PERSONALIZED SEMANTICS EXCITATION FOR FEDERATED IMAGE CLASSIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Federated learning casts a light on the collaboration of distributed local clients with privacy protected to attain a more generic global model. However, significant distribution shift in input/label space across different clients makes it challenging to well generalize to all clients, which motivates personalized federated learning (PFL). Existing PFL methods typically customize the local model by fine-tuning with limited local supervision and the global model regularizer, which secures local specificity but risks ruining the global discriminative knowledge. In this paper, we propose a novel Personalized Semantics Excitation (**PSE**) mechanism to breakthrough this limitation by exciting and fusing *personalized* semantics from the global model during local model customization. Specifically, PSE explores channel-wise gradient differentiation across global and local models to identify important low-level semantics mostly from convolutional layers which are embedded into the client-specific training. In addition, PSE deploys the collaboration of global and local models to enrich high-level feature representations and facilitate the robustness of client classifier through a cross-model attention module. Extensive experiments and analysis on various image classification benchmarks demonstrate the effectiveness and advantage of our method over the state-of-the-art PFL methods.

1 INTRODUCTION

Deep learning algorithms typically demand prolific training samples for model optimization (LeCun et al., 2015; He et al., 2016; Rao et al., 2021), which often entails crowd-sourcing from different clients. However, data privacy issue arises when transmitting data across clients (Yang et al., 2019; Wei & Liu, 2021; Ghazi et al., 2021). This motivates the exploration of federated learning (FL), which aims to learn a highly generalizable global model from the collaboration across multiple clients communicating with a centralized server to perform knowledge sharing (Li & Zhan, 2021; Huang et al., 2021a). While tremendous efforts have been made by existing approaches to produce a global model of strong generalizability to all clients, the crowd-sourcing nature of FL makes it really difficult to generate a generic model satisfying the demand of all clients with various data distributions (Chen & Chao, 2021; Sun et al., 2021).

The straightforward and efficient solution to this FL challenge is directly fine-tuning the well-learned global model to adapt the distribution property of each client (Mansour et al., 2020; Hu et al., 2020; Zhu et al., 2021). This widely-explored strategy is named as personalized federated learning (PFL), which conducts model customization per client by refining the local model with both local data and global model constraint (Achituve et al., 2021; Wu et al., 2022; Chen et al., 2022). Alternatively, **personalization** relies on the limited supervision per client refine the model to preserve client-specific patterns with the integration of global model. To reach better customization, (Hanzely & Richtárik, 2020; Deng et al., 2020) adopt additive mixture manner over the global and local network parameters to gradually adjust the local model learning. Similarly, one recent work named as Ditto (Li et al., 2021) enforces the local model parameters to be close to the global ones with ℓ_2-norm regularization term, which encourages clients to obtain generic knowledge and guarantees the convergence of training process. In addition, meta-learning mechanism has attracted much attention to overcome PFL challenges, since it enables the learning process of clients to imitate the attribution of knowledgeable in global model (Liang et al., 2020; Collins et al., 2021). Differently, FedRep (Collins et al., 2021) disentangles the top-down network architecture into a generic feature extractor and a private classifier.

Such a design manner not only preserves abundant high-level discriminative semantics related to data distribution but also gains benefits from cross-client collaboration via information integration in low-level convolutional layers. These mentioned works suggest that the communal and private semantic excitation and fusion is the key to achieve successful personalized client models.

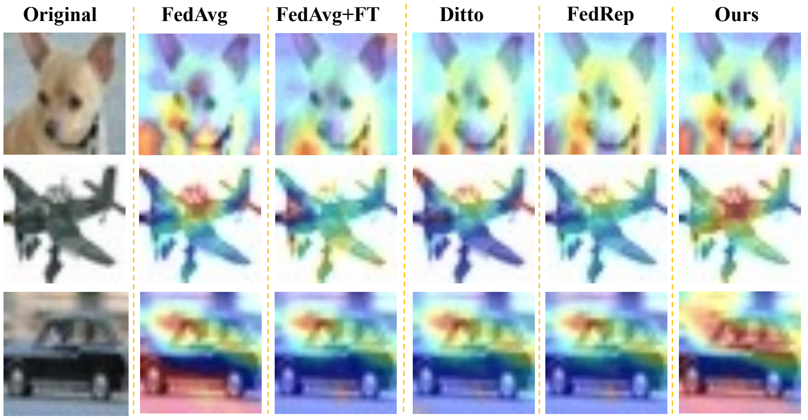


Figure 1: Comparison of attention maps drawn by global model (FedAvg), local models respectively learned by FedAvg+FT (McMahan et al., 2017), Ditto (Li et al., 2021), FedRep (Collins et al., 2021) and our proposed method.

Naturally, we post a question "how to precisely achieve personalization without hurting universality during model customization", which is promising yet underexplored. Namely, this learning process needs to determine which universal semantics are essential to improve local model performance and which are unnecessary to be overridden with local specific semantics. To explicitly answer this question, this work focuses on the **federated image classification**. First, we employ Grad-CAM (Selvaraju et al., 2017; Chen et al., 2020) to identify which patterns captured by convolutional neural network (CNN) are activated. Figure 1 shows the activated maps of image samples obtained by the global model learned by FedAvg and local models obtained from several PFL methods separately. From the comparison among FedAvg and the existing three PFL works in Figure 1, we easily observe that the global model learned by FedAvg pays more attention to the nose and mouth of the dog and utilizes these discriminative information to successfully identify the dog. However, FedAvg+FT, Ditto and FedRep show different degrees of degradation on these parts and fail to correctly classify it. This phenomenon is pretty common and illustrates that local model training introduces client-specific semantics but easily conceals or updates certain discriminative global information, which deviates from the eventual goal of PFL.

