GRACE: A Generalized and Personalized Federated Learning Method for Medical Imaging

Ruipeng Zhang[1,2], Ziqing Fan[1,2], Qinwei Xu[1,2], Jiangchao Yao[1,2(✉)], Ya Zhang[1,2], and Yanfeng Wang[1,2(✉)]

[1] CMIC, Shanghai Jiao Tong University, Shanghai, China
[2] Shanghai AI Laboratory, Shanghai, China
{sunarker,wangyanfeng}@sjtu.edu.cn

Abstract. Federated learning has been extensively explored in privacy-preserving medical image analysis. However, the domain shift widely existed in real-world scenarios still greatly limits its practice, which requires to consider both generalization and personalization, namely generalized and personalized federated learning (GPFL). Previous studies almost focus on the partial objective of GPFL: personalized federated learning mainly cares about its local performance, which cannot guarantee a generalized global model for unseen clients; federated domain generalization only considers the out-of-domain performance, ignoring the performance of the training clients. To achieve both objectives effectively, we propose a novel GRAdient CorrEction (GRACE) method. GRACE incorporates a feature alignment regularization under a meta-learning framework on the client side to correct the personalized gradients from overfitting. Simultaneously, GRACE employs a consistency-enhanced reweighting aggregation to calibrate the uploaded gradients on the server side for better generalization. Extensive experiments on two medical image benchmarks demonstrate the superiority of our method under various GPFL settings. Code available at https://github.com/MediaBrain-SJTU/GPFL-GRACE.

Keywords: Federated learning · Domain generalization · Domain shift

1 Introduction

Data-driven deep networks maintain the great potential to achieve superior performance under the large-scale medical data [23,30,34]. However, privacy concerns from patients and institutions prevent centralized training from access to the data in multiple centers. Federated learning (FL) thereby becomes a promising compromise that preserves privacy by distributing the model to data sources to train a global model without sharing their data directly [27].

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43898-1_2.

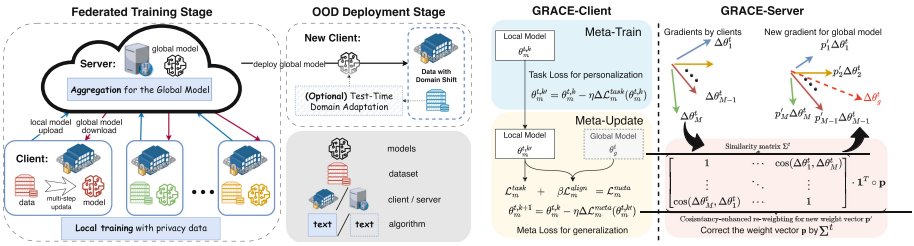
H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14222, pp. 14–24, 2023.
https://doi.org/10.1007/978-3-031-43898-1_2

A major challenge to FL in real-world scenarios is domain shift, which refers to the difference of marginal data distributions across centers and induces significant performance degradation [9,22,30]. Current methods to address the problem of domain shift can be categorized into two directions. One is federated domain generalization (FedDG) [5,22], which tackles the domain shift between training and testing clients. FedDG aims at obtaining a generalizable global model, but the optimal performance on local training clients cannot be guaranteed. Another direction is personalized federated learning (PFL) [1,7,18,20,29,31], which tackles the domain shifts among training clients by personalizing the global model locally. However, both FedDG and PFL only consider the partial objective of GPFL, ignoring either personalization or generalization in real-world scenarios.

In this paper, we focus on generalized and personalized federated learning (GPFL), which considers both generalization and personalization to holistically combat the domain shift. We notice that one recent work IOP-FL [13] also studied the problem of GPFL, but it mainly resorts to the model personalization for the unseen clients by test-time training on deployment stage, which did not directly consider enhancing both objectives of GPFL in the training phase.

We seek a more effective and efficient solution to GPFL in this work. Specifically, our intuition is based on the following conjecture: a more generalizable global model can facilitate the local models to better adapt to the corresponding local distribution, and better adapted local models can then provide positive feedbacks to the global model with improved gradients.

