Collaborative Semantic Aggregation and Calibration for Federated Domain Generalization

Junkun Yuan, *Student Member, IEEE,* Xu Ma, Defang Chen, Kun Kuang, Fei Wu, *Senior Member, IEEE,* and Lanfen Lin, *Member, IEEE*

Abstract—Domain generalization (DG) aims to learn from multiple known source domains a model that can generalize well to unknown target domains. The existing DG methods usually exploit the fusion of shared multi-source data to train a generalizable model. However, tremendous data is distributed across lots of places nowadays that can not be shared due to privacy policies. In this paper, we tackle the problem of federated domain generalization where the source datasets can only be accessed and learned locally for privacy protection. We propose a novel framework called Collaborative Semantic Aggregation and Calibration (CSAC) to enable this challenging problem. To fully absorb multi-source semantic information while avoiding unsafe data fusion, we conduct data-free semantic aggregation by fusing the models trained on the separated domains layer-by-layer. To address the semantic dislocation problem caused by domain shift, we further design cross-layer semantic calibration with an attention mechanism to align each semantic level and enhance domain invariance. We unify multi-source semantic learning and alignment in a collaborative way by repeating the semantic aggregation and calibration alternately, keeping each dataset localized, and the data privacy is carefully protected. Extensive experiments show the significant performance of our method in addressing this challenging problem.

Index Terms—Domain generalization, Federated learning, Semantic aggregation, Semantic calibration, Attention mechanism.

1 Introduction

Recently, deep learning has made revolutionary advances to visual recognition [22], under the i.i.d. assumption that training and test data is sampled from the same distribution. Since the adopted datasets could be very distinct in many real-world applications, the performance of deep models learned from one training (source) dataset may drop rapidly on another test (target) dataset. To address this *dataset/domain shift* [63] problem, *domain generalization* (DG) [4], [77], [100] is introduced to train a generalizable model to unknown target domains by learning from multiple semantically-relevant source domains.

Numerous DG methods [5], [12], [99] have been proposed recently. They popularize a variety of favorable strategies for training generalizable models by (indirectly) exploiting the fusion of *"shared"* multi-source data. For example, some alignment-based methods [33], [40], [99] match source data distributions in latent space for generating domain-invariant feature representations. Some meta-learning based strategies [12], [32], [42] utilize meta-train and meta-test datasets built by sampling from multi-source data for training a stable model to unknown domains. However, these methods may seriously violate data privacy policies, as tremendous data is stored locally in distributed places nowadays which may contain private information, e.g., the patient data from hospitals and the video recording from surveillance cameras. Therefore, a dilemma is encountered: The requirements of learning from shared multi-source data for training a highly generalizable model may hard to be met in many real scenarios due to the privacy issues. Meanwhile, without simultaneous access to the source datasets for obtaining adequate information of multi-source distribution, identifying and learning domain invariance for improving model generalization might be led astray.

In this paper, we tackle the problem of *federated domain generalization* [48] (see Fig. 1), where the source datasets are separated and can only be accessed and learned locally. It enables privacy preserving of sensitive data when employing them for improving model generalization. However, it is much more challenging than the conventional domain generalization task as: (1) The separated source datasets are private and may not be directly fused, hence the simultaneous learning of the multi-source semantic information is greatly hindered, making the identification of domain invariance tricky. (2) The heterogeneous source datasets with distinct data distributions may constitute enormous obstacles for training a generalizable model as the model is allowed to access only one local dataset each time, while the accessed dataset could contain particularly unusual bias and even bring negative gain for model generalization.

We propose a novel method called Collaborative Semantic Aggregation and Calibration (CSAC) to enable federated domain generalization. We begin by hypothesizing that the deep models extract semantic information layer-by-layer, and the model parameters in each layer are related to the corresponding level as well as the training data distribution (proof-of-concept experiments are provided to verify it in Sec. 4). In light of this, to fully absorb multi-source semantic information while avoiding risky data fusion, a data-free semantic aggregation strategy is devised to fuse the models trained on the separated domains layer-by-layer. Then a

- *J. Yuan, X. Ma, D. Chen, K. Kuang, F. Wu, L. Lin are with the College of Computer Science and Technology, Zhejiang University, Zhejiang, China. Correspondence to: Kun Kuang. Junkun Yuan and Xu Ma contributed equally to this work. E-mail: {yuanjk, maxu, defchern, kunkuang, llf}@zju.edu.cn and wufei@cs.zju.edu.cn.*

Manuscript received April 19, 2005; revised August 26, 2015.

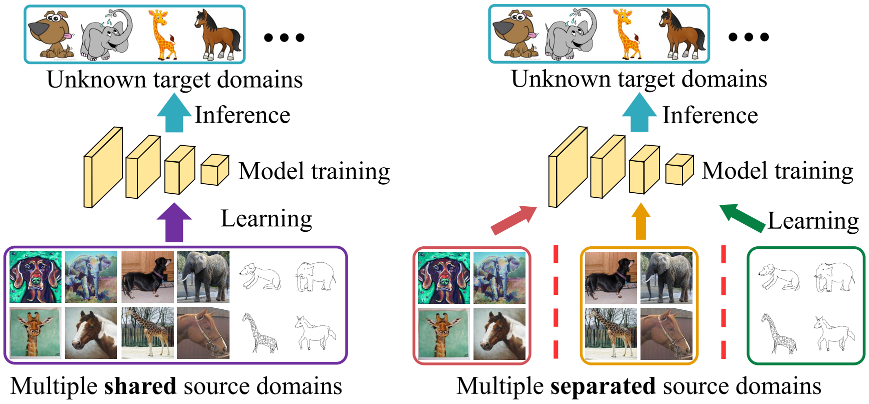

Fig. 1. Comparison of the conventional domain generalization (left) and the federated domain generalization (right). The latter trains generalizable models while protecting privacy for many real-world scenarios.

semantic dislocation problem has arisen: Due to domain shift, the same level of semantic information from different domains could distribute across model layers during the aggregation. To this end, we further design cross-layer semantic calibration with an elaborate attention mechanism for precise semantic level alignment and domain-invariance enhancement. We unify the multi-source semantic learning and alignment in a collaborative way by repeating the semantic aggregation and calibration alternately. Each source dataset contributes semantic information locally for boosting model generalization during this process, resulting in a high-quality generalizable model under privacy protection.

Our main contributions are summarized as: (1) We tackle a practical problem of federated domain generalization for addressing the dilemma between model generalization and privacy protection. This problem is important for training generalizable models in many privacy-sensitive scenarios but lacks extensive research to our knowledge. (2) To enable federated domain generalization, we propose a novel framework called Collaborative Semantic Aggregation and Calibration (CSAC) to unify the multi-source semantic learning and alignment in a collaborative way by repeating semantic aggregation and calibration alternately. (3) Extensive experiments on benchmark datasets show the significant performance of our method in addressing federated domain generalization, which is even comparable to the previous DG methods with shared source data.

The rest of the paper is organized as follows. In Sec. 2, some related works about domain adaptation, domain generalization, federated learning, and distributed domain adaptation and generalization are introduced. In Sec. 3, the problem definition of the federated domain generalization and our proposed CSAC framework and algorithm are stated. In Sec. 4, the results of the experiments on benchmark datasets as well as ablation studies and discussions are provided. We discuss the investigation of the federated domain generalization with a future outlook in Sec. 5.

2 Related Work

2.1 Domain Adaptation

To address the widespread domain shift problem, remarkable progress [3], [9], [27], [35], [36], [39], [47], [49], [50], [59], [61], [82], [86], [89] has been made in domain adaptation task. It aims to adapt a model trained on source domains to target domains by exploiting target data/information. One prevailing direction for this task is to employ adversarial learning [9], [39], [61] for reducing domain gap and generating domain-agnostic representations. Meanwhile, some algorithms [36], [49], [50] are put forward to directly minimize domain divergence with distance metric like Maximum Mean Discrepancies (MMD). However, the target data/information is assumed to be available in this task, which greatly limits its implementation in many real-world applications since collecting adequate target data and information might be extremely expensive and laborious.

2.2 Domain Generalization

Domain generalization (DG) [4], [11], [32], [34], [42], [53], [60], [77], [93], [94], [99], [100] aims to train a stable model to unknown target domains by learning invariant knowledge from multiple source domains. A direct idea for DG is to align multi-source data distributions in latent space for generating invariant semantic representation [33], [34], [40], [41], [53], [60], [62], [99]. For example, Li et al. [41] extract domain-invariant representations of multi-source joint distributions through a conditional invariant adversarial network. Another set of works [2], [12], [30], [32], [42] are based on meta-learning, they employ an episodic training paradigm that trains the model and improves its out-of-distribution generalization ability on meta-train and meta-test datasets, respectively, which are built by the shared multi-source data. For instance, Dou et al. [12] present a model-agnostic meta-learning training paradigm with two complementary losses to consider both global knowledge and local cohesion. Data augmentation [5], [66], [74], [80], [101], [102], [103] for DG is also popular which trains the model on generated novel domains for improving model generalization. Among them, JiGen [5] is a representative work that utilizes the data with disordered patches to train the model for solving a jigsaw puzzle. Some other works [6], [24], [65] optimize the regularization terms of the data or networks to obtain generalization performance gain. These methods are mostly in thrall to shared multi-source data for identifying domain invariance and boosting model generalization, while concerns about data privacy are thus raised as tremendous private data might be distributed across separated places in many real scenarios. In comparison, we investigate a more practical setting of federated domain generalization towards privacy-preserving model training by accessing and learning each source dataset locally.