To prevent such phenomenons, this paper proposes a novel **Personalized Semantics Excitation** (**PSE**) mechanism to strike a balance between personalization and universality during local model customization. Our method mainly involves two modules: adaptively personalized channel excitation module and personalized semantic enhancement module. The first module considers precisely adjusting the filter parameters of convolution w.r.t local feature extractor by discovering which channel the global model provides more discriminative information. The delicate cross-model channel excitation to the utmost extent preserves the useful global knowledge. On the other hand, the second module aims to enrich high-level features and enhance the robustness of classifier. To attain this expectation, our method introduces the cross-model attention exchange mechanism over the last convolutional layer of feature extractor, which relies on channel-wise similarity to further elevate representation of discriminative semantics. The main contributions of our work include three folds:

- First, we empirically validate that local model training for the existing PFL is likely to override essential global semantics with weak discriminative client-specific contents. To avoid such a pitfall, we develop the adaptive channel excitation module to balance personalization and universality for each local client customization.
- Second, we develop the personalized semantic enhancement module with cross-model attention exchange mechanism to reach better personalization, which explores channel-wise similarity across global and local models to produce more robust high-level semantics representation for the classifier training.
- Finally, we evaluate our method and other baselines on novel scenario with data distribution divergence as well as conventional PFL setting with label shift. Extensive experimental results and analysis comprehensively illustrate the effectiveness of our method on achieving better model customization for federated image classification.

2 The Proposed Method

2.1 Preliminary and Motivation

Federated learning (FL) typically utilizes the communication between one centralized server and many distributed clients to construct a shared model with high generalization (Hu et al., 2022). This problem setting assumes each client locally stores their own private data $\mathcal{D}_i = \{\mathbf{x}^i_j, \mathrm{y}^i_j\}_{j=1}^{n^i}$ collected from the distribution $\mathcal{P}_i(\mathbf{x}^i, \mathrm{y}^i)$, where $\mathbf{x}^i_j$ and y^i_j denote visual input and its corresponding label, respectively. Assume there are m clients, and the i-th client contains n^i samples. With the collaborative protocol, all clients usually adopt the identical network architecture $\mathcal{F}(\cdot)$ with local trainable parameters Θ_i formulated by $\mathcal{F}(\mathbf{x}^i, \Theta_i)$. The most popular strategy is FedAvg (McMahan et al., 2017), which aims to achieve model sharing across different clients with data privacy protection. Formally, FebAvg naively averages the local model parameters Θ_i to reach the integrated model Θ with its objective functions at local clients and global server as:

$$\textbf{Local}: \min_{\Theta_i} \sum\nolimits_{j=1}^{n^i} \mathcal{L}(\mathcal{F}(\mathbf{x}^i_j, \Theta_i), \mathrm{y}^i_j) \quad \Leftrightarrow \quad \textbf{Global}: \Theta = \sum\nolimits_{i=1}^{m} \frac{n^i}{\sum_{i=1}^{m} n^i} \Theta_i, \tag{1}$$

where $\mathcal{L}(\cdot)$ is usually defined as the cross-entropy loss for classification task.

However, it is difficult to make the generic global Θ suitable for all clients due to the considerable distribution discrepancy in input or label space. The dilemma motivates the exploration of personalized federated learning (PFL) (Horvath et al., 2021; Zhang et al., 2021), which attempts to customize the global model for each concrete client. Along with this direction, the most recent FedRep (Collins et al., 2021) claims that clients can privatize the classifier and enable it to be more discriminative for their local data property with $\Theta_i = \{\Theta^i_g, \Theta^i_c\}$, where Θ^i_g are the shared parameters of generic feature generator from global server while Θ^i_c are the private classifier parameters locally preserved for i-th client. Another widely-used framework named as Ditto (Li et al., 2021) learns personalized client models with the following learning objective as:

$$\min_{\Theta^i_g, \Theta^i_c} \sum\nolimits_{j=1}^{n^i} \mathcal{L}(\mathcal{F}(\mathbf{x}^i_j, \Theta^i_g, \Theta^i_c), \mathrm{y}^i_j) + \lambda\Big(\|\Theta^i_g - \Theta_g\|^2_{\ell_2} + \|\Theta^i_c - \Theta_c\|^2_{\ell_2}\Big), \tag{2}$$

where $\|\cdot\|_{\ell_2}$ denotes ℓ_2-norm and λ is the trade-off parameter to the second and third terms, which enforce the global model regularizer to conduct client-specific customization.

Although Ditto and FedRep both achieve promising performance under PFL scenario, their local learning strategy that simply updates all global network parameter with private samples hardly counterpoises personalization and universality to achieve optimal performance (See Figure 1). **First**, the low-level convolutional filters learned from color and sketch images are likely to be diverse. Thus, the second term of Eq. 2 enforcing Θ^i_g to be the averaged Θ_g with no difference fails to adapt cross-modality clients. **Second**, the local classifier Θ^i_c also needs more augmented knowledge to promote its robustness and discriminative ability, especially when the private clients are with insufficient training samples, since other clients with the similar distribution provide effective assistant (Huang et al., 2021b; Sattler et al., 2020).

To reach the better model customization, the ideal solution not only needs to actively identify the important and discriminative global semantics by maximizing their contribution for local model training, but also discovers client-specific semantics to generate discriminative representations. Consequently, we develop 1) one adaptive personalized excitation mechanism within feature extractor Θ^i_g and 2) one personalization enhancement module with cross-model attention in private classifier module Θ^i_c. It is worth noting that such shared feature extractor parameters from all clients will be sent to the server, which would further conduct model integration as FedAvg in this paper for simplicity.

2.2 Adaptively Personalized Channel Excitation

Recent works on explainable deep learning (Chen et al., 2019; You et al., 2021; Nauta et al., 2021; Wickramanayake et al., 2021) suggest that different convolutional filters lying in the same layer focus on various regions of the input feature map and propagate their captured semantics into the next layer

(Sun et al., 2013; Dong et al., 2014; Farha & Gall, 2019), i.e., $\mathbf{F}^{(l)} = \mathbf{W}_l \otimes \mathbf{F}^{(l-1)}$, where $\otimes$ denotes the convolutional operation and $\mathbf{F}^{(l)}$ represents the 3D feature map of the l-layer with the total channel number as c_l. Based on this property, Grad-CAM (Selvaraju et al., 2017) attempts to learn the attention map from the output of the last convolutional (L-th) layer via $\mathbf{A} = \mathbf{ReLU}(\sum_k \alpha_k \mathbf{F}_k^{(L)})$, where $\mathbf{F}_k^{(L)} \in \mathbb{R}^{w \times h}$ represents the k-th channel with w, h as the index of the width and height, and $\alpha_k = \sum_w \sum_h \frac{\partial \mathrm{y}}{\partial \mathbf{F}_{k,w,h}^{(L)}}$. The combination weight α_k indicates the "importance" of the k-th feature map to the final prediction. Matching the attention matrix $\mathbf{A}$ over the original input image easily explains which regions lead the model to make the final decision.