Based on the above intuition, we propose a novel method named GRAdient CorrEction (GRACE) that can achieve both generalization and personalization during training by enhancing the model consistency on both the client side and the server side. By analyzing the federated training stage in Fig. 1(a), we discover a significant discrepancy in the feature distributions between the global and local models due to domain shifts, which will influence the training process on both client and server side. To address this problem, we aim to correct the inconsistent gradients on both sides. On the client side, we leverage a meta-learning strategy to align the feature spaces of global and local models while fitting the local data distribution, as depicted in Fig. 1(b). Furthermore, on the server side,



(a) Federated learning framework (b) GRACE on the client and server sides

Fig. 1. The overall framework of (a) the federated learning system which has two stages for algorithm design and (b) our GRACE method on both the client and server sides.

we estimate the gradient consistency by computing the cosine similarity among the gradients from clients and re-weight the aggregation weights to mitigate the negative effect of domain shifts on the global model update. Through these two components, GRACE preserves the generalizability in global model as much as possible when personalizing it on local distributions. Comprehensive experiments are conducted on two medical FL benchmarks, which show that GRACE outperforms state-of-the-art methods in terms of both generalization and personalization. We also perform insightful analysis to validate the rationale of our algorithm design.

2 Method

2.1 Overview of the GPFL Framework

Consider a federated learning scenario where M clients possess local private data with one unique domain per client. Let $\mathcal{D} = \{\mathcal{D}_1, \mathcal{D}_2, \cdots, \mathcal{D}_M\}$ be the set of M distributed training source domains, and $\mathcal{D}_o$ be the distribution of the unseen client. We denote $P(\mathcal{X}, \mathcal{Y})$ as the joint input and label space of a task. For client $m = 1, 2, \ldots, M$, the local dataset $\mathcal{D}_m = \{(\mathbf{x}_i^m, y_i^m)\}_{i=1}^{N_m}$, with $N_m = |\mathcal{D}_m|$ and $N = \sum_{m=1}^{M} N_m$ being the number of local samples and total samples respectively. Let $\mathcal{L}_m$ be the task loss function of the local model $f(\mathbf{x}^m; \theta_m)$, and $f_g(\mathbf{x}; \theta_g)$ be the global model. The goal of GPFL is to learn a generalized global model f_g that minimizes the expected risk $\mathbb{E}_{(\mathbf{x},y)\in\mathcal{D}_o}[\mathcal{L}_o(f(\mathbf{x}; \theta_g), y)]$ on unseen clients while enabling each participating client to have a personalized local model f that adapts to its local data distribution $\mathcal{D}_m$.

We characterize the GPFL system in Fig. 1(a), which consists of the federated training stage and the OOD deployment stage. For federated training, there are two iterative phases, namely local training on the client's private data and server aggregation for the global model. The Standard method FedAvg [27] optimizes a global model by an empirical objective $\min_\theta \sum_{m=1}^{M} p_m \mathcal{L}_m(\ f(x; \theta), y)$, which is implemented as a weighted aggregation of local models $\{\theta_m\}_{m=1}^{M}$ trained in the clients ($\theta_g = \sum_{m=1}^{M} p_m \theta_m$). Here $p_m = \frac{N_m}{N}$ and we denote the weight vector as $\mathbf{p} = [p_1, \ldots, p_M] \in \mathbb{R}^M$. This method implicitly assumes that all clients share the same data distribution, thus failing to adapt to domain shift scenarios.

To solve the GPFL problem, we propose a GRAdient CorrEction (GRACE) method for both local client training and server aggregation during the federated training stage. Unlike IPO-FL [13], our method constrains the model's generalizability and personality only during the federated training stage. Our motivation comes from an observation that domain shift causes a significant mismatch between the feature spaces of the local models and the initial global model after client training at each round. It leads to inconsistent gradients uploaded from the biased local models, which hurts the generalizability of the global model. Therefore, in GRACE, we alleviate the inconsistency between gradients obtained from different clients. The details of GRACE are elaborated in the following sections.

2.2 Local Training Phase: Feature Alignment & Personalization

We calibrate the gradient during local training via a feature alignment constraint by meta-learning, which preserves the generalizable feature while adapting to the local distributions. We conduct the client-side gradient correction in two steps.
Meta-Train: We denote $\theta_m^{t,k}$ as the personalized model parameter at the local update step k of client m in round t and η as the local learning rate. The first step is a personalization step by the task loss:

$$\theta_m^{t,k\prime} = \theta_m^{t,k} - \eta \Delta \mathcal{L}_m^{task}(\theta_m^{t,k}) = \theta_m^{t,k} - \eta \Delta \mathcal{L}_m(f(\mathbf{x}_{tr}, \theta_m^{t,k}), y_{tr}) \tag{1}$$

where $(\mathbf{x}_{tr}, y_{tr}) \in \mathcal{D}_m$ is the sampled data and $\theta_m^{t,k\prime}$ is the updated parameter that will be used in the second step.

