\delta \in [0, 1]$ between $\boldsymbol{U}$ and $\boldsymbol{V}$ can be expressed as follows:

$$\delta = \frac{\mathrm{HSIC}(\boldsymbol{K}, \boldsymbol{L})}{\sqrt{\mathrm{HSIC}(\boldsymbol{K}, \boldsymbol{K})\,\mathrm{HSIC}(\boldsymbol{L}, \boldsymbol{L})}}, \quad \mathrm{HSIC}(\boldsymbol{K}, \boldsymbol{L}) = \frac{\mathrm{vec}\left(\boldsymbol{K}'\right) \cdot \mathrm{vec}\left(\boldsymbol{L}'\right)}{(n-1)^2}, \tag{2}$$

where $\boldsymbol{K}'$ and $\boldsymbol{L}'$ are centered $\boldsymbol{K}$ and $\boldsymbol{L}$, $\mathrm{vec}(\cdot)$ is vectorization operation. For convenience, we use $\mathrm{CKA}(\cdot)$ to represent the above process.

2.2 Fed-LWR: Layer-wise Re-weighting for Federated Learning

In this section, we propose a novel layer-wise re-weighting method, `Fed-LWR`, by estimating the differences among clients from the perspective of feature shift. The overview of `Fed-LWR` is presented in Fig. 1 (a). Compared to standard FL framework [21], it incorporates additional representation difference estimation on the hospital and layer-wise re-aggregation process on the server after local training. The detailed algorithm is presented in Alg. 1.

Representation Difference Estimation. After the local training stage of t-th round, we first obtain a global model through Eq. (1) and send it to each hospital, which serves as an anchor. Then, we compute the CKA similarity between each layer of the local model and the anchor for every client. The above process at client k can be written as:

$$\delta_k = [\delta_k^1, \ldots, \delta_k^m, \ldots, \delta_k^M], \qquad \text{and} \quad \delta_k^m = \mathrm{CKA}(\boldsymbol{U}_k^m, \boldsymbol{V}_k^m), \tag{3}$$

where $\boldsymbol{U}_k^m$ and $\boldsymbol{V}_k^m$ are the feature matrices from the m-th layer of local model and anchor on dataset D_k.

$$\begin{aligned} \boldsymbol{U}_k^m &= h^m \circ \cdots \circ h^2 \circ h^1([\boldsymbol{X_1}, \boldsymbol{X_2}, \ldots, \boldsymbol{X_{n_k}}], \boldsymbol{w}_{k,m}^t \cup \cdots \cup \boldsymbol{w}_{k,2}^t \cup \boldsymbol{w}_{k,1}^t), \\ \boldsymbol{V}_k^m &= h^m \circ \cdots \circ h^2 \circ h^1([\boldsymbol{X_1}, \boldsymbol{X_2}, \ldots, \boldsymbol{X_{n_k}}], \boldsymbol{w}_{G,m}^t \cup \cdots \cup \boldsymbol{w}_{G,2}^t \cup \boldsymbol{w}_{G,1}^t), \end{aligned} \tag{4}$$

where $\boldsymbol{w}_{k,m}^t$ and $\boldsymbol{w}_{G,m}^t$ are the m-th layer parameters of local model and anchor. CKA can determine correspondences between hidden layers of neural networks

Algorithm 1: Fed-LWR

Input: K hospitals, communication rounds T, learning rate η,
Output: $\hat{\boldsymbol{w}}_G^T$

Initialize $\boldsymbol{w}_G^0$ // $\hat{\boldsymbol{w}}_G^0 = \boldsymbol{w}_G^0$
for *round* $t = 1, 2, ..., T$ **do**
  for *hospital* $k = 1, 2, ..., K$ ***parallelly*** **do**
    $\boldsymbol{w}_k^t \leftarrow \hat{\boldsymbol{w}}_G^{t-1}$
    $\boldsymbol{w}_k^t \leftarrow \boldsymbol{w}_k^t - \eta \nabla \mathcal{L}_k$ // Local training
  end
  $\boldsymbol{w}_G^t \leftarrow \frac{1}{K} \sum_{k=1}^{K} \boldsymbol{w}_k^t$ // Averaging aggregation for the anchor
  for *hospital* $k = 1, 2, ..., K$ ***parallelly*** **do**
    $\delta_k \leftarrow [\delta_k^1, \ldots, \delta_k^m, \ldots, \delta_k^M]$, $\delta_k^m \leftarrow \mathrm{CKA}(\boldsymbol{U}_k^m, \boldsymbol{V}_k^m)$ // Eq. (3)
  end
  $\rho_k \leftarrow [\rho_k^1, \ldots, \rho_k^m, \ldots, \rho_k^M]$, $\rho_k^m \leftarrow \frac{1-\delta_k^m}{\sum_{i=1}^{K}(1-\delta_i^m)}$ // Eq. (6)
  $\hat{\boldsymbol{w}}_G^t \leftarrow \{\hat{\boldsymbol{w}}_{G,m}^t\}_{m=1}^M$, $\hat{\boldsymbol{w}}_{G,m}^t \leftarrow \sum_{k=1}^{K} \rho_k^m \boldsymbol{w}_{k,m}^t$ // Eq. (3)
end
return $\hat{\boldsymbol{w}}_G^T$