Beneficial from Grad-CAM, we easily observe how does the PFL strategy achieve model customization. For example, we draw the attention map captured from FedRep (Collins et al., 2021) and FedAvg+FT (McMahan et al., 2017) adding the simple fine-tuning on well-learned global model. As Figure 1 shows, we can attain two important observations via the corresponding comparison. **First**, the local personalized learning in FedRep is likely to conceal or update certain important task-relevant information which are helpful for object classification yet activated by the global model from FedAvg. The reason we speculate lies in that the insufficient local training samples difficultly guide model to capture these patterns while the global model can integrate models across all clients to enrich them. That is also why Ditto (Li et al., 2021) in Eq. (2) attempts to reduce the distance between each client model parameters and global ones during the local training stage. **Second**, the client model actually can intensify the representations of certain regions around objects of our interest when compared with the attention map achieved by global model. The phenomenon results from that server is averaging the contributions of each client to realize global optimal solution. Therefore, imitating all patterns from server side as Ditto (Li et al., 2021) is also unsuitable for reaching personalized federated learning. With these findings, the ideal model customization not only preserves the local learned discriminative information but also borrows task-relevant semantics from global model.

To approximate the vision, the intuitive manner is to discover which convolutional filters of the global model can be activated to emphasize our interested pattern and embed them into the local learning process. Thus, we develop the adaptive channel excitation mechanism in client side with $\widetilde{\Theta} = \{\widetilde{\Theta}_g, \widetilde{\Theta}_c\}$ [1]. To this end, given arbitrary one training sample at any local client, we can feed it into local and global models to get the corresponding predictions via $\widetilde{\mathbf{p}}_j = \mathcal{F}(\mathbf{x}_j; \widetilde{\Theta}_g, \widetilde{\Theta}_c)$ and $\mathbf{p}_j = \mathcal{F}(\mathbf{x}_j; \Theta_g, \widetilde{\Theta}_c)$. On the other hand, we consider global model and local model would have different channel activation score given the same input sample. With the ground-truth label of the local training samples, we are able to deploy Grad-CAM (Selvaraju et al., 2017; Chattopadhay et al., 2018) to estimate the contribution of each feature map $\widetilde{\mathbf{F}}_k^{(l)}/\mathbf{F}_k^{(l)}$ at l-th layer k-th channel to the correct prediction per mini-batch as:

$$\widetilde{\alpha}_k^{(l)} = \sum_{j=1}^{b_s} \sum_w \sum_h \frac{\partial \widetilde{\mathbf{p}}_j^c}{\partial \widetilde{\mathbf{F}}_{k,(wh),j}^{(l)}}, \qquad \alpha_k^{(l)} = \sum_{j=1}^{b_s} \sum_w \sum_h \frac{\partial \mathbf{p}_j^c}{\partial \mathbf{F}_{k,(wh),j}^{(l)}}, \tag{3}$$

where $\widetilde{\mathbf{p}}_j^c$ / $\mathbf{p}_j^c$ denotes the predictive output of the j-th training sample on the c-th category (ground-truth), and b_s is the batch size. Intuitively, we can compare $\widetilde{\alpha}_k^{(l)}$ with $\alpha_k^{(l)}$ to identify the k-th channel's importance locally and globally at layer l. Since we hope the highly excited channels only resided in global model to compensate the local one, thus, we calculate $\Delta_k^{(l)} = \alpha_k^{(l)} - \widetilde{\alpha}_k^{(l)}$ with only positive difference. In the practical implementation, we first adopt **Sigmoid**$(\cdot)$ function to separately normalize the contribution coefficients of the same layer over client and global models. Thus, the personalized channel excitation is formulated as:

$$\widetilde{\mathbf{W}}_k^{(l)} \Leftarrow \widetilde{\mathbf{W}}_k^{(l)} + \Big\{ \mathbb{I}(\Delta^{k_l} \geq \bar{\Delta}) \cdot \xi \cdot \big(\mathbf{W}_k^{(l)} - \widetilde{\mathbf{W}}_k^{(l)}\big) \Big\}, \tag{4}$$

where $\bar{\Delta} = \mathbf{mean}(\sum_{k_l} \Delta_k^{(l)})$, and $\mathbb{I}(\cdot)$ is the indicator function. $\xi > 0$ controls the ratio of accepting external novel knowledge with its value as 0.01 by default.

2.3 Personalized Semantic Enhancement via Cross-model Attention

The adaptive channel excitation mechanism effectively fuses discriminative semantics from local and global sides to promote the generalization of feature. To examine the activation difference more

[1] Note that we remove the client index i of Θ_i with $\widetilde{\Theta}$ for easy observation.

correctly through Grad-CAM, we certainly expect the private classifier module to be more robust and discriminative in terms of generic feature representation. To achieve this, we are not fully relying on the generic feature extractor, but also measure the high-level semantic feature representation from both local and global models. The intuition is the global model is generic for all different tasks across various clients, which contributes feature robustness. With this thought, we propose the cross-model attention exchange module which adopts and advances the traditional self-attention components.