Meta-Update: After optimizing the local task objective, we need a meta-update to virtually evaluate the updated parameters $\theta_m^{t,k\prime}$ on the held-out meta-test data $(\mathbf{x}_{te}, y_{te}) \in \mathcal{D}_m$ with a meta-objective $\mathcal{L}_m^{meta}$. We add the feature alignment regularizer into the loss function of meta-update:

$$\mathcal{L}_m^{meta}(\mathbf{x}_{te}, y_{te}; \theta_m^{t,k'}) = \mathcal{L}_m(f(\mathbf{x}_{te}, \theta_m^{t,k}), y_{te}) + \beta \mathcal{L}_m^{align}(h(\mathbf{x}_{te}; \phi_g^t), h(\mathbf{x}_{te}; \phi_m^{t,k'}))) \tag{2}$$

where $h(\cdot;\phi)$ is the feature extractor part of model $f(\cdot;\theta)$ and ϕ is the corresponding parameter, and β is the weight for alignment loss which has a default value of 1.0. Here, we apply three widely-used alignment losses to minimize the discrepancy in the feature space: CORAL [32], MMD [16], and adversarial training [11]. We show that varying alignment loss functions can boost the generalization capability during local training and report the results in Table 4.

2.3 Aggregation Phase: Consistency-Enhanced Re-weighting

We introduce a novel aggregation method on the server side that corrects the global gradient by enhancing the consistency of the gradients received from training clients. We measure the consistency of two gradient vectors by their cosine similarity and use the average cosine similarity among all uploaded gradients as an indicator of gradient quality on generalization. The main idea is to balance the quantity and quality of data represented by each client's gradients and generate more robust global model parameters by reducing the influence of inconsistent gradients. It is vital for medical scenarios, and existing methods that only measure gradient importance by data size need to be revised.

Let $\Delta\theta_m^t = \theta_m^t - \theta_g^t$ be the gradient of t round from each client after local training, where θ_m^t and θ_g^t are the local and global models on client m at round t. The similarity matrix is $\Sigma^t = \{\sigma_{ij}^t\}_{i,j=1,\cdots,M}$, where $\sigma_{ij}^t = \cos(\Delta\theta_i^t, \Delta\theta_j^t) \in \mathbb{R}^{M\times M}$. The corrected global gradient is:

$$\Delta\theta_g^t = [\Delta\theta_1^t, \cdots, \Delta\theta_M^t] \cdot \mathbf{p'}^T, \mathbf{p'} = \mathbf{Norm}[(\Sigma^t \cdot \mathbf{1}^T) \circ \mathbf{p}]. \tag{3}$$

"$\circ$" means element multiplication of two vectors and "**Norm**" means to normalize the new weight vector $\mathbf{p'}$ with $\sum_{m=1}^{M} p'_m = 1, p'_m \in (0,1)$. Then the updated

global model for round $t+1$ will be $\theta_g^{t+1} = \theta_g^t - \eta_g \Delta\theta_g^t$, where η_g is the global learning rate with default value 1.

Theoretical Analysis: We prove that the re-weighting method in Eq.(3) will enhance the consistency of the global gradient based on the FedAvg. First, we define the averaged cosine similarity of $\Delta\theta_m^t$ as c_m^t, where $c_m^t = \frac{1}{M}\sum_{n=1}^{M}\sigma_{mn}^t$. Then, the consistency degree of the global gradient in FedAvg is $\sum_{m=1}^{M} p_m c_m^t$. Thus, the consistency degree after applying our GRACE method has

$$\sum_{j=1}^{M} \frac{p_j {c_j^t}^2}{\sum_{m=1}^{M} p_m c_m^t} \geq \sum_{m=1}^{M} p_m c_m^t, \text{ if } \forall m \in [1, M], p_m \leq \frac{1}{\sqrt{M}}, . \tag{4}$$

Note that, we can easily prove the inequality in Eq. (4) by using the *generalized arithmetic-geometric mean inequality*, which holds when $c_1^t = \cdots = c_M^t$.

Table 1. Personalization results for Fed-ISIC2019 and Fed-Prostate benchmark.