trained under different conditions compared to traditional similarity metrics [12]. This best aligns with the situation of FL, as local models are trained on diverse datasets with different distributions.

Layer-wise Re-aggregation. The server collects the CKA similarity scores from each hospital, which reflect the local-global feature differences. Lower scores mean a greater difference, indicating that the global model is further from the local optimum, resulting in poorer performance at that hospital. Based on this, we convert the CKA similarity scores into aggregation weights:

$$\rho_k = [\rho_k^1, \ldots, \rho_k^m, \ldots, \rho_k^M], \qquad \text{and} \quad \rho_k^m = \frac{1 - \delta_k^m}{\sum_{i=1}^{K}(1 - \delta_i^m)}. \tag{5}$$

Eq. (5) assigns higher weights to hospitals with greater feature differences, and the sum of the total weight for each layer is 1. Finally, we use the new weights for layer-wise re-aggregation to get the fair global model:

$$\hat{\boldsymbol{w}}_G^t = \hat{\boldsymbol{w}}_{G,M}^t \cup \cdots \cup \hat{\boldsymbol{w}}_{G,m}^t \cup \cdots \cup \hat{\boldsymbol{w}}_{G,1}^t, \quad \text{and} \quad \hat{\boldsymbol{w}}_{G,m}^t = \sum_{k=1}^{K} \rho_k^m \boldsymbol{w}_{k,m}^t. \tag{6}$$

The fair global model $\hat{\boldsymbol{w}}_G^t$ will serve as the initial weights for the next round of local training.

3 Experiment

3.1 Experimental Setup

Datasets. To evaluate the effectiveness of our method, we conducted experiments on two medical image segmentation datasets: **ProstateMRI** [18] and

RIF [29], which are widely used in medical FL [8,10,27,28,32]. ProstateMRI collects T2-weighted MRI images from six different data sources for prostate segmentation. The size of all images has been processed to 384 × 384. We treat each data source as a client and divide the data into training, validation, and testing sets with a ratio of 6:2:2 [32]. RIF contains the retinal fundus images from four different clinical institutions for optic disc and cup segmentation. Following [27,28], we resize all images to 384 × 384. Since each sub-dataset has already been pre-divided into training and testing sets [29], we further split the training set into training and validation sets at a ratio of 4:1.

Baselines. We compared our method with following baselines: ❶ **Solo**: clients train models locally without communication; ❷ **FedAvg** [21]: the most popular FL method which updates the global model through parameter averaging; and several state-of-the-art fair FL methods including ❸ **q-FedAvg** [16]: a method re-weights the clients through empirical loss; ❹ **CFFL** [20]: it re-allocates the received model from the server by estimating the contribution of clients from validation results; ❺ **CGSV** [31]: it proposed a cosine gradient Shapley value to estimate the contribution of clients from the gradient; ❻ **Ditto** [14]: a personalized fair FL method to set different optimization objectives for clients; ❼ **FedCE** [8]: a fair federated medical image segmentation method by estimating the weights of clients simultaneously from gradients and validation loss. There are two different versions, *i.e.*, FedCE (Sum.) and FedCE (Multi.), that use addition and multiplication to merge contributions from gradients and losses, respectively. We use the Dice coefficient, a popular metric in medical image segmentation tasks, to evaluate the performance of the method. Following [16,14], we evaluate the performance fairness of methods by the standard deviation of testing performance across all clients.

Implementation Details. We implement all methods using PyTorch and conduct all experiments using an NVIDIA RTX 4090 GPU with 24GB of memory. Besides, we use U-Net [26] for the segmentation task and Dice loss as the optimization objective for the clients. The network is optimized by the Adam optimizer with a learning rate of 1e-3 and weight decay of 1e-4. The batch size is set to 8. We run 200 communication rounds with 1 local epoch and ensure all methods have converged stably. For a fair comparison, all methods adopt the same experimental settings.