2.3 Federated Learning

As an active research field towards modern privacy protection, *federated learning* (FL) [16], [23], [37], [45], [55], [91], [92], [97], [104] makes local clients jointly train a model with a central server and keeps data decentralized. Take a representative paradigm FedAvg [55] as an example. In each communication round, a subset of the clients is chosen to receive the parameters of a global model from the server and trains it on their local data. The trained models are then transmitted back to the server for updating the global model with data-size based weights. Our investigated federated domain generalization task is closely related to federated learning as the data is decentralized, but the former is much more challenging: FL mainly focuses on guaranteeing model convergence when training on non-i.i.d. data [91], and improving model performance on the *"known"* clients. In contrast, our goal is to capture domain invariance from the separated source domains and train a generalizable model for the *"unknown"* out-of-distribution target domains.

2.4 Federated Domain Adaptation and Generalization

Source-free domain adaptation [29], [38] improves performance on the target domain by using a source pretrained model. Federated learning-based domain adaptation [26], [58] adapts models from distributed source domains to target domains. However, these methods are limited to the strong assumption of available target data/information, as we discussed previously. To tackle this issue, lots of federated learning-based domain generalization methods [7], [8], [10], [13], [21], [57], [67], [69], [71], [73], [79], [83], [85], [87], [96] have been proposed. [57] aims to learn simple representation of the distributed data with L2-norm and conditional mutual information constraints. FedDG [48] is a representative method, which augments the distributed data in frequency space. However, it builds an amplitude spectrum distribution bank from the source data and shares it to all the clients, which might be time-consuming and needs high communication costs. Meanwhile, it shares the amplitude spectrum of the source data to all the clients that may increase the risks of data privacy leakage. In comparison, our proposed method do not have extra time-consuming procedures like building such a distribution bank. More importantly, we do not share any data (or parts of its information) during training for efficient communication and effective data privacy protection.

Beyond computer vision, distributed training a generalizable model has also been explored in other fields, like natural language processing [20], [44], [46], [76], [84], recommender system [25], [78], [90], speech recognition [64], edge computing [68], etc. For example, [76] proposes a plug-and-play knowledge composition module to share knowledge across clients for non-i.i.d. multilingual natural language understanding. [78] proposes a recommendation model on decentralized domains, which learns data from user devices and trains a robust model by clipping training gradient. Compared with these works, we aim to solve the visual federated domain generalization problem by distributed learning invariant semantics of images through semantic aggregation and calibration processes for privacy protection.

3 Method

We begin with the problem definition of the federated domain generalization and its challenges for generalizable model learning. We then introduce our method CSAC (see Fig. 2) for addressing this challenging problem in detail.

3.1 Federated Domain Generalization

In federated domain generalization, given source datasets $\{\mathcal{D}^1, ..., \mathcal{D}^H\}$ from H distributed domains. There are N^h data sampled from domain-specific distribution $P(X^h, Y^h)$ in each dataset $\mathcal{D}^h$, i.e., $\mathcal{D}^h = \{(x_i^h, y_i^h)\}_{i=1}^{N^h}$, defined on image and label spaces $\mathcal{X} \times \mathcal{Y}$. The goal is to utilize the distributed source datasets for training a generalizable model, which can perform well on unknown target domains.

The challenges of this task are: (1) The source datasets are separated and can only be utilized locally, which greatly hinders the simultaneous learning of the multi-source semantic information and even leads to invalid domain invariance identification; (2) The heterogeneous source datasets with distinct data distributions constitute enormous obstacles for generalizable model learning, since the model can only access one local dataset each time. That is, if the exploited dataset contains unusual domain-specific bias, the trained model may even exhibit a negative generalization gain.

3.2 Overview of Our Method

The key idea of our method CSAC (Fig. 2) is to fully absorb multi-source information and precisely align semantic levels, which contains three main processes: (1) *Local semantic acquisition* for learning distribution information of local data. (2) *Data-free semantic aggregation* for semantic information gathering from the trained models. (3) *Cross-layer semantic calibration* for semantic level alignment and domain invariance enhancement. After obtaining local distribution information in process (1), the latter two processes are repeated alternately, unifying the multi-source semantic learning and alignment in a collaborative way for generalizable model training. Note that by following the algorithms of federated learning [16], [45], [55], [91], [92], we only transmit models among the distributed domains, and neither data nor its information is shared, which adequately preserves privacy. To our knowledge, the practice of using the same model structure for heterogeneous data is widely adopted in domain generalization and federated learning researches, like [48], [70], [92]. Therefore, we argue that it is feasible for our method to adopt the same model architecture for heterogeneous source data due to the powerful representation learning ability of deep models, as demonstrated by both the previous works and our experiments.

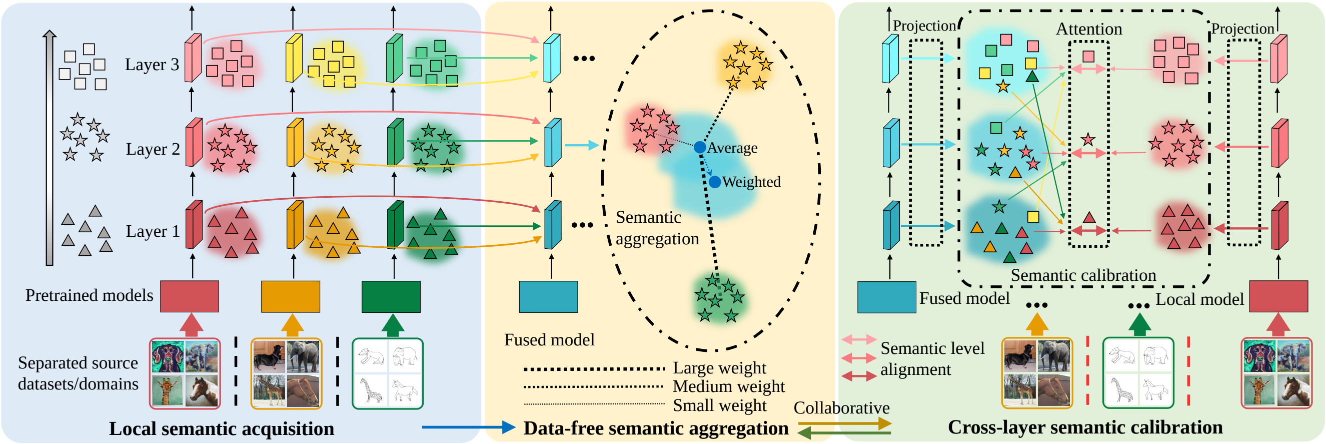

Fig. 2. Overview of the proposed framework of Collaborative Semantic Aggregation and Calibration (CSAC). Left: To learn data distribution information of the local data, one model is trained on each source domain to extract semantics layer-by-layer, i.e., ▲, ★, and ■. Middle: The trained models are fused layer-by-layer with semantic divergence based weights for data-free semantic gathering. Right: Cross-layer feature pairs between the fused model and a local model are matched with attention on each source domain for semantic level alignment and domain invariance enhancement. After semantic acquisition, the semantic aggregation and calibration processes are repeated alternately.

3.3 Local Semantic Acquisition

Before gathering and aligning the multi-source semantic information, we need to fully obtain the data distribution information of the separated source datasets. To avoid unsafe data fusion, we assign one model on each of the separated domains to impose data distribution learning. Given H separated source datasets, y^h with C categories is the groud-truth label of the image x^h in dataset $\mathcal{D}^h$, where $h \in \{1, ..., H\}$. Let $\{G^h\}_{h=1}^{H}$ be the trained models, each model G^h can be optimized on the local source data $\mathcal{D}^h$ with the following cross-entropy loss:

$$\mathcal{L}_{CE}^{h} = -\mathbb{E}_{(x^h, y^h)\in \mathcal{D}^h}[\sum_{c=1}^{C} \mathbb{1}(y^h = c) \log G^{h,c}(x^h)], \qquad (1)$$

where $G^{h,c}$ is the c-th dimension of the output of model G^h. $\mathbb{1}(\cdot)$ is an indicator function that equals to 1 for the correct condition and 0 for the rest. To facilitate the following semantic aggregation and calibration processes, we further introduce label smoothing [43], [56] to encourage data representations to group in tight evenly clusters, preventing the trained models from being over-confident. The updated learning loss for each trained model G^h is

$$\mathcal{L}_{LS}^{h} = -\mathbb{E}_{(x^h, y^h)\in \mathcal{D}^h}[\sum_{c=1}^{C} p^{h,c} \log G^{h,c}(x^h)], \qquad (2)$$

where $p^{h,c} = (1-\alpha)\mathbb{1}(y^h = c) + \alpha/C$ is the smoothed label. α is a smoothing hyper-parameter empirically set to 0.1 [56].

3.4 Data-Free Semantic Aggregation

After acquiring distribution information of local data, we devise a data-free semantic aggregation strategy with the trained models $\{G^h\}_{h=1}^{H}$. Inspired by recent researches [81], [98] on the interpretability of deep neural networks, we hypothesize that the deep models extract semantic information layer-by-layer, and the model parameters in each layer are related to the corresponding semantic level as well as the training data distribution (proof-of-concept experiments are shown in Sec. 4). To this end, we propose to fuse the trained models layer-by-layer for gathering different levels of semantics from the separated source domains. Since each model G^h trained on the data $\mathcal{D}^h$, it extracts hierarchical semantics from distribution of $\mathcal{D}^h$. Let G_l^h be the l-th layer of G^h, we have the average model parameter distribution in the l-th layer, that is,

$$G_l^{AVG} = \frac{1}{H}\sum_{h=1}^{H} G_l^h. \qquad (3)$$

We find that if a source data have a distribution far from the others, the parameters of the model trained on it would be distinct from the parameters of the other models. Therefore, this model would be considered less as it is far from the average distribution G_l^{AVG} if we directly use G_l^{AVG} as the final model. To fairly fuse the information of the source datasets for precise semantic calibration, we assign weight to each model based on its semantic divergence to G_l^{AVG}:

$$M_l = \sum_{h=1}^{H} \frac{\mathrm{dist}(G_l^h, G_l^{AVG})}{\sum_{h=1}^{H} \mathrm{dist}(G_l^h, G_l^{AVG})} G_l^h, \qquad (4)$$

where M_l is the l-th layer of the fused model M, $\mathrm{dist}(\cdot,\cdot)$ is distance metric and we empirically use L_2 distance (more discussions about it are in Sec. 4). Models with distinct parameters, or training data distributions, will be paid more attention to by being given a large weight. As the trained models are fused layer-by-layer, different levels of semantics from the separated source domains are aggregated for domain invariance learning in the semantic calibration.