Method	Fed-ISIC2019							Fed-Prostate						
	0	1	2	3	4	5	avg	A	B	C	D	E	F	avg
FedAvg	66.28	67.36	63.27	69.71	71.36	65.29	67.21	90.71	90.29	89.65	91.36	89.50	89.30	90.14
FedRep	68.03	67.93	65.41	69.15	68.59	74.96	69.01	91.20	91.00	89.95	91.86	91.60	90.40	91.00
Ditto	67.05	70.39	68.79	69.16	69.75	75.52	70.11	91.09	90.83	90.91	91.55	91.23	89.87	90.91
FedMTL	69.45	71.07	69.37	71.85	72.29	76.77	71.97	91.09	90.83	90.91	91.55	91.23	89.87	90.91
FedPer	71.90	73.00	73.18	75.05	74.45	76.54	74.02	93.20	93.66	93.44	93.66	93.25	92.99	93.37
FedBABU	74.40	74.34	73.34	75.96	76.98	77.64	75.44	92.09	93.25	91.33	91.54	92.48	91.12	91.97
FedBN	74.74	77.13	75.42	78.19	78.80	78.75	**77.17**	93.08	92.82	93.20	93.40	93.07	93.17	93.13
FedRoD	71.98	71.48	72.05	75.11	75.70	76.64	73.83	92.90	92.92	92.90	93.37	93.74	93.04	93.14
Per-FedAvg	75.50	77.64	75.00	76.52	75.72	76.80	76.20	93.54	93.93	93.75	94.22	93.89	93.40	*93.79*
pFedMe	71.75	71.08	68.40	71.60	74.02	77.55	72.40	89.20	88.86	88.97	89.97	89.66	88.73	89.24
GRACE	75.71	75.40	74.57	79.08	79.48	78.60	*77.14*	93.56	93.72	94.08	94.23	93.91	93.44	**93.82**

3 Experiments

3.1 Dataset and Experimental Setting

We evaluate our approach on two open source federated medical benchmarks: **Fed-ISIC2019** and **Fed-Prostate**. The former is a dermoscopy image classification task [9] among eight different melanoma classes, which is a 6-client federated version of ISIC2019 [6,8,34]. And the latter is a federated prostate segmentation task [13] with T2-weighted MRI images from 6 different domains [15,21,23,28]. We follow the settings of [9] for Fed-ISIC2019 and [22] for Fed-Prostate.

3.2 Comparison with SOTA Methods

We conduct the leave-one-client-out experiment for both benchmarks. In each experiment, one client is selected as the unseen client and the model is trained on the remaining clients. The average performance on all internal clients' test set is the in-domain **personalization** results, while the unseen client's performance is the out-of-domain **generalization** results. The final results of each method are the average of all leave-one-domain-out splits, and all results are over three independent runs. (Please see the details of experimental setup in open-source code.)

Performance of In-Domain Personalization. For a fair comparison, the baseline method **FedAvg** [27] and several current SOTA personalized FL methods are chosen. **Ditto** [18] and **FedMTL** [31] treat the personalized process as a kind of multi-task learning and generate different model parameters for each client. **FedBN** [20] keeps the parameters of BatchNorm layers locally as considering those parameters to contain domain information. **FedRep** [1], **FedBABU** [29], **FedPer** [1] and **FedRoD** [4] all use a personalized head to better fit the local data distribution, where FedRoD also retains a global head for OOD generalization. Besides FedRoD, **PerFedAvg** [10] and **pFedMe** [33] can also implement personalized and generalized in FL by federated meta-learning. These three methods are also involved in the comparison of out-of-domain generalization. Table 1 presents the results of two tasks. GRACE achieves comparable personalization performance on the in-domain clients with SOTA methods. Note that GRACE outperforms most of PFL methods in the table and other methods like FedRoD and Per-FedAvg also achieve good results, which means that improving generalization is also beneficial for model personalization.

Table 2. Generalization results for Fed-ISIC2019 and Fed-Prostate benchmark.