3.2 Comparison with State-of-the-Arts

In Tables 1 and 2, we present the results of quantitative comparison for all methods on ProstateMRI and RIF, including the Dice scores on the testing sets of each client, as well as their average results and standard deviation. Apparently, with the proposed layer-wise re-weighting strategy, `Fed-LWR` significantly improves the performance of FedAvg from 88.78% to **93.25%** on ProstateMRI and from 84.63% to **87.17%** on RIF. Meanwhile, it also outperforms other fair FL methods on both two datasets. The above results indicate that our method can effectively alleviate the feature shift problem. Notably, FedAvg achieved lower performance on RIF compared to Solo, which is due to the high heterogeneity

Table 1: **Quantitative comparison** with Dice coefficient (%) on the testing set of ProstateMRI [18]. We report the performance of six clients, as well as their average result (`Avg.`) and standard deviation (`Std.`). The best results are marked in bold.

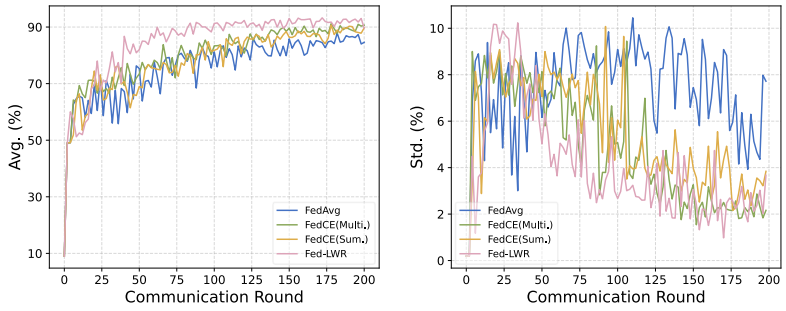
Method	`Site1`	`Site2`	`Site3`	`Site4`	`Site5`	`Site6`	`Avg.`	`Std.`
Solo	84.41	85.02	**93.70**	91.06	89.33	85.43	88.16	3.46
FedAvg [21]	85.84	92.27	93.58	87.89	92.47	80.66	88.78	4.54
q-FedAvg [16]	85.84	**94.93**	90.22	88.53	90.43	83.32	88.88	3.67
CFFL [20]	82.84	94.49	89.83	86.89	90.83	82.59	87.91	4.29
CGSV [31]	82.36	91.02	90.14	87.25	91.06	82.71	87.42	3.68
Ditto [14]	88.15	93.49	92.72	89.99	92.62	83.64	90.10	3.42
FedCE (Sum.) [8]	88.13	94.02	91.14	90.63	92.27	87.93	90.69	2.15
FedCE (Multi.) [8]	89.85	93.49	92.21	90.67	93.66	89.53	91.57	1.65
`Fed-LWR`	**91.92**	94.73	93.22	**93.34**	**94.06**	**92.21**	**93.25**	**0.97**

Table 2: **Quantitative comparison** with Dice coefficient (%) on the testing set of RIF [29]. We report the performance of four clients, as well as their average result (`Avg.`) and standard deviation (`Std.`). The best results are marked in bold.

Method	`Site1`	`Site2`	`Site3`	`Site4`	`Avg.`	`Std.`
Solo	82.01	74.57	91.26	**92.56**	85.10	7.31
FedAvg [21]	82.70	72.68	91.19	91.93	84.63	7.79
q-FedAvg [16]	77.83	**80.48**	91.72	91.62	85.41	6.32
CFFL [20]	81.49	74.97	91.33	89.53	84.33	6.55
CGSV [31]	78.79	77.30	89.77	91.78	84.41	6.42
Ditto [14]	82.11	78.91	90.90	91.95	85.97	5.58
FedCE (Sum.) [8]	**87.02**	76.60	90.92	90.28	86.21	5.73
FedCE (Multi.) [8]	86.00	77.42	90.69	89.93	86.01	5.26
`Fed-LWR`	85.89	79.43	**92.42**	90.96	**87.17**	**5.08**

of client data. In terms of performance fairness, all fair FL methods show better fairness compared with FedAvg. However, the fairness of some methods (*e.g.*, CFFL and CGSV) is achieved by sacrificing global performance, indicating that their fairness mechanisms may hinder the convergence of the global model to some extent. Compared to other methods, `Fed-LWR` achieved the best performance fairness on two datasets. This is attributed to `Fed-LWR` improving the performance of worse clients, *e.g.*, it improves the performance of `Site1` from 82.70% to **85.89%** and `Site2` from 72.68% to **79.43%** on RIF. The above results demonstrate the effectiveness of `Fed-LWR` in improving the fairness performance of the FL system.