3.5 Cross-Layer Semantic Calibration

Due to the different data distributions of the source datasets, i.e., domain shift, the same level of semantic information from different domains could be distributed across the layers of the fused model M during the aggregation process, which we call the *semantic dislocation* problem. To calibrate each level of semantic information for improving model generalization, we align each cross-layer semantic feature pair between the fused model M and a local model L^h (trained on each local source domain like G^h). Take Fig. 2 (right) as an example. The second layer of each local model mainly contains the second level of semantic information, i.e., ★. We match it with each layer of the fused model to align the second semantic level, i.e., match ★ in different layers of the fused model, where each cross-layer pair is weighted by their semantic similarities. Since the hierarchical semantic features may have different size, we first project them to the same size (we use the size of the semantic features in the last adopted layer in the experiments):

$$\begin{aligned} M_l(f_l)' &= Proj(M_l(f_l)), \\ L_m^h(f_m^h)' &= Proj(L_m^h(f_m^h)), \end{aligned} \tag{5}$$

where $Proj(\cdot)$ is the projection function with one convolution layer (see experiments for details), f_l and f_m^h are the input features for the l-th of the fused model and the m-th layer of the local model, respectively. We align each cross-layer semantic feature pair (l, m) after projection:

$$\mathcal{L}_{AL}^h = \sum_{l \in \mathcal{R}} \sum_{m \in \mathcal{R}} \alpha_{l,m} D(M_l(f_l)', L_m^h(f_m^h)'), \tag{6}$$

where $D(\cdot, \cdot)$ is used to measure the distribution discrepancy. We minimize $\mathcal{L}_{AL}^h$ to optimize the fused model M on each domain h for performing semantic feature alignment. We adopt Maximum Mean Discrepancies (MMD) [19] for $D(\cdot, \cdot)$ by following [51], [72]. The set of layers $\mathcal{R}$ for alignment is given in experiments. A dynamic weight $\alpha_{l,m}$ for the layer pair (l, m) is based on the semantic similarity learned by the attention mechanism introduced in the following.

Attention mechanism. To encourage the cross-layer pairs with larger semantic similarity to be matched for precise semantic level alignment and domain invariance enhancement, meanwhile, weakening the pairs with less similarity for avoiding further semantic dislocation, we then introduce a semantic similarity based attention mechanism for the dynamic weight $\alpha_{l,m}$. Attention [1] is a widely adopted technique [14], [75], [95] for deciding which parts of the input features should be paid more attention to. Here, we consider the semantic inter-dependencies in both *position* and *channel* dimensions. Let c, g, and w be the channel, height, and width of the semantic features after projection, respectively. We first reshape $M_l(x)' \in \mathcal{R}^{c \times g \times w}$ and $L_m^h(x)' \in \mathcal{R}^{c \times g \times w}$ to $A_l \in \mathcal{R}^{c \times d}$ and $B_m \in \mathcal{R}^{c \times d}$, respectively, where $d = g \times w$ is the number of pixels in an image. Then, we have a position-wise weight

$$\alpha_{l,m}^p = \frac{\exp\left(\operatorname{avg}\left(A_l^\top B_m\right)\right)}{\sum_{m \in \mathcal{R}} \exp\left(\operatorname{avg}\left(A_l^\top B_m\right)\right)}, \tag{7}$$

where $A_l^\top B_m$ is the position-wise attention map, measuring the response of A_l to B_m, i.e., the l-th layer of the fused model M_l to the m-th layer of the local model L_m^h. The operator $\operatorname{Avg}(\cdot)$ averages the attention map to a real number, and the weight $\alpha_{l,m}^p$ is the normalization of the average position-wise semantic similarity. Similarly, we then have

$$\alpha_{l,m}^c = \frac{\exp\left(\operatorname{avg}\left(A_l B_m^\top\right)\right)}{\sum_{m \in \mathcal{R}} \exp\left(\operatorname{avg}\left(A_l B_m^\top\right)\right)}. \tag{8}$$

$\alpha_{l,m}^c$ measures the channel-wise semantic similarities. We average them to get the final weight for each pair (l, m):

$$\alpha_{l,m} = \frac{1}{2}\left(\alpha_{l,m}^p + \alpha_{l,m}^c\right). \tag{9}$$

$\alpha_{l,m}$ characterizes the inter-dependencies between the fused model G_l and each local model L_m^h in both position and channel dimensions, weighting cross-layer pairs for calibrating semantic levels and boosting model generalization. We adopt attention to help the model identify which feature pairs are semantically-related and which are unrelated by calculating their semantic inter-dependencies in both position and channel dimensions. Through this design, the features from different layers between the global and the local model, which have similar semantic representation, would be automatically given a larger weight for alignment, and vice versa. Therefore, we argue that our method would not mix-ups the semantic features of different layers, but align and calibrate the semantics of different layers through the dynamic attention mechanism. Extensive experiments and ablation studies also demonstrate its effectiveness.

Since the semantic information from each domain is aggregated with others, the fused model may suffer from the *catastrophic forgetting* problem [18], [54], i.e., knowledge from one domain in the model is gradually forgotten when incrementally updating models with knowledge from other domains. We thus employ an auxiliary retraining loss $\mathcal{L}_{AR}^h$ for the model M on each source dataset $\mathcal{D}^h$, that is,

$$\mathcal{L}_{AR}^h = -\mathbb{E}_{(x^h, y^h) \in \mathcal{D}^h}\left[\sum_{c=1}^{C} \mathbb{1}(y^h = c) \log M^c(x^h)\right], \tag{10}$$

where M^c is the c-th dimension of the output of the model M. Then we have the calibration loss on each dataset $\mathcal{D}^h$:

$$\mathcal{L}_{CB}^h = \lambda \mathcal{L}_{AL}^h + \mathcal{L}_{AR}^h, \tag{11}$$

where λ is a hyper-parameter for semantic calibration. We minimize $\mathcal{L}_{CB}^h$ to optimized M on each domain h. Note that optimizing the retraining loss of L_{AR}^h on each client (domain) is the common practice of federated learning, like FedAvg. Therefore, we keep this loss unchanged, i.e., without using a hyper-parameter, such that we can conduct a clear ablation study and make a fair comparison with FedAvg. In addition, we try to avoid using more hyper-parameters to simplify our method.

3.6 Model Optimization

We first perform local semantic acquisition by training the models $\{G^h\}_{h=1}^H$ with Equation (2). The trained models are employed to calculate the fused model M for semantic aggregation through Equation (4). We then copy and transmit M to each domain and optimize it through Equation (11) for semantic calibration, then fuse the calibrated fused models again. We repeat the semantic aggregation and calibration alternately to gather semantic information from the distributed source domains and calibrate it to enhance domain-invariant information, resulting in a highly generalizable model $\hat{M}$ for inference on unknown target domains.

Remark. In practice, we assign the parameters of the fused model M to each trained model G^h for simultaneous semantic calibration on each domain h, and fuse them again.

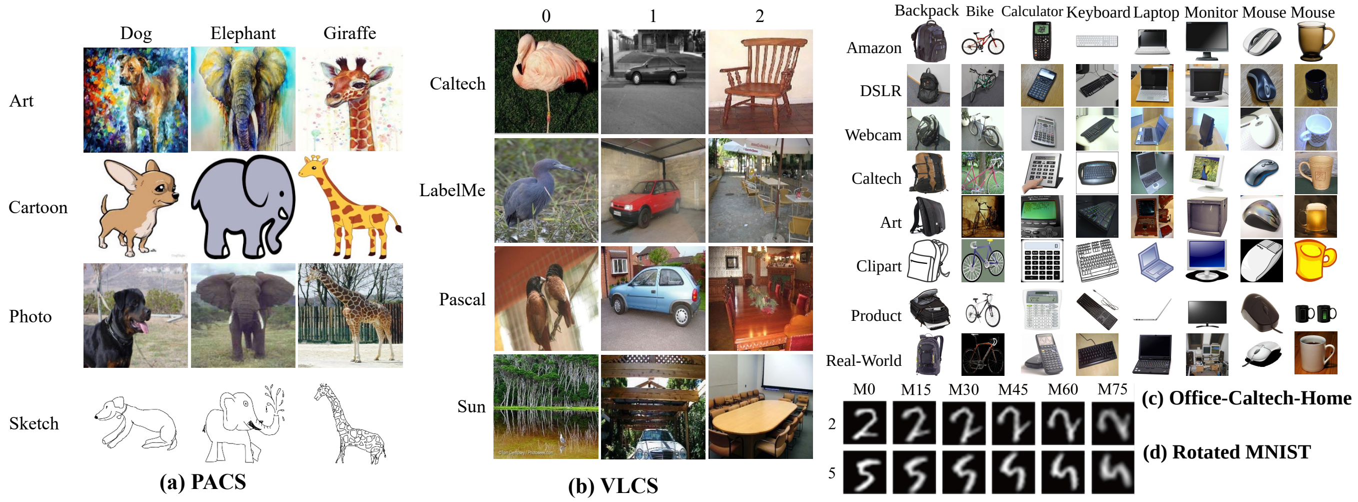

Fig. 3. Some example images of the adopted datasets for experiments, i.e., PACS (a), VLCS (b), Office-Caltech-Home (c), and Rotated MNIST (d).

4 Experiments

4.1 Setup

In this section, we evaluate the proposed CSAC method for the federated domain generalization task on multiple datasets, and give in-depth ablation studies and discussions.