Method	Fed-ISIC2019							Fed-Prostate						
	0	1	2	3	4	5	avg	A	B	C	D	E	F	avg
FedAvg	36.94	62.70	54.25	40.86	39.68	73.37	51.30	89.57	88.63	82.69	85.39	79.21	89.74	85.87
ELCFS	37.08	69.37	62.63	38.48	38.44	72.47	53.08	89.91	90.80	84.89	88.19	83.88	87.24	*87.47*
FedProx	34.13	62.77	64.35	36.52	40.25	75.47	52.25	90.55	87.69	83.27	85.42	79.05	90.18	86.03
HarmoFL	36.80	74.00	65.74	43.63	46.75	70.21	*56.19*	92.07	89.17	83.60	85.55	81.86	90.00	87.04
Scaffold	36.31	60.83	70.60	40.11	41.37	73.10	53.72	90.47	87.98	84.15	85.27	81.56	89.37	86.47
MOON	35.54	61.25	71.53	38.82	44.22	68.26	53.27	89.04	84.12	82.64	85.22	75.38	87.05	83.91
FedRoD	45.36	65.12	66.33	47.36	39.36	69.79	55.55	90.09	88.67	81.59	86.88	78.95	87.90	85.68
Per-FedAvg	39.44	61.88	69.50	42.63	41.13	70.89	54.24	92.08	89.63	84.94	87.11	78.17	89.35	86.88
pFedMe	38.05	63.40	64.08	42.82	42.63	74.33	54.22	83.86	82.68	82.52	78.93	78.54	87.30	82.30
GRACE	43.57	76.52	73.58	44.80	41.74	68.99	**58.20**	91.53	90.15	84.28	87.55	81.39	92.37	**87.88**

Performance of Out-of-Domain Generalization. For out-of-domain comparison, we select several FL methods that aim to solve the data heterogeneity problem, such as **FedProx** [19], **Scaffold** [14] and **MOON** [17]. Some FedDG methods are also chosen for comparison, like **ELCFS** [22] and **HarmoFL** [12].

Table 3. Generalization results for two benchmarks with test-time adaptation.

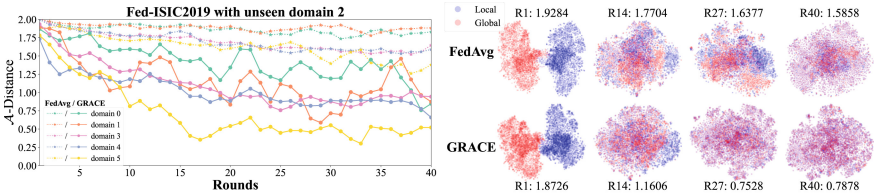
Fed-ISIC2019	0	1	2	3	4	5	avg
FedAvg	$36.94_{0.52}$	$62.70_{0.71}$	$54.25_{3.17}$	$40.86_{2.48}$	$39.68_{3.34}$	$73.37_{4.49}$	$51.30_{2.45}$
DSBN	$30.26_{0.28}$	$69.48_{1.91}$	$66.22_{1.34}$	$39.71_{0.97}$	$38.91_{0.08}$	$75.69_{0.84}$	$53.38_{0.90}$
Tent	$32.46_{0.34}$	$65.45_{1.64}$	$65.87_{1.23}$	$37.51_{1.24}$	$41.38_{0.84}$	$75.20_{1.07}$	$52.98_{1.06}$
IOP-FL	$33.88_{0.03}$	$65.05_{0.75}$	$71.49_{1.88}$	$40.91_{0.90}$	$41.30_{0.77}$	$70.32_{0.35}$	$53.82_{0.78}$
GRACE + DSBN	$38.21_{0.06}$	$78.91_{2.06}$	$71.28_{2.27}$	$48.62_{2.34}$	$44.29_{1.18}$	$70.18_{1.32}$	$\underline{58.58}_{1.54}$
GRACE + Tent	$38.64_{0.02}$	$74.49_{2.03}$	$73.10_{1.04}$	$51.93_{1.45}$	$44.96_{1.01}$	$78.19_{1.37}$	$\mathbf{60.22}_{1.15}$
Fed-Prostate	A	B	C	D	E	F	avg.
FedAvg	$89.57_{0.95}$	$88.63_{1.10}$	$82.69_{1.77}$	$85.39_{0.40}$	$79.21_{1.38}$	$89.74_{0.90}$	$85.87_{1.08}$
DSBN	$89.44_{0.50}$	$88.24_{0.12}$	$85.08_{0.34}$	$85.75_{0.22}$	$81.50_{0.32}$	$89.83_{0.24}$	$86.64_{0.29}$
Tent	$90.35_{0.12}$	$85.93_{0.68}$	$86.06_{0.13}$	$88.22_{0.88}$	$81.59_{0.19}$	$91.76_{0.08}$	$87.32_{0.35}$
IOP-FL	$90.52_{0.33}$	$90.52_{0.64}$	$88.32_{0.41}$	$89.39_{0.36}$	$84.33_{0.21}$	$92.61_{0.23}$	$\underline{89.28}_{0.36}$
GRACE + DSBN	$92.60_{0.28}$	$91.11_{0.33}$	$87.13_{0.74}$	$88.65_{0.76}$	$84.45_{0.15}$	$92.82_{0.58}$	$\mathbf{89.46}_{0.47}$
GRACE + Tent	$92.59_{0.28}$	$90.32_{0.45}$	$87.39_{0.47}$	$89.03_{0.35}$	$84.58_{0.58}$	$91.63_{0.48}$	$89.26_{0.44}$

Table 4. Ablation studies on the client & server side of our GRACE.