Given the rephrased 3-D feature map $\mathbf{F} \in \mathbb{R}^{h\times w\times c}$ from the last layer of Θ_g, we aim to automatically discover the channel-wise similarity in last convolutional layer to capture addition discriminative knowledge with the cross-model attention exchange. On the other hand, each client in federated learning typically consists of insufficient training samples for model optimization. Under this condition, the abundant linear projections to the keys $\mathbf{K} \in \mathbb{R}^{d_{hw}\times c}$, queries $\mathbf{Q} \in \mathbb{R}^{d_{hw}\times c}$ and values $\mathbf{V} \in \mathbb{R}^{d_{hw}\times c}$ (where $d_{hw} = h \times w$) in conventional self-attention module (Zhao et al., 2020; Dosovitskiy et al., 2020; Han et al., 2021; Liu et al., 2021) easily result in significant overfitting issue. To avoid it, we adopt lightweight convolutional kernel over feature maps to obtain the projections:

$$\mathbf{Q} = \mathbf{W}_q \otimes \mathbf{F}, \quad \mathbf{K} = \mathbf{W}_k \otimes \mathbf{F}, \quad \mathbf{V} = \mathbf{W}_v \otimes \mathbf{F}, \tag{5}$$

where $\mathbf{W}_{q/k/v} \in \mathbb{R}^{1\times 1}$ are 1-D convolutional filter with convolutional operator $\otimes$. Thus, we follow the tensor multiplication of (Wu et al., 2021) to obtain the output as the weight sum of the values:

$$\mathbf{O} = \textbf{Softmax}(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{wh}})\mathbf{V}, \tag{6}$$

where the self-attention weights $\textbf{Softmax}(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{wh}}) \in \mathbb{R}^{d_{hw}\times d_{hw}}$ highlight the important semantics within per channel and $\mathbf{O} \in \mathbb{R}^{d_{hw}\times c}$ will be reshaped into the same size as $\mathbf{F}$. Similarly, these convolutional kernel will be deployed over the local feature maps $\widetilde{\mathbf{F}}$ from $\widetilde{\Theta}_g$ to obtain the corresponding outputs $\{\widetilde{\mathbf{Q}}, \widetilde{\mathbf{K}}, \widetilde{\mathbf{V}}\} \Rightarrow \widetilde{\mathbf{O}}$.

To obtain more discriminative knowledge from global to local client, we consider to exchange $\widetilde{\mathbf{Q}}$ with $\mathbf{Q}$ to deepen their consensus on the high-level channel-wise features. Similarly, we can replace $\mathbf{Q}$ with $\widetilde{\mathbf{Q}}$ in Eq. (6) to increase input diversity which further improve the robustness of classifier. In addition, the relative positions of channels in the same network layer is important information. With this consideration, beyond the exchange queries, we further introduce a learnable position variable parameterized as $\mathbf{P} \in \mathbb{R}^{w\times c\times h}$ into cross-model attention exchange module:

$$\widetilde{\mathbf{O}} = \textbf{Softmax}(\frac{\mathbf{Q}\widetilde{\mathbf{K}}^\top + \mathbf{Q}\mathbf{P}^\top}{\sqrt{wh}})\widetilde{\mathbf{V}}, \quad \mathbf{O} = \textbf{Softmax}(\frac{\widetilde{\mathbf{Q}}\mathbf{K}^\top + \widetilde{\mathbf{Q}}\mathbf{P}^\top}{\sqrt{wh}})\mathbf{V}. \tag{7}$$

Finally, we utilize max-pooling on the outputs $\widetilde{\mathbf{O}}$ and $\mathbf{O}$ of cross-model attention module, then flatten and feed them into two fully-connected layer θ_{fc} to access their logits, i.e., $\widetilde{\mathbf{p}} = \theta_{\text{fc}}(\widetilde{\mathbf{O}})$ and $\mathbf{p} = \theta_{\text{fc}}(\mathbf{O})$. Note that the classifier parameter $\widetilde{\Theta}_c$ is specified as $\{\mathbf{W}_{q/k/v}, \mathbf{P}, \theta_{\text{fc}}\}$. The attention exchange enables the local model to pay attention to these informative and discriminative channels as the global network does and utilizes its values ($\mathbf{V}$) to heavily preserve local well-learned knowledge. Meanwhile, we also conduct the similar operation over the feature maps of global model and feed the outputs $\widetilde{\mathbf{O}}$ and $\mathbf{O}$ into the last fully-connected layer of classifier, which further promote the generalization of classifier.

2.4 Overall Objective and Discussion

Therefore, we can deduce the objective function of our local model learning as the following via the integrating Eq. (4) as a regularizer with respect to $\widetilde{\mathbf{W}}_k^{(l)}$ as well as enhanced classification loss:

$$\min_{\widetilde{\Theta}_g,\widetilde{\Theta}_c} \widetilde{\mathcal{L}} = \underbrace{\sum\nolimits_j \mathcal{L}(\widetilde{\mathbf{p}}_j, \mathrm{y}_j) + \mathcal{L}(\mathbf{p}_j, \mathrm{y}_j)}_{\textbf{Obj } 1} + \underbrace{\sum\nolimits_l \sum\nolimits_k \frac{\xi}{2} \cdot \mathbb{I}(\Delta^{k_l} \geq \bar{\Delta}) \cdot \|\widetilde{\mathbf{W}}_k^{(l)} - \mathbf{W}_k^{(l)}\|_{\ell_2}^2}_{\textbf{Obj } 2}. \tag{8}$$

The global model is frozen during the overall local training process. Note that in the inference stage, the client model only depends on the local network $\widetilde{\Theta}_g, \widetilde{\Theta}_c$ to achieve classification task without the assistance of global model, which means the cross-model attention exchange is degenerated into the self-attention mode.

Remark: Two strategies mutually work together to enhance the personalization from shared channel-wise semantic (Obj 2 in Eq. (8)) and private semantic information (Obj 1 in Eq. (8)). Actually, our model is very relevant to Ditto (Li et al., 2021) but with two most significant improvements. First, we aim to achieve optimal balance of personalization and universality to improve local model performance via channel-wise excitation instead of simply regularizing all local parameters indifferently. Second, we explore cross-model high-level semantic correlation to trigger the private classifier more robust and discriminative. In addition, we provide the explicit theorem for model convergence and convergence rate as follows.

Assumption 1. The stochastic gradient $\mathbf{g}_t = \nabla\widetilde{\mathcal{L}}(\widetilde{\Theta}_t, \mathbf{x}_t)$ at time t is an unbiased estimator of the local gradient with the expectation as $\mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[\mathbf{g}_t] = \nabla\widetilde{\mathcal{L}}_t$ and variance as $\mathbb{E}[\|\mathbf{g}_t - \nabla\widetilde{\mathcal{L}}_t\|_2^2] \leq \delta^2$.