Benchmark datasets. We first adopt two popular datasets of object recognition. One is *PACS* [31] that covers 7 categories within 4 domains, i.e., *Art*, *Cartoon*, *Sketch*, and *Photo*. Another is *VLCS* [15] that contains 5 classes from 4 domains, i.e., *Pascal*, *LabelMe*, *Caltech*, and *Sun*. A simulated digit dataset *Rotated MNIST* [15] is then employed. It has 6 domains, i.e., *M0*, *M15*, *M30*, *M45*, *M60*, and *M75* through clock-wise rotation of the original images (M0) five times with fifteen degree intervals. We use 100 images per class for Rotated MNIST dataset by following [15], [99]. We process the data by following the previous works [24], [99]. To evaluate the performance under the scenarios with more domains, we further construct a dataset *Office-Caltech-Home* [93] by choosing the common classes from Office-Caltech [17] and Office-Home [72] datasets, and merge them to get 7 domains (the domain DSLR is discarded since it only contains a few images), i.e., Amazon (*Am*), Webcam (*We*), Caltech (*Ca*), Art (*Ar*), Clipart (*Cl*), Product (*Pr*), and Real-World (*Rw*). Some representative example images of these adopted datasets are shown in Fig. 3. We conduct leave-one-domain-out experiments, i.e., choosing one domain from each dataset to hold out as the target domain, the others are used as the (distributed) source domains. We train the model on each domain, and only transmit model parameters among domains by following [16], [45], [55], [91], [92].

Baseline methods. We compare our method CSAC against the representative federated learning method *FedAvg* [55] and the federated learning based generalizable model learning method *FedDG* [48] in the separated domain generalization task. We also show the performance of the state-of-the-art DG methods (see Table 1, 2, and 3) introduced in the Sec. 2 for the domain generalization task with shared source data. Following the previous works [5], [12], [99], we implement a baseline method *DeepAll* by employing the fusion of the shared source datasets for model training.

TABLE 1
Accuracy (%) on PACS dataset. "Sep.": whether using separated source datasets. "*": the methods implemented by us. The best results are emphasized in bold.

Methods	Sep.	Art	Cartoon	Photo	Sketch	Average
DeepAll*	✗	78.95±0.48	74.90±1.82	94.08±0.55	73.02±0.80	80.24±0.50
JiGen [5]	✗	79.42	75.25	96.03	71.35	80.51
MASF [12]	✗	80.29	77.17	94.99	71.69	81.04
DGER [99]	✗	80.70	76.40	**96.65**	71.77	81.38
DMG [6]	✗	76.90	**80.38**	92.35	75.21	81.46
FACT [88]	✗	**85.37**	78.38	95.15	**79.15**	**84.51**
Epi-FCR [32]	✗	82.1	77.0	93.9	73.0	81.5
MixSyle [103]	✗	84.1	78.8	96.1	75.9	83.7
EISNet [80]	✗	81.89	76.44	95.93	74.33	82.15
FedADG [96]	✓	77.8±0.5	74.7±0.4	92.9±0.3	79.5±0.4	81.2
FedCMI [57]	✓	80.8±0.4	73.7±0.2	92.8±0.5	79.5±0.2	81.7
FedSR [57]	✓	**83.2**±0.3	76.0±0.3	93.8±0.5	**81.9**±0.2	83.7
FedL2R [57]	✓	82.2±0.4	75.8±0.3	92.8±0.4	81.6±0.1	83.1
FedAvg* [55]	✓	77.49±0.10	**77.21**±0.52	93.56±0.38	81.19±0.80	82.36±0.44
FedDG* [48]	✓	78.46±0.20	75.98±0.28	93.23±0.43	80.92±0.72	82.15±0.35
CSAC* (ours)	✓	81.98±0.87	76.41±0.49	**95.20**±0.29	81.64±0.49	**83.81**±0.33

Implementation details. We use the pretrained ResNet-18 network [22] for PACS, VLCS, and Office-Caltech-Home datasets and also use the AlexNet network [28] for VLCS by following [5], [12], [24]. We use standard MNIST CNN architecture with two convolution layers and two fully-connected layers for Rotated MNIST dataset by following [15], [99]. We extract the convolution layers of the last three blocks of ResNet-18, or the last three convolution layers of AlexNet, or the last two convolution layers of MNIST CNN, as the layer set for semantic calibration. We implement the methods according to their public code, where the boundary component (it is used for segmentation task) of FedDG is discarded for fair comparison. We use SGD optimizer with learning rate 0.01 and momentum 0.5 for ResNet-18 and MNIST CNN, and learning rate 0.001 for AlexNet. The training epochs for semantic acquisition are set to 30, collaboration rounds for aggregation and calibration are set to 40, for all the datasets. In each round, the calibration epochs are set to 5 and 10 for the Rotated MNIST and the other datasets, respectively. The hyper-parameter λ is set to 0.6 for all the experiments, its sensitivity is further analyzed. We run the experiments on a device with CPU Xeon Gold 6254

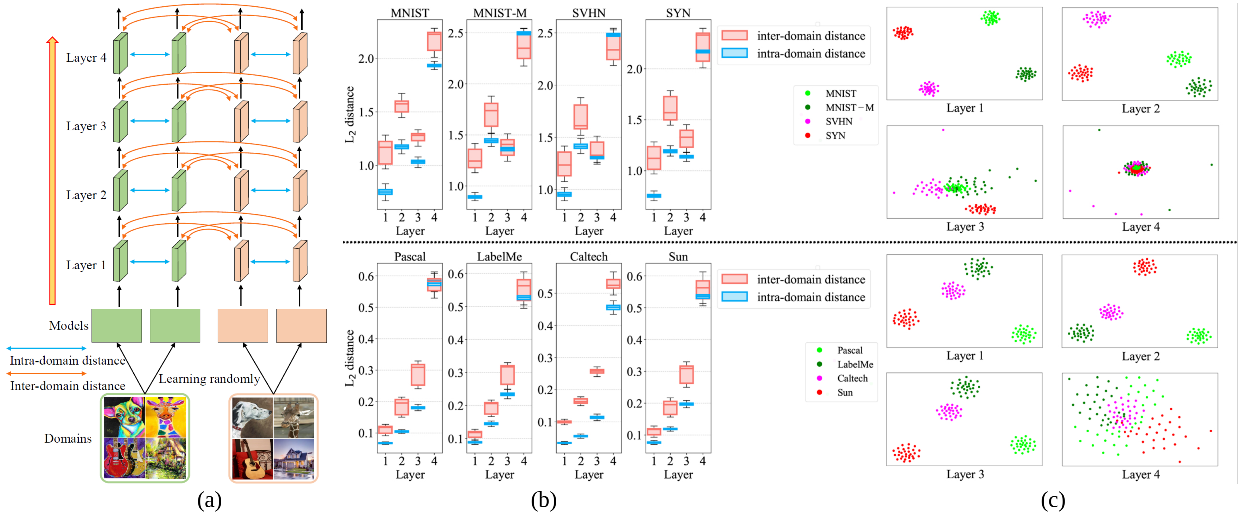

Fig. 4. Results of proof-of-concept experiments. 50 models with ResNet-18 architecture are randomly run on each domain of Digits-DG (domains: MNIST, MNIST-M, SVHN, and SYN) and VLCS datasets (domains: Pascal, LabelMe, Caltech, and Sun). (a): the model parameters of the last convolution layer of the four blocks of the trained models are extracted to be layer {1, 2, 3, 4} and their intra- and inter-domain L_2 distance are calculated. (b): the calculated intra- and inter-domain L_2 distance of the model parameters in each layer (above: Digits-DG, below: VLCS). (c): t-SNE visualization [52] of the model parameters (each point represents a trained model) in each layer (above: Digits-DG, below: VLCS).

TABLE 2
Accuracy (%) on VLCS dataset. "Sep.": whether using separated source datasets. "*": the methods implemented by us. The best results are emphasized in bold.

Methods	Sep.	Pascal	LabelMe	Caltech	Sun	Average
			AlexNet			
DeepAll*	✗	71.67±0.26	59.64±0.81	**97.48**±0.14	67.58±0.68	74.09±0.17
Epi-FCR [32]	✗	67.1	64.3	94.1	65.9	72.9
JiGen [5]	✗	70.62	60.90	96.93	64.30	73.19
MASF [12]	✗	69.14	**64.90**	94.78	67.64	74.11
DGER [99]	✗	**73.24**	58.26	96.92	**69.10**	74.38
EISNet [80]	✗	69.83	63.49	97.33	68.02	**74.67**
FedAvg* [55]	✓	67.92±0.26	**60.23**±0.81	96.85±0.26	66.88±0.22	72.97±0.16
FedDG* [48]	✓	67.27±0.07	58.48±0.04	96.83±0.47	**68.20**±0.12	72.69±0.16
CSAC* (ours)	✓	**70.21**±0.32	58.99±0.29	**97.13**±0.35	67.27±0.54	**73.40**±0.17
			ResNet-18			
DeepAll*	✗	71.40±0.32	59.77±0.95	97.54±0.54	69.01±0.25	74.43± 0.25
JiGen* [5]	✗	73.97±0.21	61.94±0.74	97.40±1.03	66.90±0.64	75.05±0.26
COPA [85]	✓	71.50±1.05	61.00±0.89	93.83±0.41	71.72±0.74	74.51
FedAvg* [55]	✓	71.95±0.06	63.29±0.06	96.48±0.18	72.37±0.06	76.02±0.08
FedDG* [48]	✓	**72.59**±0.30	60.33±0.07	96.70±0.20	**73.61**±0.17	75.81±0.16
CSAC* (ours)	✓	71.97±0.56	**63.45**±0.73	**97.24**±0.57	72.06±0.80	**76.18**±0.42

TABLE 3
Accuracy (%) on Rotated MNIST dataset. "Sep.": whether using separated source datasets. "*": the methods implemented by us. The best results are emphasized in bold.