Table 3: **Results of ablation study** on the testing set of ProstateMRI [18].



(a) (b)

Fig. 2: **(a) Avg. and (b) Std.** versus the number of communication rounds on the testing set of ProstateMRI [18].

3.3 Analytical Studies

Ablation Study. To provide more insights into `Fed-LWR`, we further delve into the design of our approach, *i.e.*, measurement of feature difference and layer-wise aggregation. First, we build up two variants of our method as follows: ❶ `Fed-LWR-v1`: we utilize cosine similarity, widely used in various applications [24,30], instead of CKA similarity to measure feature difference for `Fed-LWR`; ❷ `Fed-LWR-v2`: we only use the CKA similarity of the last layer of the U-Net encoder to estimate the weights of clients and aggregate the entire model. As shown in Table 3, the performance significance of cosine similarity is significantly lower than that of CKA similarity because cosine similarity is more suitable as an optimization objective to minimize the feature distance, and is difficult to estimate the similarity of two unrelated features. Besides, the comparison results between `Fed-LWR` and `Fed-LWR-v2` demonstrate the importance of layer-wise aggregation.

Convergence. We visualize the curve of testing performance versus communication rounds for FedAvg, FedCE, and our method. As shown in Fig. 2 (a), we can observe that `Fed-LWR` converges faster (stable after 90 rounds) compared to the other two methods. Interestingly, from Fig. 2 (b), `Fed-LWR` exhibits a larger standard deviation between rounds 20 to 50, as it is converging rapidly. This indicates that the fairness mechanism of `Fed-LWR` does not slow its convergence.

4 Conclusion

In this work, we improved the performance fairness of medical FL from the perspective of feature shift. Specifically, we proposed a novel fair FL framework, `Fed-LWR`, by estimating the feature differences between local models and the global model for layer-wise re-aggregation.

Future Work. We primarily focused our evaluation on segmentation tasks, as it is the most representative task in medical image analysis. Nevertheless, `Fed-LWR` is a versatile framework, and we will further evaluate its effectiveness on more medical image tasks (*e.g.*, classification or reconstruction) in future work.

Acknowledgments. This work is supported by the Guangzhou-HKUST(GZ) Joint Funding Program (No. 2023A03J0671), the Guangzhou Municipal Science and Technology Project (Grant No. 2023A03J0671), National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG2-TC-2021-003), and the Agency for Science, Technology and Research (A*STAR) through its AME Programmatic Funding Scheme Under Project A20H4b0141.

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

References

1. Cortes, C., Mohri, M., Rostamizadeh, A.: Algorithms for learning kernels based on centered alignment. The Journal of Machine Learning Research **13**(1), 795–828 (2012) 3, 4
2. Du, W., Xu, D., Wu, X., Tong, H.: Fairness-aware agnostic federated learning. In: Proceedings of the 2021 SIAM International Conference on Data Mining (SDM). pp. 181–189. SIAM (2021) 2
3. Feng, C.M., Yan, Y., Wang, S., Xu, Y., Shao, L., Fu, H.: Specificity-preserving federated learning for mr image reconstruction. IEEE Transactions on Medical Imaging (2022) 1
4. Gretton, A., Bousquet, O., Smola, A., Schölkopf, B.: Measuring statistical dependence with hilbert-schmidt norms. In: International conference on algorithmic learning theory. pp. 63–77. Springer (2005) 4
5. Guan, H., Liu, M.: Federated learning for medical image analysis: A survey. arXiv preprint arXiv:2306.05980 (2023) 1
6. Hu, Z., Shaloudegi, K., Zhang, G., Yu, Y.: Federated learning meets multi-objective optimization. IEEE Transactions on Network Science and Engineering **9**(4), 2039–2051 (2022) 2
7. Huang, W., Ye, M., Shi, Z., Li, H., Du, B.: Rethinking federated learning with domain shift: A prototype view. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16312–16322. IEEE (2023) 3
8. Jiang, M., Roth, H.R., Li, W., Yang, D., Zhao, C., Nath, V., Xu, D., Dou, Q., Xu, Z.: Fair federated medical image segmentation via client contribution estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16302–16311 (2023) 2, 6, 7