Assumption 2. The objective function optimized in each client is L_1-Lipschitz smooth. In other words, the gradient of Eq. (8) is L_1-Lipschitz continuous (Malherbe & Vayatis, 2017), i.e., $\|\nabla\widetilde{\mathcal{L}}_{t_1} - \nabla\widetilde{\mathcal{L}}_{t_2}\|_2 \leq L_1\|\widetilde{\Theta}_{t_1} - \widetilde{\Theta}_{t_2}\|_2$, where $\mathcal{L}_{t_{1/2}}$ means the loss values at local iteration time $t_{1/2}$.

Theorem 1. When assumption 1 and 2 hold, we have the following conclusion in any arbitrary client after per communication round (r):

$$\mathbb{E}[\widetilde{\mathcal{L}}_{(r+1)\tau}] \leq \widetilde{\mathcal{L}}_{r\tau+1} - (\eta - \frac{L_1\eta^2}{2})\sum_{e=1}^{\tau-1}\|\nabla\widetilde{\mathcal{L}}_{r\tau+e}\|_2^2 + \frac{L_1\tau\eta^2}{2}\delta^2, \tag{9}$$

where τ is the total iteration of local model update and η is the learning rate. This theorem suggests that selecting appropriate η can achieve our expected gradient decrease in one communication round so that it finally can guarantee the convergence of model.

Theorem 2. Given any ϵ, after R round communication, we infer that

$$\frac{1}{R\tau}\sum_{r=1}^{R-1}\sum_{e=1}^{\tau-1}\mathbb{E}[\|\nabla\widetilde{\mathcal{L}}_{r\tau+e}\|_2^2] \leq \epsilon, \quad R \geq \frac{2(\widetilde{\mathcal{L}}_0 - \widetilde{\mathcal{L}}_*)}{\tau\epsilon(2\eta - L_1\eta^2) - \tau\eta^2 L_1\delta^2}, \tag{10}$$

where $\eta < \frac{2\epsilon}{L_1(\epsilon+\delta^2)}$ and $\widetilde{\mathcal{L}}_*$ denotes the loss of the optimal solution for the local model. This theorem illustrates the convergence rate of model, which is related to the overall communication round and the expectation of ℓ_2-norm of gradient. Sufficient communication rounds make the bound tighter. Please refer to the supplementary material for the proofs of two theorems.

3 Experiments

3.1 Experimental Setup

Datasets. In practical experiments, we not only consider label distribution shift across various clients as the traditional PFL works (Tan et al., 2022; Fallah et al., 2020) but also attempt to explore the interference of cross-client data distribution mismatch. In terms of the label shift, we follow the protocol of FedRep (Collins et al., 2021) to randomly divide 50,000 training images of Cifar-10 and Cifar-100 (Krizhevsky et al., 2009) into 20/50 clients and each client contains the same category number. And the 10,000 test samples are also split into each client according to their categories. Similarly, the original FashionMNIST consists of 60,000 training gray images and 10,000 test ones, which are also randomly distributed into 100 client terminals, whose category number varies from two to four per client. For the data shift experiments, we first convert gray images of FashionMNIST and FeMNIST into colorful or edge images. Specifically, we arbitrarily crop the 28×28 patch from color images of BSD500 (Martin et al., 2001) and add them into the gray images to generate colorful digit or fashion images. Moreover, the edge fashion images are synthesized using classical canny detector over the gray images. Given another modality dataset, we adopt the same manner to split the newly-created samples into the additional 100 clients. To this end, we can evaluate PFL algorithms over 200 clients with significant data/label distribution divergence.

Implementation Details. For our proposed method, the network architecture for all experiments includes one feature extractor and one classifier. Concretely, the feature extractor involves three convolutional layers with the specific channel numbers (1/3→32→64→128)[2]. The classifier consists

[2] Note that if there exist color images for training, the channel number of input will be three and that of gray or edge images is also converted into three-channel input.

Table 1: Average Recognition Accuracy (%) under novel joint label and data shift scenarios.

Datasets	FEMNIST			FashionMNIST			FashionMNIST		
Modality	(Gray, Color)			(Gray, Color)			(Color, Edge)		
(#M, #C)	(200,3)	(200,4)	(200,5)	(100,3)	(100,4)	(100,5)	(200,3)	(200,4)	(200,5)
Local	81.97	80.40	79.44	81.46	79.62	76.95	83.53	81.95	80.57
FedAvg+FT	83.17	81.97	81.53	84.28	82.36	79.80	86.39	84.08	82.99
FedProx+FT	82.87	81.36	81.03	84.51	82.31	79.50	87.05	84.25	83.09
SCAFFOLD+FT	84.00	81.54	82.04	84.79	82.12	79.94	85.54	83.19	83.05
Fed-MTL	81.14	80.30	79.24	78.70	77.14	78.39	81.12	79.49	79.89
LG-Fed	83.27	81.40	80.03	81.59	79.23	75.89	83.86	80.90	78.31
L2GD	81.88	80.53	79.68	80.16	78.90	77.46	81.75	80.86	79.52
APFL	82.85	81.17	81.14	85.25	81.16	78.73	85.96	82.46	79.22
Ditto	85.23	82.94	82.34	88.11	85.76	84.46	87.82	84.77	84.13
FedRep	84.43	83.54	83.51	86.71	83.01	83.49	84.78	85.10	84.46
Ours	**88.81**	**87.86**	**87.98**	**89.58**	**88.12**	**86.61**	**89.97**	**87.95**	**85.69**

of one multi-head (4-heads) cross-model attention block and two fully-connected layers. The local model training within each client adopts stochastic gradient descent (SGD) to optimize the model with momentum 0.5 and the learning rate as 0.01. Moreover, in each round of communication, 50% clients of Cifar-10/100 or 10% ones of other experiments are randomly selected to update their local model for 5 epochs and send their feature extractors to the global server for model integration. The server will conduct 100 rounds of communication with local clients.