Methods	Sep.	M0	M15	M30	M45	M60	M75	Average
DeepAll*	✗	86.73±0.45	98.27±0.40	98.63±0.15	97.50±0.89	97.47±0.25	87.20±0.95	94.30±0.29
CrossGrad	✗	86.03	98.92	98.60	98.39	98.68	88.94	94.93
MetaReg [2]	✗	85.70	98.87	98.32	98.58	98.93	89.44	94.97
FeaCri [42]	✗	87.04	**99.53**	**99.41**	**99.52**	99.23	**91.52**	96.04
DGER [99]	✗	**90.09**	99.24	99.27	99.31	**99.45**	90.81	**96.36**
FedAvg* [55]	✓	82.60±0.44	98.56±0.27	**98.97**±0.29	93.66±0.03	95.78±0.27	86.30±0.10	92.65±0.08
FedDG* [48]	✓	73.07±0.67	94.37±1.03	95.60±0.19	89.43±0.38	94.61±0.39	84.50±0.10	88.60±0.30
CSAC* (ours)	✓	**84.57**±0.31	**98.87**±0.23	98.63±0.15	**95.06**±0.48	**96.57**±0.40	**90.73**±0.25	**94.07**±0.02

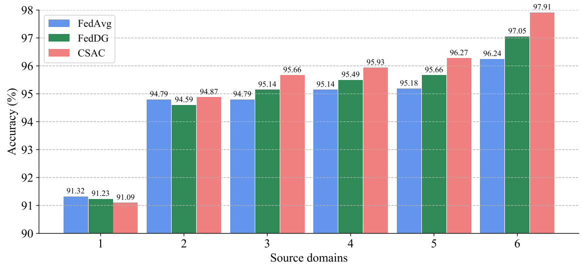

Fig. 5. Results on Office-Caltech-Home dataset with 7 domains. We let domain Rw be the target domain, and add one source domain each time from a source domain set {Am, Cl, Pr, We, Ar, Ca}, i.e., using 1 source domain: {Am}, using 2 source domains: {Am, Cl}, and so on.

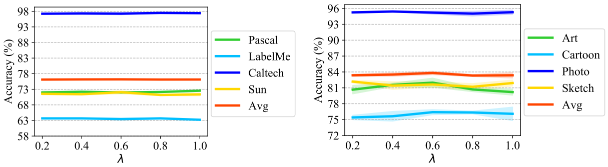

Fig. 6. Sensitivity analysis of the hyper-parameter λ for semantic calibration on VLCS (left) and PACS (right) datasets.

× 2, and GPU Nvidia RTX 2080 TI × 4. We report the mean and standard error of the classification accuracy over five runs with random seeds for the experiments implemented by us (marked with *). And we cite other results of the DG methods from the published papers like the previous works.

4.2 Proof-of-Concept Experiments

We first provide proof-of-concept experiments to verify our hypothesis: the deep models extract semantic information layer-by-layer, and the model parameters in each layer are related to the corresponding level as well as the training data distribution. We randomly run 50 models with ResNet-18 architecture on each domain of VLCS [15] and Digits-DG [101] datasets. After training, we extract the model parameters of the last convolution layer of the four blocks of the trained models to be layer {1, 2, 3, 4}. We calculate

TABLE 4
Effect of semantic aggregation with semantic divergence (strategy) and L_2 distance (metric). The best results are emphasized in bold.

Type	Case	PACS					VLCS				
		Art	Cartoon	Photo	Sketch	Average	Pascal	LabelMe	Caltech	Sun	Average
Strategy	Semantic similarity	79.22±0.19	72.44±0.53	93.00±0.14	74.34±0.37	79.75±0.22	70.80±0.16	62.86±0.39	96.61±0.26	69.49±1.03	74.94±0.46
	Semantic average	80.25±0.36	74.54±0.44	93.74±0.15	77.88±0.69	81.85±0.07	**72.44**±0.17	61.20±0.32	96.54±0.30	70.83±0.45	75.25±0.21
Metric	Cosine distance	80.89±0.23	73.57±0.41	94.43±0.03	76.59±1.31	81.37±0.41	72.27±0.11	62.94±0.44	97.05±0.15	71.41±0.06	75.92±0.11
	L_1 distance	81.75±0.51	74.93±0.09	94.45±0.28	77.63±0.42	82.19±0.27	72.16±0.16	**64.11**±0.28	96.64±0.13	71.67±0.21	76.15±0.04
	CSAC	**81.98**±0.87	**76.41**±0.49	**95.20**±0.29	**81.64**±0.49	**83.81**±0.33	71.97±0.56	63.45±0.73	**97.24**±0.57	**72.06**±0.80	**76.18**±0.42

TABLE 5
Effect of semantic calibration with cross-layer alignment (strategy) and MMD discrepancy (metric). The best results are emphasized in bold.

Type	Case	PACS					VLCS				
		Art	Cartoon	Photo	Sketch	Average	Pascal	LabelMe	Caltech	Sun	Average
Strategy	Without alignment	80.34±0.24	76.07±0.14	**95.45**±0.49	81.11±0.12	83.24±0.15	70.11±1.21	**65.73**±2.31	97.03±0.15	71.39±0.16	76.06±0.75
	Same-layer alignment	80.49±0.23	74.19±0.11	94.85±0.55	80.34±0.44	82.46±0.17	**73.51**±0.21	62.26±0.14	97.38±0.08	71.38±0.14	76.14±0.12
	Without attention (position)	79.57±0.44	73.57±0.61	94.46±0.12	79.66±0.47	81.82±0.16	72.57±0.12	63.85±0.78	96.09±0.26	69.16±0.21	75.42±0.31
	Without attention (channel)	79.80±0.58	73.77±0.03	94.43±0.09	79.32±0.19	81.83±0.14	72.40±0.52	63.54±0.52	95.92±0.30	69.24±0.12	75.27±0.20
	Without attention	80.35±0.87	76.27±0.11	93.59±0.24	77.88±1.53	82.02±0.58	71.63±0.06	64.48±2.46	96.27±0.40	67.67±1.03	75.01±0.44
	Without label smoothing	81.78±1.00	76.37±0.43	95.00±0.71	81.46±0.80	83.65±0.52	71.70±0.78	63.33±0.82	97.17±0.61	71.34±1.18	75.89±0.54
	Without cross-entropy	81.35±0.09	76.04±0.17	94.98±0.39	**82.60**±0.50	83.74±0.13	72.12±0.23	62.70±0.16	**97.55**±0.21	71.76±0.23	76.03±0.10
Metric	Mean Square Error (MSE)	77.08±0.28	71.05±1.26	94.61±0.14	75.99±0.33	79.68±0.49	72.10±0.79	62.17±0.19	96.47±0.35	71.09±0.14	75.46±0.26
	CSAC	**81.98**±0.87	**76.41**±0.49	95.20±0.29	81.64±0.49	**83.81**±0.33	71.97±0.56	63.45±0.73	97.24±0.57	**72.06**±0.80	**76.18**±0.42

TABLE 6
Run time (hours) on PACS and VLCS datasets.

Methods	PACS					VLCS				
	Photo	Art	Cartoon	Sketch	Average	Pascal	LabelMe	Caltech	Sun	Average
FedAvg [55]	3.85	3.95	4.03	4.32	4.04	3.03	3.10	3.08	3.11	3.08
FedDG [48]	11.65	11.84	11.92	12.01	11.86	10.81	10.43	10.50	10.19	10.48
CSAC (ours)	4.81	4.81	4.48	4.34	4.61	3.61	3.61	3.57	3.59	3.60

intra-domain distance, i.e., the L_2 distance of the model parameters between each pair of the models trained on the same domain, and *inter-domain distance*, i.e., the L_2 distance of the model parameters between each pair of the models trained on different domains, as shown in Fig. 4 (a). From the distance results in Fig. 4 (b), we observe that the models trained on the same domain have closer parameter distance than the models trained on different domains, which is also verified by the t-SNE visualization [52] in Fig. 4 (c) that the points of the model parameters from the same domain gather together. It verifies that the model parameters are related to the distributions of training data. Meanwhile, we find that the inter-domain distance becomes closer to the intra-domain distance in the high layers in Fig. 4 (b) and the points with different colors cluster together in the high layers in Fig. 4 (c). It indicates that the model parameters are not only related to the training data distributions but also the corresponding semantic level, since the lower semantic level is more related to the data distribution while the higher level is more related to the object categories that is invariant to the domains. We argue that it is also the latent assumption of domain generalization task that one can extract high-level discriminative yet domain-agnostic semantics for training a highly generalizable model.

4.3 Main Results

We first report the results on PACS dataset in Table 1. We observe that our method CSAC achieves the highest average accuracy for the federated domain generalization task. Moreover, CSAC with separated source data even outperforms the domain generalization methods (except FACT) with shared source data on the average accuracy. It shows the effectiveness of our collaborative semantic aggregation and calibration strategy. Moreover, it indicates that we may learn a generalizable model by sharing information among domains through the model parameters. In this way, we can improve generalization performance of the model under careful data privacy protection, which is important for many privacy-sensitive real-world scenarios.

We further use models with AlextNet and ResNet-18 architecture for the experiments on VLCS dataset and report the results in Table 2. In the experiments of the two network architecture, our method CSAC surpasses the FedAvg and FedDG methods. It shows excellent generalization learning ability of CSAC with separated source data. With the AlexNet network architecture, CSAC slightly outperforms Epi-FCR and JiGen methods with shared source data. However, by using the larger network ResNet-18, CSAC performs better than DeepAll and JiGen. We attribute it to the semantic calibration process and the attention mechanism which may need large network to show their advantages.