9. Jiang, M., Wang, Z., Dou, Q.: Harmofl: Harmonizing local and global drifts in federated learning on heterogeneous medical images. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 1087–1095 (2022) 1
10. Jiang, M., Yang, H., Cheng, C., Dou, Q.: Iop-fl: Inside-outside personalization for federated medical image segmentation. IEEE Transactions on Medical Imaging (2023) 1, 6
11. Kim, D., Han, B.: On the stability-plasticity dilemma of class-incremental learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20196–20204 (2023) 4
12. Kornblith, S., Norouzi, M., Lee, H., Hinton, G.: Similarity of neural network representations revisited. In: International conference on machine learning. pp. 3519–3529. PMLR (2019) 3, 4, 5
13. Li, T., Beirami, A., Sanjabi, M., Smith, V.: Tilted empirical risk minimization. In: International Conference on Learning Representations (2020) 2
14. Li, T., Hu, S., Beirami, A., Smith, V.: Ditto: Fair and robust federated learning through personalization. In: International Conference on Machine Learning. pp. 6357–6368. PMLR (2021) 2, 6, 7
15. Li, T., Sahu, A.K., Talwalkar, A., Smith, V.: Federated learning: Challenges, methods, and future directions. IEEE signal processing magazine **37**(3), 50–60 (2020) 1
16. Li, T., Sanjabi, M., Beirami, A., Smith, V.: Fair resource allocation in federated learning. International Conference on Learning Representations (2020) 2, 3, 4, 6, 7
17. Li, X., JIANG, M., Zhang, X., Kamp, M., Dou, Q.: Fedbn: Federated learning on non-iid features via local batch normalization. In: International Conference on Learning Representations (2020) 3
18. Liu, Q., Dou, Q., Yu, L., Heng, P.A.: Ms-net: multi-site network for improving prostate segmentation with heterogeneous mri data. IEEE transactions on medical imaging **39**(9), 2713–2724 (2020) 5, 7, 8
19. Liu, Q., Yang, H., Dou, Q., Heng, P.A.: Federated semi-supervised medical image classification via inter-client relation matching. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24. pp. 325–335. Springer (2021) 1
20. Lyu, L., Xu, X., Wang, Q., Yu, H.: Collaborative fairness in federated learning. Federated Learning: Privacy and Incentive pp. 189–204 (2020) 2, 6, 7
21. McMahan, B., Moore, E., Ramage, D., Hampson, S., y Arcas, B.A.: Communication-efficient learning of deep networks from decentralized data. In: Artificial intelligence and statistics. pp. 1273–1282. PMLR (2017) 1, 4, 6, 7, 8
22. Mohri, M., Sivek, G., Suresh, A.T.: Agnostic federated learning. In: International Conference on Machine Learning. pp. 4615–4625. PMLR (2019) 2
23. Nguyen, T., Raghu, M., Kornblith, S.: Do wide and deep networks learn the same things? uncovering how neural network representations vary with width and depth. In: International Conference on Learning Representations (2020) 4
24. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021) 8
25. Raghu, M., Unterthiner, T., Kornblith, S., Zhang, C., Dosovitskiy, A.: Do vision transformers see like convolutional neural networks? Advances in Neural Information Processing Systems **34**, 12116–12128 (2021) 4

26. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. pp. 234–241. Springer (2015) 6
27. Wang, J., Jin, Y., Stoyanov, D., Wang, L.: Feddp: Dual personalization in federated medical image segmentation. IEEE Transactions on Medical Imaging (2023) 1, 6
28. Wang, J., Jin, Y., Wang, L.: Personalizing federated medical image segmentation via local calibration. In: European Conference on Computer Vision. pp. 456–472. Springer (2022) 6
29. Wang, S., Yu, L., Li, K., Yang, X., Fu, C.W., Heng, P.A.: Dofe: Domain-oriented feature embedding for generalizable fundus image segmentation on unseen datasets. IEEE Transactions on Medical Imaging **39**(12), 4237–4248 (2020) 6, 7
30. Xie, J., Xiang, J., Chen, J., Hou, X., Zhao, X., Shen, L.: C2am: Contrastive learning of class-agnostic activation map for weakly supervised object localization and semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 989–998 (2022) 8
31. Xu, X., Lyu, L., Ma, X., Miao, C., Foo, C.S., Low, B.K.H.: Gradient driven rewards to guarantee fairness in collaborative machine learning. Advances in Neural Information Processing Systems **34**, 16104–16117 (2021) 2, 6, 7
32. Zhou, T., Konukoglu, E.: Fedfa: Federated feature augmentation. In: International Conference on Learning Representations (2023) 3, 6