Baselines. To evaluate the effectiveness of our method, we compare with the state-of-the-art PFL algorithms. Generally, they are divided into two branches. One manner is utilizing conventional federated learning methods such as FedAvg (McMahan et al., 2017), FedProx (Li et al., 2018) and SCAFFOLD (Karimireddy et al., 2019) to attain their global models and then fine-tuning (FT) them to customize the local network named as "**X**"+FT. The other direction is to design the specific customized model training approaches as Fed-MTL (Smith et al., 2017), LG-Fed (Liang et al., 2020), L2GD (Hanzely & Richtárik, 2020), APFL (Deng et al., 2020), Ditto (Li et al., 2021), and FedRep (Collins et al., 2021). For a fair comparison, we perform experiments with their public available implementations and replace the network architecture with the above mentioned design, e.g., three-layer CNNs, one self-attention module (Eq. (6)) and two FC layers, where only our cross-model attention mechanism is not deployed.

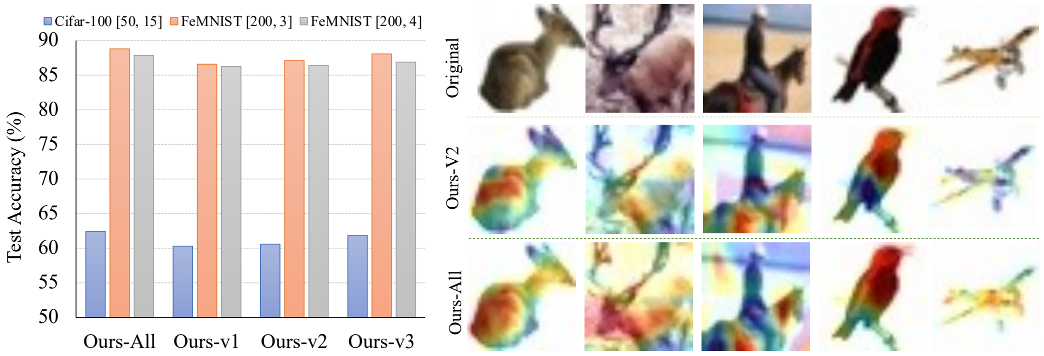
3.2 Comparison Results

In PFL experiments, all training and test samples are randomly allocated into multiple clients. To reduce the uncertain influence of random partition, we carry out many times for each task and report the average accuracy. It is worth nothing that each client will evaluate local model with its private test samples and access the corresponding accuracy. The above test accuracy refers to average the test classification accuracy across all clients. Table 1 and Table 2 show the performances of our method and other baselines over various datasets under different partitions. According to them, we can easily achieve several valuable conclusions.

First, it is straightforward to observe that our method obtains the state-of-the-art performance in all mentioned tasks. This convincingly illustrates the effectiveness of our method on customizing client model under federated learning scenario. In terms of the experiments on Cifar-100 with 50 clients, there exists considerable label space divergence across different clients. In other words, arbitrary clients have a little category information overlap. Under the difficult situation, our method outperforms others by a large margin, especially for the case (50, 15), (Ours *v.s.* FedRep)$\sim$(62.46% *v.s.* 58.94%). These comparisons suggest our proposed method significantly overcomes the negative effect of label distribution shift when conducting knowledge sharing. **Second**, compared with several personalized training manners as Fed-MTL, LG-Fed, the naive fine-tuning mechanism over the global model well-learned from FedAvg or FedProx produces promising results in many tasks. And Ditto heavily depends on the global models and attains stable performances in these experiments. It demonstrates that the local personalized learning is likely to conceal or update useful knowledge from

Table 2: Average Recognition Accuracy (%) under conventional label shift scenarios.

Datasets	CIFAR-10			CIFAR-100			FashionMNIST		
(#M, #C)	(20,2)	(20,3)	(20,4)	(50,5)	(50,10)	(50,15)	(100,3)	(100,4)	(100,5)
Local	79.65	73.97	67.54	73.35	58.76	49.79	89.65	86.37	85.75
FedAvg+FT	82.94	78.23	74.62	77.01	61.96	55.40	91.43	89.00	87.36
FedProx+FT	82.44	76.74	73.63	74.10	60.40	53.35	88.35	87.05	85.51
SCAFFOLD+FT	82.03	76.51	72.92	75.09	59.92	51.54	90.33	87.68	85.22
Fed-MTL	83.19	75.81	69.57	65.28	54.84	48.72	84.65	82.59	82.86
LG-Fed	84.24	77.1	71.23	67.17	54.31	50.63	87.07	84.51	81.19
L2GD	83.76	76.26	69.8	67.15	55.30	50.12	85.50	83.88	82.84
APFL	82.09	78.80	74.29	72.81	61.77	54.04	90.58	86.83	85.67
Ditto	84.74	80.34	76.25	75.23	65.40	56.14	91.21	89.91	88.81
FedRep	84.12	80.39	76.28	78.30	63.52	58.94	92.71	90.73	89.56
Ours	**86.95**	**82.98**	**78.03**	**79.58**	**67.10**	**62.46**	**94.03**	**91.77**	**90.47**



(a) (b)

Figure 2: (a) Comparison of multiple variants over three tasks, (b) Comparison of attention map drawn by our proposed method and one variant.

the external collaborators. Therefore, our proposed learning mechanism not only discovers valuable global information but also gradually adjusts local model to augment private data distribution. **Third**, through the comparison of Table 1 and Table 2 with respect to FashionMNIST, we notice that all involved methods suffer from performance degradation when introducing distribution shift on inputs across various clients. This domain shift scenario brings in more challenges to personalized federated learning. However, our method still significantly outperforms other competitors. The main reason lies in the collaboration of adaptive channel excitation and cross-model attention mechanisms, which effectively captures more discriminative information to promote the robustness of model.