We report the results on Rotated MNIST dataset in Table 3. It demonstrates that our method CSAC defeats FedAvg and FedDG methods for the distributed domain generalization task. We also observe that CSAC achieves slightly worse performance than the domain generalization methods with shared data. It is probably because we use a much smaller

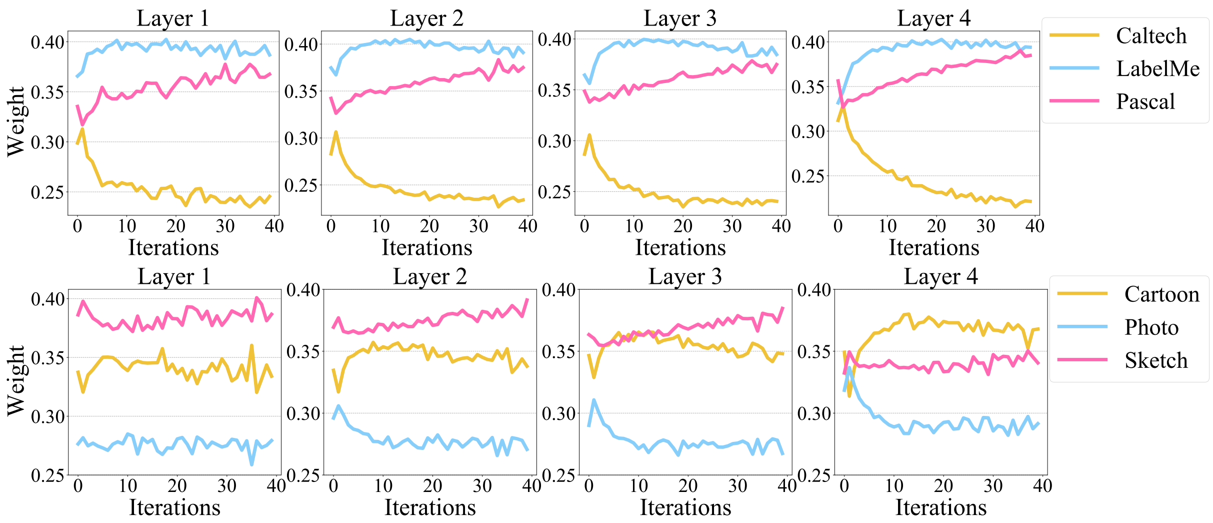

Fig. 7. Semantic divergence based weight for the last layer of the four blocks of ResNet-18 of each trained model (marked with the corresponding domain) during training on VLCS (above) and PACS (below).

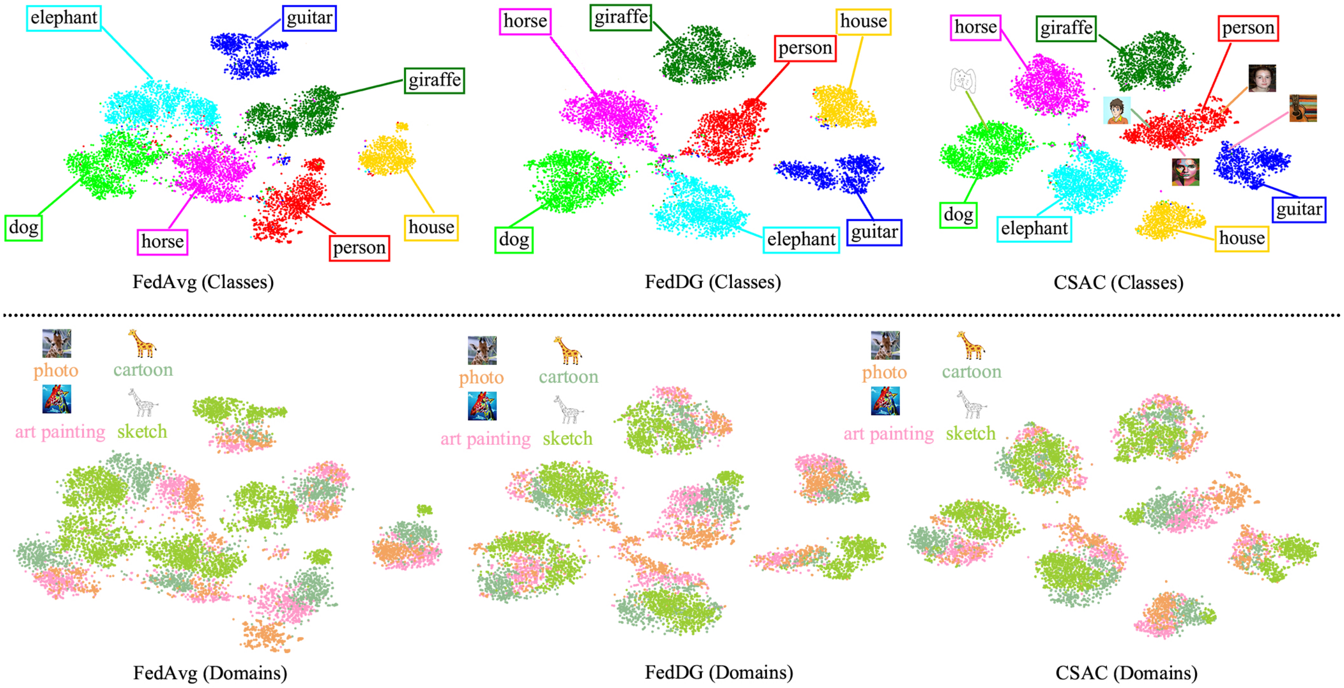

Fig. 8. T-SNE visualization of the learned semantic feature distributions of the data points on PACS dataset (target domain is photo). Different colors represent different classes (above) or domains (below).

network here for the Rotated MNIST dataset. It is similar to our previous conclusion that CSAC needs large network to unleash the potential of domain invariance learning.

In order to evaluate the methods under the scenarios with more domains, we further conduct experiments on the Office-Caltech-Home dataset with 7 domains. We let domain Rw be the target domain, and add one source domain each time from a source domain set {Am, Cl, Pr, We, Ar, Ca}, i.e., using 1 source domain: {Am}, using 2 source domains: {Am, Cl}, and so on. The results are shown in Fig. 5. We have two observations: (1) Our method CSAC surpasses other methods when given more than two source domains. (2) Giving more source domains enable CSAC to achieve much better performance. We argue that it is because the adequate semantic information provided by multiple source domains facilitate the semantic level alignment and invariance enhancement in the CSAC framework.

4.4 In-Depth Ablation Studies

Semantic aggregation. Table 4 reports ablation results for semantic aggregation. By replacing the strategy with semantic similarity (larger weights for the models that are closer to the average distribution) and average (equal weights), we find that it is important to pay more attention to the domain with the semantic distribution far from the others for fairly

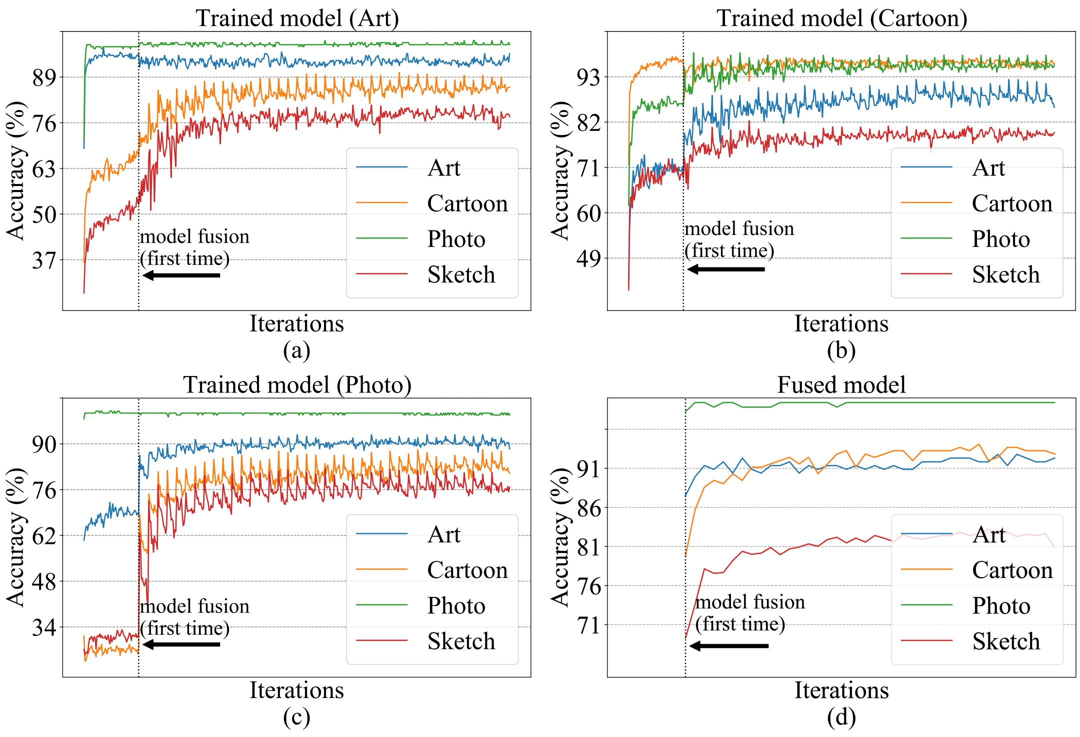

Fig. 9. Accuracy (on all the domains) of the model trained on domain Art (a), Cartoon (b), and Photo (c), and the fused model (d), during the training process, i.e., semantic acquisition and the repeat of semantic aggregation and calibration. The adopted dataset (target domain): PACS (Sketch).

absorbing knowledge from all the source domain, facilitating valid domain invariance learning. The experiments with other metrics show the effectiveness of the used L_2 distance.

Semantic calibration. Table 5 reports the ablation results for semantic calibration. We compare the results without alignment and using the same-layer alignment and find that it is necessary to consider cross-layer semantic relationships for addressing the semantic dislocation problem. By conducting the attention ablations, we demonstrate the attention mechanism with both position and channel interdependency consideration is important for precise semantic level alignment. The label smoothing is showed useful for a final generalizable model learning, which is may because it leads to more smooth models for stable model fusion. Cross-entropy displays its importance for catastrophic forgetting. The MMD metric is more effective for alignment than the MSE, which may also be the reason that MMD is widely adopted in alignment-based domain adaptation works.