Method	GRACE	GRACE-Client				Fed-ISIC2019		Fed-Prostate	
	-Server	Adv	CORAL	MMD	MOON	P	G	P	G
FedAvg	–	–	–	–	–	$67.21_{0.95}$	$51.30_{2.45}$	$90.14_{0.36}$	$85.87_{1.08}$
MOON	–	–	–	–	✓	$67.78_{3.39}$	$53.27_{2.84}$	$91.72_{0.65}$	$83.91_{1.75}$
+GRACE	✓	–	–	–	✓	$70.43_{2.67}$	$56.04_{2.58}$	$92.20_{0.43}$	$85.90_{0.66}$
Model A	✓	–	–	–	–	–	$55.97_{3.35}$	-	$86.64_{1.13}$
Model B	–	✓	–	–	–	$75.26_{3.71}$	$53.86_{1.71}$	$93.57_{0.22}$	$87.02_{0.75}$
Model C	–	–	✓	–	–	$77.00_{3.14}$	$54.55_{2.21}$	$93.72_{0.17}$	$87.63_{0.97}$
Model D	–	–	–	✓	–	$75.75_{2.99}$	$52.60_{3.32}$	$93.03_{0.30}$	$87.31_{0.62}$
Model E	✓	✓	–	–	–	$77.11_{2.79}$	$57.09_{2.23}$	$93.71_{0.45}$	$\mathbf{88.04}_{0.48}$
Model F	✓	–	✓	–	–	$\mathbf{77.14}_{2.64}$	$\mathbf{58.20}_{1.92}$	$93.51_{0.59}$	$87.39_{0.68}$
Model G	✓	–	–	✓	–	$76.33_{2.85}$	$58.06_{2.95}$	$\mathbf{93.82}_{0.41}$	$87.88_{0.46}$

The results are summarized in Table 2, and according to the comparison, GRACE shows a significant improvement in the unseen client. Combining Table 1 and 2, it indicates that our gradient correction on both the local and global sides effectively enhances the generalization of the global model based on personalization.

Generalization with TTDA. Considering the promise of recent test-data domain adaptation (TTDA) techniques for domain generalization in medical imaging [13,25], we also compare GRACE under the TTDA scenario with **IOP-FL** [13], **DSBN** [3] and **Tent** [35]). As shown in Table 3, GRACE obtains better results combined with DSBN and Tent, which means our method is orthogonal with TTDA, and TTDA can benefit from our generalized global model.



(a) Curves of $\mathcal{A}$-Distance (b) t-SNE of global and local features

Fig. 2. $\mathcal{A}$-distance between global and local models on FedAvg and GRACE. (a) Comparison of change curves over training rounds on each training client (dashed line: FedAvg; solid line: GRACE). (b) t-SNE of global and local features at different rounds. (We mark the $\mathcal{A}$-distance corresponding to the round.)

3.3 Further Analysis

Ablation Study on Different Parts of Our Method. The detailed ablation studies are shown in Table 4 to further validate the effectiveness of each component in GRACE. Three widely-used approaches from domain adaptation/generalization area are used for feature alignment loss in local training. In Table 4, "Adv." [11] means using adversarial training between features from the global model and the local model, and "CORAL" [32] and "MMD" [16] are classic regularization losses for domain alignment. From the table, our framework can obtain performance improvements on different alignment approaches. Considering both performance and efficiency, we prefer "CORAL" for alignment. Add the server-side correction on top of MOON can also obtain some gains, but the overall effect is limited, since its alignment loss might push the current feature away from features in previous round and thus reduces the discriminativeness.