3.3 Empirical Analysis

Ablation Study. The cooperation of adaptive channel excitation and cross-model attention exchange assists our model in achieving better recognition performance. To clearly understand the contribution of each component, we design three variants for our method by separately removing one of the following components: a) the effect of channel excitation module (Ours-v1), 2) cross-model attention (Ours-v2) and 3) position information of Eq. (7) (Ours-v3). The results in Figure 2(a) show their difference under three scenarios. On one hand, by removing two important modules, Ours-v1 and Ours-v2 suffer from significant performance degradation, which inversely testifies the effectiveness of them on personalization. On the other hand, the position information also provides a little positive effect on performance improvement by intensifying valuable channel representation. In addition, we also explicitly analyze how does the cross-model attention help the model to promote its discriminative ability and robustness. We also visualize the heat map of Ours-v2 in Figure 2(b). From these visualizations, we achieve the conclusion that cross-model attention exchange explores the channel-wise similarity to find novel discriminative knowledge and instructs low-level convolutional operation to achieve them. For example, for the "elk" in the 2-nd column, Ours-v2 merely focuses

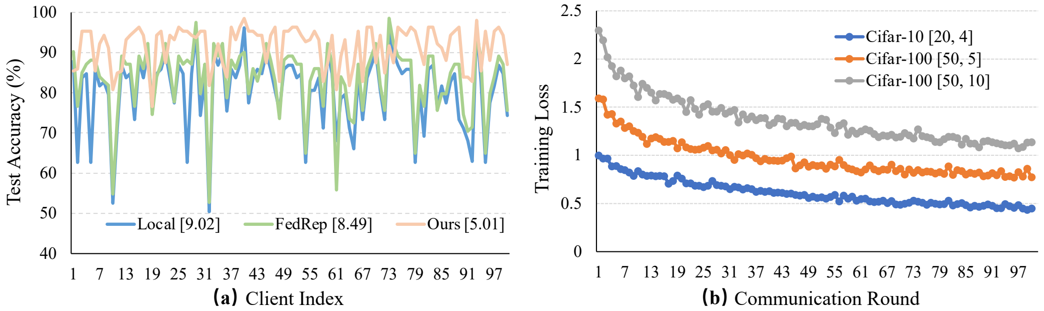


Figure 3: (a) Fairness analysis by calculating the performance per client and overall standard deviation attached behind the method, (b) Convergence analysis over different communication rounds.

neck of “elk” and provides a little discriminative information for the final decision. Differently, our integrated method can pursue more useful knowledge as head and antlers to object classification.

Fairness & Convergence. PFL setting not only customizes local models to attain performance improvement but also expects that all clients are able to benefit from the model sharing with fair performance improvement, which is also defined as “Fairness” (Li et al., 2021). Thus, we utilize the well-learned local model of each client to do evaluation on test samples from gray FashionMNIST (100, 4) and record them in Figure 3(a). Compared with FedAvg+FT and FedRep, our performance divergence across all clients is relatively slight. Specifically, the standard deviations of all client test accuracy for FedAvg+FT, FedRep and ours are 9.02%, 8.49% and 5.01%. Thus, our proposed method generates better fairness when solving PFL challenge. Moreover, PFL scenario generally concerns the training convergence. For this point, we record the training loss in each communication process on three cases and show them in Figure 3(b). By observing them, it is simple to know that the training process of our method is stable and easily achieves convergence which is consistent with the theorems.

Confusion Matrix. To clearly understand how our method benefits the various categories in each client, we randomly select one client from Cifar-100 (50, 15) by comparing our model and local training only. It is worth noting that there are only 15 categories per client. The confusion matrices for the local test samples are shown in Figure 4, where we highlight the significant improved categories in red, and slightly decreased categories in blue. From it, we find that our method significantly improves the ratio of correct classification in most categories, which illustrates our method captures more discriminative semantics when preserving certain valuable global information.

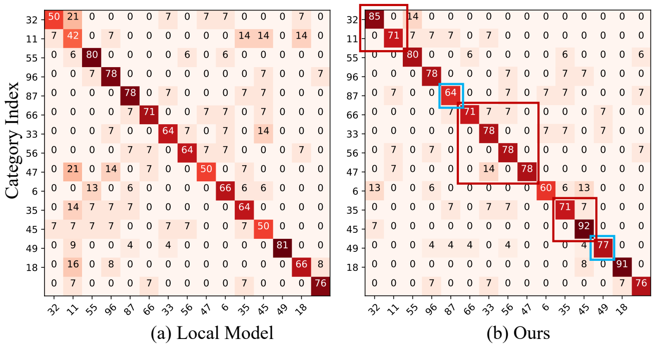


Figure 4: Confusion matrix of local training manner and our method in one client from Cifar-100 (50, 15).

4 Conclusion

Personalized federated learning not only utilizes the collaboration of numerous distributed clients to achieve knowledge sharing without private data leakage but also customizes local model to adapt the property of private data. Through the empirical studies on the existing PFL solutions, we observed that their local personalization easily conceals certain important patterns captured by global model, leading to incorrect classification. To solve this, we proposed a novel algorithm to attain better customization including two modules, i.e., adaptive personalized channel excitation and personalized semantic enhancement. The first component attempts to discover valuable knowledge from global model and precisely adjust the parameters of convolutional filters in local model to achieve semantics fusion. The second one explores the cross-model attention exchange mechanism to discover complementary and discriminative knowledge to enhance the robustness of local features. In practical implementation, we evaluate the performance of algorithm on conventional PFL setting with label shift and novel scenario with input distribution shift. The experimental comparisons with baselines and analysis verify the effectiveness of our method on solving PFL issue.

References

Idan Achituve, Aviv Shamsian, Aviv Navon, Gal Chechik, and Ethan Fetaya. Personalized federated learning with gaussian processes. *Advances in Neural Information Processing Systems*, 34, 2021.

Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pp. 839–847. IEEE, 2018.

Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. This looks like that: deep learning for interpretable image recognition. *Advances in neural information processing systems*, 32, 2019.

Hong-You Chen and Wei-Lun Chao. On bridging generic and personalized federated learning for image classification. In *International Conference on Learning Representations*, 2021.

Lei Chen, Jianhui Chen, Hossein Hajimirsadeghi, and Greg Mori. Adapting grad-cam for embedding networks. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2794–2803, 2020.

Zhen Chen, Chen Yang, Meilu Zhu, Zhe Peng, and Yixuan Yuan. Personalized retrogress-resilient federated learning towards imbalanced medical data. *IEEE Transactions on Medical Imaging*, 2022.

Liam Collins, Hamed Hassani, Aryan Mokhtari, and Sanjay Shakkottai. Exploiting shared representations for personalized federated learning. *arXiv preprint arXiv:2102.07078*, 2021.

Yuyang Deng, Mohammad Mahdi Kamani, and Mehrdad Mahdavi. Adaptive personalized federated learning. *arXiv preprint arXiv:2003.13461*, 2020.

Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *European conference on computer vision*, pp. 184–199. Springer, 2014.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach. *Advances in Neural Information Processing Systems*, 33:3557–3568, 2020.

Yazan Abu Farha and Jurgen Gall. Ms-tcn: Multi-stage temporal convolutional network for action segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3575–3584, 2019.

Badih Ghazi, Noah Golowich, Ravi Kumar, Pasin Manurangsi, and Chiyuan Zhang. Deep learning with label differential privacy. *Advances in Neural Information Processing Systems*, 34, 2021.

Kai Han, An Xiao, Enhua Wu, Jianyuan Guo, Chunjing Xu, and Yunhe Wang. Transformer in transformer. *Advances in Neural Information Processing Systems*, 34, 2021.

Filip Hanzely and Peter Richtárik. Federated learning of a mixture of global and local models. *arXiv preprint arXiv:2002.05516*, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Samuel Horvath, Stefanos Laskaridis, Mario Almeida, Ilias Leontiadis, Stylianos Venieris, and Nicholas Lane. Fjord: Fair and accurate federated learning under heterogeneous targets with ordered dropout. *Advances in Neural Information Processing Systems*, 34, 2021.

Rui Hu, Yuanxiong Guo, Hongning Li, Qingqi Pei, and Yanmin Gong. Personalized federated learning with differential privacy. *IEEE Internet of Things Journal*, 7(10):9530–9539, 2020.

Zeou Hu, Kiarash Shaloudegi, Guojun Zhang, and Yaoliang Yu. Federated learning meets multi-objective optimization. *IEEE Transactions on Network Science and Engineering*, 2022.

Baihe Huang, Xiaoxiao Li, Zhao Song, and Xin Yang. Fl-ntk: A neural tangent kernel-based framework for federated learning analysis. In *International Conference on Machine Learning*, pp. 4423–4434. PMLR, 2021a.

Yutao Huang, Lingyang Chu, Zirui Zhou, Lanjun Wang, Jiangchuan Liu, Jian Pei, and Yong Zhang. Personalized cross-silo federated learning on non-iid data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 7865–7873, 2021b.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank J Reddi, Sebastian U Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for on-device federated learning. 2019.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *arXiv preprint arXiv:1812.06127*, 2018.

Tian Li, Shengyuan Hu, Ahmad Beirami, and Virginia Smith. Ditto: Fair and robust federated learning through personalization. In *International Conference on Machine Learning*, pp. 6357–6368. PMLR, 2021.

Xin-Chun Li and De-Chuan Zhan. Fedrs: Federated learning with restricted softmax for label distribution non-iid data. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 995–1005, 2021.

Paul Pu Liang, Terrance Liu, Liu Ziyin, Nicholas B Allen, Randy P Auerbach, David Brent, Ruslan Salakhutdinov, and Louis-Philippe Morency. Think locally, act globally: Federated learning with local and global representations. *arXiv preprint arXiv:2001.01523*, 2020.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10012–10022, 2021.

Cédric Malherbe and Nicolas Vayatis. Global optimization of lipschitz functions. In *International Conference on Machine Learning*, pp. 2314–2323. PMLR, 2017.

Yishay Mansour, Mehryar Mohri, Jae Ro, and Ananda Theertha Suresh. Three approaches for personalization with applications to federated learning. *arXiv preprint arXiv:2002.10619*, 2020.

David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, volume 2, pp. 416–423. IEEE, 2001.

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*, pp. 1273–1282. PMLR, 2017.

Meike Nauta, Annemarie Jutte, Jesper Provoost, and Christin Seifert. This looks like that, because... explaining prototypes for interpretable image recognition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 441–456. Springer, 2021.

Yongming Rao, Wenliang Zhao, Zheng Zhu, Jiwen Lu, and Jie Zhou. Global filter networks for image classification. *Advances in Neural Information Processing Systems*, 34, 2021.

Felix Sattler, Klaus-Robert Müller, and Wojciech Samek. Clustered federated learning: Model-agnostic distributed multitask optimization under privacy constraints. *IEEE transactions on neural networks and learning systems*, 32(8):3710–3722, 2020.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

Virginia Smith, Chao-Kai Chiang, Maziar Sanjabi, and Ameet Talwalkar. Federated multi-task learning. *arXiv preprint arXiv:1705.10467*, 2017.

Benyuan Sun, Hongxing Huo, Yi Yang, and Bo Bai. Partialfed: Cross-domain personalized federated learning via partial initialization. *Advances in Neural Information Processing Systems*, 34, 2021.

Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deep convolutional network cascade for facial point detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3476–3483, 2013.

Alysa Ziying Tan, Han Yu, Lizhen Cui, and Qiang Yang. Towards personalized federated learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

Wenqi Wei and Ling Liu. Gradient leakage attack resilient deep learning. *IEEE Transactions on Information Forensics and Security*, 2021.

Sandareka Wickramanayake, Wynne Hsu, and Mong Li Lee. Explanation-based data augmentation for image classification. *Advances in Neural Information Processing Systems*, 34, 2021.

Chuhan Wu, Fangzhao Wu, Lingjuan Lyu, Yongfeng Huang, and Xing Xie. Communication-efficient federated learning via knowledge distillation. *Nature communications*, 13(1):1–8, 2022.

Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing convolutions to vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22–31, 2021.

Qiang Yang, Yang Liu, Yong Cheng, Yan Kang, Tianjian Chen, and Han Yu. Federated learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 13(3):1–207, 2019.

Zunzhi You, Yi-Hsuan Tsai, Wei-Chen Chiu, and Guanbin Li. Towards interpretable deep networks for monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12879–12888, 2021.

Jie Zhang, Song Guo, Xiaosong Ma, Haozhao Wang, Wenchao Xu, and Feijie Wu. Parameterized knowledge transfer for personalized federated learning. *Advances in Neural Information Processing Systems*, 34, 2021.

Hengshuang Zhao, Jiaya Jia, and Vladlen Koltun. Exploring self-attention for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10076–10085, 2020.

Linghui Zhu, Xinyi Liu, Yiming Li, Xue Yang, Shu-Tao Xia, and Rongxing Lu. A fine-grained differentially private federated learning against leakage from gradients. *IEEE Internet of Things Journal*, 2021.