Sensitivity analysis. Fig. 6 shows that CSAC is generally robust to the weight of semantic calibration, i.e., hyper-parameter λ. CSAC might be practical and effective without the time-consuming hyper-parameter fine-tuning.

4.5 Run Time

We report the run time of the methods (implemented locally) in Table 6. FedDG is computationally inefficient by using about three times the run time of FedAvg and CSAC, which is may because of the time-consuming process of the distribution bank building. Besides, FedDG transmits the bank to all the domains, which needs high communication costs and might increase the risks of privacy leakage (although we can not verify it with experiments).

4.6 Why Does CSAC Work?

In Fig. 7, we observe that the weight curves have the similar trend, i.e., the four extracted layers of each model have the similar semantic divergence to others, which indirectly verifies our hypothesis that parameters are related to the data. The model with divergent semantics is given large weight layer-by-layer for adequate semantic gathering, facilitating the semantic calibration as shown in ablation studies.

Fig. 8 shows comparisons on the learned semantic feature distributions. CSAC obtains more discriminative and domain-agnostic information and generates class clear and domain compact semantic feature representations. We attribute it to the effectiveness of the collaborative semantic aggregation and calibration strategy for domain invariance learning with distributed source domains.

We then present insights on the proposed CSAC via showing the accuracy curves of the models on all the source datasets during training in Fig. 9 (note that the target dataset is only used for testing the model performance). During semantic acquisition, each trained model is assigned to each separated domain for data distribution learning, and its accuracy on the learned domain improves rapidly (see the parts before model fusion in the subfigures (a-c)). The accuracy of the trained models assigned to Art, Cartoon, and Photo domain, on the target dataset, i.e., Sketch (red curve at dotted line), is 52.74%, 68.47%, 32.02%, respectively, before model fusion. Then, the trained models are fused for semantic aggregation, each domain knowledge is fully gathered. The parameters of the fused model are then assigned to each trained model again for semantic calibration with local datasets. We unify semantic learning and alignment by repeating semantic aggregation and calibration alternately, the domain invariance from the separated domains is indirectly captured, making the accuracy of the fused model (subfigure (d)) on all the datasets improve gradually. By comparing the results before model fusion, the accuracy of the trained local model on the target dataset finally reaches improvement of more than 29%, 13%, and 50%.

5 Conclusions

Training a generalizable model is a vital issue for the deep learning community. However, common practices of domain generalization rely on shared multi-source data, which may violate privacy policies in many real-world applications. This paper tackles the privacy-preserving problem of federated domain generalization, and presents a novel method for this challenging task with collaborative semantic aggregation and calibration. Our method unifies multi-source semantic learning and alignment in a collaborative way, distributed improving model generalization under careful privacy protection. In future, one may be demanded to collaboratively train a generalizable model by exploiting thousands of separated source datasets. Thus, our work sheds some light on this promising direction which lacks extensive research. It is important for many privacy-sensitive scenarios like finance and medical care.

Acknowledgments

This work was supported by Key R & D Projects of the Ministry of Science and Technology (No. 2020YFC0832500), Young Elite Scientists Sponsorship Program by CAST, National Natural Science Foundation of China (No. 62006207), Zhejiang Province Natural Science Foundation (No. LQ21F020020, LZ22F020012).

References

[1] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. In *ICLR*, 2015.

[2] Y. Balaji, S. Sankaranarayanan, and R. Chellappa. Metareg: Towards domain generalization using meta-regularization. In *NeurIPS*, pages 998–1008, 2018.

[3] A. Bitarafan, M. S. Baghshah, and M. Gheisari. Incremental evolving domain adaptation. *IEEE Transactions on Knowledge and Data Engineering*, 28(8):2128–2141, 2016.

[4] G. Blanchard, G. Lee, and C. Scott. Generalizing from several related classification tasks to a new unlabeled sample. *NeurIPS*, 24:2178–2186, 2011.

[5] F. M. Carlucci, A. D'Innocente, S. Bucci, B. Caputo, and T. Tommasi. Domain generalization by solving jigsaw puzzles. *CVPR*, pages 2224–2233, 2019.

[6] P. Chattopadhyay, Y. Balaji, and J. Hoffman. Learning to balance specificity and invariance for in and out of domain generalization. In *ECCV*, pages 301–318. Springer, 2020.

[7] H. Chen, A. Frikha, D. Krompass, and V. Tresp. Fraug: Tackling federated learning with non-iid features via representation augmentation. *arXiv preprint arXiv:2205.14900*, 2022.

[8] J. Chen, M. Jiang, Q. Dou, and Q. Chen. Federated domain generalization for image recognition via cross-client style transfer. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 361–370, 2023.

[9] Q. Dai, X.-M. Wu, J. Xiao, X. Shen, and D. Wang. Graph transfer learning via adversarial domain adaptation with graph convolution. *IEEE Transactions on Knowledge and Data Engineering*, 2022.

[10] A. B. de Luca, G. Zhang, X. Chen, and Y. Yu. Mitigating data heterogeneity in federated learning with data augmentation. *arXiv preprint arXiv:2206.09979*, 2022.

[11] Z. Ding and Y. Fu. Deep domain generalization with structured low-rank constraint. *IEEE Transactions on Image Processing (TIP)*, 27(1):304–313, 2017.

[12] Q. Dou, D. C. de Castro, K. Kamnitsas, and B. Glocker. Domain generalization via model-agnostic learning of semantic features. In *NeurIPS*, 2019.

[13] A. Frikha, H. Chen, D. Krompaß, T. Runkler, and V. Tresp. Towards data-free domain generalization. *arXiv preprint arXiv:2110.04545*, 2021.

[14] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu. Dual attention network for scene segmentation. In *CVPR*, pages 3146–3154, 2019.

[15] M. Ghifary, W. Bastiaan Kleijn, M. Zhang, and D. Balduzzi. Domain generalization for object recognition with multi-task autoencoders. In *ICCV*, pages 2551–2559, 2015.

[16] A. Ghosh, J. Chung, D. Yin, and K. Ramchandran. An efficient framework for clustered federated learning. *NeurIPS*, 2020.

[17] B. Gong, Y. Shi, F. Sha, and K. Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *CVPR*, pages 2066–2073. IEEE, 2012.

[18] I. J. Goodfellow, M. Mirza, D. Xiao, A. Courville, and Y. Bengio. An empirical investigation of catastrophic forgetting in gradient-based neural networks. *arXiv*, 2013.

[19] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *JMLR*, 13(1):723–773, 2012.

[20] X. Guo, H. Yu, B. Li, H. Wang, P. Xing, S. Feng, Z. Nie, and C. Miao. Federated learning for personalized humor recognition. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 13(4):1–18, 2022.

[21] S. Gupta, K. Ahuja, M. Havaei, N. Chatterjee, and Y. Bengio. Fl games: A federated learning framework for distribution shifts. *arXiv preprint arXiv:2205.11101*, 2022.

[22] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[23] W. Huang, J. Liu, T. Li, T. Huang, S. Ji, and J. Wan. Feddsr: Daily schedule recommendation in a federated deep reinforcement learning framework. *IEEE Transactions on Knowledge and Data Engineering*, 2021.

[24] Z. Huang, H. Wang, E. P. Xing, and D. Huang. Self-challenging improves cross-domain generalization. In *ECCV*, pages 124–140, 2020.

[25] M. Imran, H. Yin, T. Chen, Q. V. H. Nguyen, A. Zhou, and K. Zheng. Refrs: Resource-efficient federated recommender system for dynamic and diversified user preferences. *ACM Transactions on Information Systems*, 41(3):1–30, 2023.

[26] E. Jiang, Y. J. Zhang, and O. Koyejo. Federated domain adaptation via gradient projection. *arXiv preprint arXiv:2302.05049*, 2023.

[27] W. Jiang, H. Gao, W. Lu, W. Liu, F.-L. Chung, and H. Huang. Stacked robust adaptively regularized auto-regressions for domain adaptation. *IEEE Transactions on Knowledge and Data Engineering*, 31(3):561–574, 2018.

[28] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Commun. ACM*, 60(6):84–90, 2017.

[29] J. N. Kundu, N. Venkat, R. V. Babu, et al. Universal source-free domain adaptation. In *CVPR*, pages 4544–4553, 2020.

[30] D. Li, Y. Yang, Y.-Z. Song, and T. Hospedales. Learning to generalize: Meta-learning for domain generalization. In *AAAI*, 2018.

[31] D. Li, Y. Yang, Y.-Z. Song, and T. M. Hospedales. Deeper, broader and artier domain generalization. In *ICCV*, pages 5542–5550, 2017.

[32] D. Li, J. Zhang, Y. Yang, C. Liu, Y.-Z. Song, and T. M. Hospedales. Episodic training for domain generalization. *ICCV*, pages 1446–1455, 2019.

[33] H. Li, S. Jialin Pan, S. Wang, and A. C. Kot. Domain generalization with adversarial feature learning. In *CVPR*, pages 5400–5409, 2018.

[34] H. Li, Y. Wang, R. Wan, S. Wang, T. Li, and A. C. Kot. Domain generalization for medical imaging classification with linear-dependency regularization. In *NeurIPS*, 2020.

[35] J. Li, M. Jing, H. Su, K. Lu, L. Zhu, and H. T. Shen. Faster domain adaptation networks. *IEEE Transactions on Knowledge and Data Engineering*, 2021.

[36] K. Li, J. Lu, H. Zuo, and G. Zhang. Dynamic classifier alignment for unsupervised multi-source domain adaptation. *IEEE Transactions on Knowledge and Data Engineering*, 2022.

[37] Q. Li, Z. Wen, Z. Wu, S. Hu, N. Wang, Y. Li, X. Liu, and B. He. A survey on federated learning systems: vision, hype and reality for data privacy and protection. *IEEE Transactions on Knowledge and Data Engineering*, 2021.