Visualization of the Feature Alignment Loss in Eq. (2). The $\mathcal{A}$-distance measurement is used to evaluate the dissimilarity between the local and global models, which is suggested to measure the cross-domain discrepancy in the domain adaptation theory [2]. We follow the proxy implementation in [24] and trace the curve of $\mathcal{A}$-distance on FedAvg and our method on each client throughout the training process in Fig. 2(a). The curves demonstrate a substantial reduction of feature discrepancy compared with FedAvg. It validates the efficacy of our algorithm design and corroborates our claim that generalization and personalization are compatible objectives. Our personalized local models can close the distance with the global model while preserving a good fit for local data distribution. In addition, we use t-SNE [26] to visualize the feature distributions in Fig. 2(b) and GRACE can reduce the discrepancy between global and local features.

Loss Curves Comparison of FedAvg and Our Method. Fig. 3 shows the loss curves of FedAvg and our GRACE method with only server-side aggregation (Model A in Table 4). Our method can achieve better global minimization.

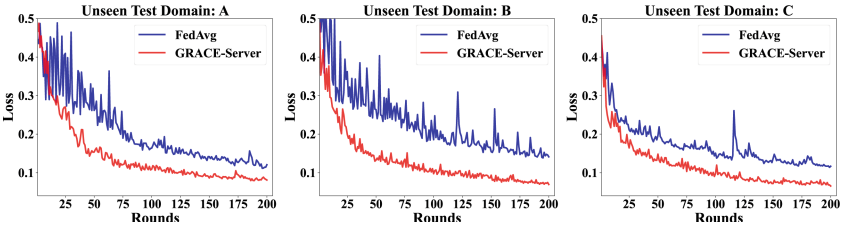


Fig. 3. Loss curves of FedAvg and GRACE-Server in 3 experiments of Fed-Prostate.

4 Conclusion

We introduce GPFL for multi-center distributed medical data with domain shift problems, which aims to achieve both generalization for unseen clients and personalization for internal clients. Existing approaches only focus on either generalization (FedDG) or personalization (PFL). We argue that a more generalizable global model can facilitate the local models to adapt to the clients' distribution, and the better-adapted local models can contribute higher quality gradients to the global model. Thus, we propose a new method GRAdient CorrEction (GRACE), which corrects the model gradient at both the client and server sides during training to enhance both the local personalization and the global generalization. The experimental results on two medical benchmarks show that GRACE can enhance both local adaptation and global generalization and outperform existing SOTA methods in generalization and personalization.

Acknowledgments. The data used in this paper are open by previous works, *i.e.*, Fed-Prostate in [22]; Fed-ISIC2019 in [9] with licence CC-BY-NC 4.0. This work is supported by the National Key R&D Program of China (No. 2022ZD0160702), STCSM (No. 22511106101, No. 18DZ2270700, No. 21DZ1100-100), 111 plan (No. BP0719010), and State Key Laboratory of UHD Video and Audio Production and Presentation. Ruipeng Zhang is partially supported by Wu Wen Jun Honorary Doctoral Scholarship, AI Institute, Shanghai Jiao Tong University.

References

1. Arivazhagan, M.G., Aggarwal, V., Singh, A.K., Choudhary, S.: Federated learning with personalization layers. arXiv preprint arXiv:1912.00818 (2019)
2. Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., Vaughan, J.W.: A theory of learning from different domains. Mach. Learn. **79**, 151–175 (2010)
3. Chang, W.G., You, T., Seo, S., Kwak, S., Han, B.: Domain-specific batch normalization for unsupervised domain adaptation. In: CVPR, pp. 7354–7362 (2019)
4. Chen, H.Y., Chao, W.L.: On bridging generic and personalized federated learning for image classification. In: ICLR (2022)
5. Chen, J., Jiang, M., Dou, Q., Chen, Q.: Federated domain generalization for image recognition via cross-client style transfer. In: WACV, pp. 361–370 (2023)

6. Codella, N.C., Gutman, D., et al.: Skin lesion analysis toward melanoma detection: a challenge at the 2017 International Symposium on Biomedical Imaging (ISBI), Hosted by the International Skin Imaging Collaboration (ISIC). In: ISBI (2018)
7. Collins, L., Hassani, H., Mokhtari, A., Shakkottai, S.: Exploiting shared representations for personalized federated learning. In: ICML, pp. 2089–2099 (2021)
8. Combalia, M., et al.: Bcn20000: dermoscopic lesions in the wild. arXiv preprint arXiv:1908.02288 (2019)
9. Du Terrail, J.O., et al.: FLamby: datasets and benchmarks for cross-silo federated learning in realistic healthcare settings. In: NeurIPS, Datasets and Benchmarks Track (2022)
10. Fallah, A., Mokhtari, A., Ozdaglar, A.: Personalized federated learning: a meta-learning approach. arXiv preprint arXiv:2002.07948 (2020)
11. Ganin, Y., et al.: Domain-adversarial training of neural networks. J. Mach. Learn. Res. **17**(1), 2030–2096 (2016)
12. 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)
13. Jiang, M., Yang, H., Cheng, C., Dou, Q.: IOP-FL: inside-outside personalization for federated medical image segmentation. arXiv preprint arXiv:2204.08467 (2022)
14. Karimireddy, S.P., Kale, S., Mohri, M., et al.: Scaffold: stochastic controlled averaging for federated learning. In: ICML, pp. 5132–5143 (2020)
15. Lemaître, G., Martí, R., Freixenet, J., et al.: Computer-aided detection and diagnosis for prostate cancer based on mono and multi-parametric MRI: a review. Comput. Biol. Med. **60**, 8–31 (2015)
16. Li, H., Pan, S.J., Wang, S., Kot, A.C.: Domain generalization with adversarial feature learning. In: CVPR, pp. 5400–5409 (2018)
17. Li, Q., He, B., Song, D.: Model-contrastive federated learning. In: CVPR, pp. 10713–10722 (2021)
18. Li, T., Hu, S., Beirami, A., Smith, V.: Ditto: fair and robust federated learning through personalization. In: ICML, pp. 6357–6368 (2021)
19. Li, T., Sahu, A.K., Zaheer, M., Sanjabi, M., Talwalkar, A., Smith, V.: Federated optimization in heterogeneous networks. MLSys **2**, 429–450 (2020)
20. Li, X., JIANG, M., Zhang, X., Kamp, M., Dou, Q.: FedBN: federated learning on non-IID features via local batch normalization. In: ICLR (2021)
21. Litjens, G., Toth, R., et al.: Evaluation of prostate segmentation algorithms for MRI: the promise12 challenge. Med. Image Anal. **18**(2), 359–373 (2014)
22. Liu, Q., Chen, C., Qin, J., Dou, Q., Heng, P.A.: FedDG: federated domain generalization on medical image segmentation via episodic learning in continuous frequency space. In: CVPR, pp. 1013–1023 (2021)
23. Liu, Q., Dou, Q., et al.: MS-net: multi-site network for improving prostate segmentation with heterogeneous MRI data. IEEE TMI **39**(9), 2713–2724 (2020)
24. Long, M., Cao, Y., Cao, Z., Wang, J., Jordan, M.I.: Transferable representation learning with deep adaptation networks. IEEE TPAMI **41**(12), 3071–3085 (2019)
25. Ma, W., Chen, C., Zheng, S., Qin, J., Zhang, H., Dou, Q.: Test-time adaptation with calibration of medical image classification nets for label distribution shift. In: MICCAI, pp. 313–323 (2022)
26. van der Maaten, L., Hinton, G.: Visualizing data using t-SNE. In: JMLR 9 (2008)
27. McMahan, B., Moore, E., et al.: Communication-efficient learning of deep networks from decentralized data. In: AISTATS, pp. 1273–1282 (2017)
28. Nicholas, B., et al.: NCI-proc. IEEE-ISBI conference 2013 challenge: automated segmentation of prostate structures. Can. Imaging Arch. (2015)

29. Oh, J., Kim, S., Yun, S.Y.: FedBABU: toward enhanced representation for federated image classification. In: ICLR (2022)
30. Roth, H.R., Chang, K., et al.: Federated learning for breast density classification: a real-world implementation. In: MICCAI Workshop, pp. 181–191 (2020)
31. Smith, V., et al.: Federated multi-task learning. In: NeurIPS, vol. 30 (2017)
32. Sun, B., Saenko, K.: Deep coral: correlation alignment for deep domain adaptation. In: ECCV, pp. 443–450 (2016)
33. T Dinh, C., Tran, N., Nguyen. J.: Personalized federated learning with moreau envelopes. NeurIPS **33**, 21394–21405 (2020)
34. Tschandl, P., Rosendahl, C., Kittler, H.: The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Sci. Data **5**(1), 1–9 (2018)
35. Wang, D., Shelhamer, E., Liu, S., Olshausen, B., Darrell, T.: Tent: fully test-time adaptation by entropy minimization. In: ICLR (2021)