[38] R. Li, Q. Jiao, W. Cao, H.-S. Wong, and S. Wu. Model adaptation: Unsupervised domain adaptation without source data. In *CVPR*, pages 9641–9650, 2020.

[39] S. Li, W. Ma, J. Zhang, C. H. Liu, J. Liang, and G. Wang. Meta-reweighted regularization for unsupervised domain adaptation. *IEEE Transactions on Knowledge and Data Engineering*, 2021.

[40] Y. Li, M. Gong, X. Tian, T. Liu, and D. Tao. Domain generalization via conditional invariant representations. In *AAAI*, 2018.

[41] Y. Li, X. Tian, M. Gong, Y. Liu, T. Liu, K. Zhang, and D. Tao. Deep domain generalization via conditional invariant adversarial networks. In *ECCV*, pages 624–639, 2018.

[42] Y. Li, Y. Yang, W. Zhou, and T. Hospedales. Feature-critic networks for heterogeneous domain generalization. In *ICML*, pages 3915–3924. PMLR, 2019.

[43] J. Liang, D. Hu, and J. Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *ICML*, pages 6028–6039. PMLR, 2020.

[44] B. Y. Lin, C. He, Z. Zeng, H. Wang, Y. Huang, C. Dupuy, R. Gupta, M. Soltanolkotabi, X. Ren, and S. Avestimehr. Fednlp: Benchmarking federated learning methods for natural language processing tasks. *arXiv preprint arXiv:2104.08815*, 2021.

[45] T. Lin, L. Kong, S. U. Stich, and M. Jaggi. Ensemble distillation for robust model fusion in federated learning. *NeurIPS*, 2020.

[46] Z. Lit, S. Sit, J. Wang, and J. Xiao. Federated split bert for heterogeneous text classification. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2022.

[47] H. Liu, M. Shao, Z. Ding, and Y. Fu. Structure-preserved unsupervised domain adaptation. *IEEE Transactions on Knowledge and Data Engineering*, 31(4):799–812, 2018.

[48] Q. Liu, C. Chen, J. Qin, Q. Dou, and P.-A. Heng. Feddg: Federated domain generalization on medical image segmentation via episodic learning in continuous frequency space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1013–1023, 2021.

[49] M. Long, J. Wang, Y. Cao, J. Sun, and S. Y. Philip. Deep learning of transferable representation for scalable domain adaptation. *IEEE Transactions on Knowledge and Data Engineering*, 28(8):2027–2040, 2016.

[50] M. Long, J. Wang, G. Ding, S. J. Pan, and S. Y. Philip. Adaptation regularization: A general framework for transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 26(5):1076–1089, 2013.

[51] M. Long, H. Zhu, J. Wang, and M. I. Jordan. Unsupervised domain adaptation with residual transfer networks. In *NeurIPS*, 2016.

[52] L. V. D. Maaten and G. E. Hinton. Visualizing data using t-sne. *JMLR*, 9:2579–2605, 2008.

[53] T. Matsuura and T. Harada. Domain generalization using a mixture of multiple latent domains. In *AAAI*, 2020.

[54] M. McCloskey and N. J. Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier, 1989.

[55] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas. Communication-efficient learning of deep networks from decentralized data. In *AISTATS*, pages 1273–1282. PMLR, 2017.

[56] R. Müller, S. Kornblith, and G. E. Hinton. When does label smoothing help? In *NeurIPS*, 2019.

[57] A. T. Nguyen, P. Torr, and S.-N. Lim. Fedsr: A simple and effective domain generalization method for federated learning. In *Advances in Neural Information Processing Systems*, 2022.

[58] X. Peng, Z. Huang, Y. Zhu, and K. Saenko. Federated adversarial domain adaptation. *ICLR*, 2020.

[59] M. Pilanci and E. Vural. Domain adaptation on graphs by learning aligned graph bases. *IEEE Transactions on Knowledge and Data Engineering*, 2020.

[60] V. Piratla, P. Netrapalli, and S. Sarawagi. Efficient domain generalization via common-specific low-rank decomposition. In *ICML*, 2020.

[61] W. Qiang, J. Li, C. Zheng, B. Su, and H. Xiong. Robust local preserving and global aligning network for adversarial domain adaptation. *IEEE Transactions on Knowledge and Data Engineering*, 2021.

[62] F. Qiao, L. Zhao, and X. Peng. Learning to learn single domain generalization. In *CVPR*, pages 12556–12565, 2020.

[63] J. Quionero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence. *Dataset shift in machine learning*. The MIT Press, 2009.

[64] F. Sattler, S. Wiedemann, K.-R. Müller, and W. Samek. Robust and communication-efficient federated learning from non-iid data. *IEEE transactions on neural networks and learning systems*, 31(9):3400–3413, 2019.

[65] S. Seo, Y. Suh, D. Kim, J. Han, and B. Han. Learning to optimize domain specific normalization for domain generalization. In *ECCV*, 2020.

[66] S. Shankar, V. Piratla, S. Chakrabarti, S. Chaudhuri, P. Jyothi, and S. Sarawagi. Generalizing across domains via cross-gradient training. In *ICLR*, 2018.

[67] Y. Sun, N. Chong, and H. Ochiai. Feature distribution matching for federated domain generalization. *Proceedings of Machine Learning Research*, 189(2022), 2022.

[68] A. Z. Tan, H. Yu, L. Cui, and Q. Yang. Towards personalized federated learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[69] X. Tang, S. Guo, and J. Zhang. Exploiting personalized invariance for better out-of-distribution generalization in federated learning. *arXiv preprint arXiv:2211.11243*, 2022.

[70] Z. Tang, Y. Zhang, S. Shi, X. He, B. Han, and X. Chu. Virtual homogeneity learning: Defending against data heterogeneity in federated learning. In *International Conference on Machine Learning*, pages 21111–21132. PMLR, 2022.

[71] C. X. Tian, H. Li, Y. Wang, and S. Wang. Privacy-preserving constrained domain generalization for medical image classification. *arXiv preprint arXiv:2105.08511*, 2021.

[72] H. Venkateswara, J. Eusebio, S. Chakraborty, and S. Panchanathan. Deep hashing network for unsupervised domain adaptation. In *CVPR*, pages 5018–5027, 2017.

[73] P. Venkateswaran, V. Isahagian, V. Muthusamy, and N. Venkatasubramanian. Fedgen: Generalizable federated learning. *arXiv preprint arXiv:2211.01914*, 2022.

[74] R. Volpi, H. Namkoong, O. Sener, J. C. Duchi, V. Murino, and S. Savarese. Generalizing to unseen domains via adversarial data augmentation. In *NeurIPS*, pages 5334–5344, 2018.

[75] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang. Residual attention network for image classification. In *CVPR*, pages 3156–3164, 2017.

[76] H. Wang, H. Zhao, Y. Wang, T. Yu, J. Gu, and J. Gao. Fedkc: Federated knowledge composition for multilingual natural language understanding. In *Proceedings of the ACM Web Conference 2022*, pages 1839–1850, 2022.

[77] J. Wang, C. Lan, C. Liu, Y. Ouyang, T. Qin, W. Lu, Y. Chen, W. Zeng, and P. Yu. Generalizing to unseen domains: A survey on domain generalization. *IEEE Transactions on Knowledge and Data Engineering*, 2022.

[78] Q. Wang, H. Yin, T. Chen, J. Yu, A. Zhou, and X. Zhang. Fast-adapting and privacy-preserving federated recommender system. *The VLDB Journal*, pages 1–20, 2021.

[79] R. Wang, W. Huang, M. Shi, J. Wang, C. Shen, and Z. Zhu. Federated adversarial domain generalization network: A novel machinery fault diagnosis method with data privacy. *Knowledge-Based Systems*, 256:109880, 2022.

[80] S. Wang, L. Yu, C. Li, C.-W. Fu, and P. Heng. Learning from extrinsic and intrinsic supervisions for domain generalization. In *ECCV*, 2020.

[81] Y. Wang, H. Su, B. Zhang, and X. Hu. Interpret neural networks by identifying critical data routing paths. *CVPR*, pages 8906–8914, 2018.

[82] P. Wei, Y. Ke, and C. K. Goh. A general domain specific feature transfer framework for hybrid domain adaptation. *IEEE Transactions on Knowledge and Data Engineering*, 31(8):1440–1451, 2018.

[83] Y. Wei, L. Yang, Y. Han, and Q. Hu. Multi-source collaborative contrastive learning for decentralized domain adaptation. *IEEE Transactions on Circuits and Systems for Video Technology*, 2022.

[84] O. Weller, M. Marone, V. Braverman, D. Lawrie, and B. Van Durme. Pretrained models for multilingual federated learning. *arXiv preprint arXiv:2206.02291*, 2022.

[85] G. Wu and S. Gong. Collaborative optimization and aggregation for decentralized domain generalization and adaptation. In

Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 6484–6493, 2021.

[86] H. Wu, Y. Yan, G. Lin, M. Yang, M. K.-P. Ng, and Q. Wu. Iterative refinement for multi-source visual domain adaptation. *IEEE Transactions on Knowledge and Data Engineering*, 2020.

[87] A. Xu, W. Li, P. Guo, D. Yang, H. R. Roth, A. Hatamizadeh, C. Zhao, D. Xu, H. Huang, and Z. Xu. Closing the generalization gap of cross-silo federated medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20866–20875, 2022.

[88] Q. Xu, R. Zhang, Y. Zhang, Y. Wang, and Q. Tian. A fourier-based framework for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14383–14392, 2021.

[89] Y. Yan, H. Wu, Y. Ye, C. Bi, M. Lu, D. Liu, Q. Wu, and M. K.-P. Ng. Transferable feature selection for unsupervised domain adaptation. *IEEE Transactions on Knowledge and Data Engineering*, 2021.

[90] L. Yang, B. Tan, V. W. Zheng, K. Chen, and Q. Yang. Federated recommendation systems. *Federated Learning: Privacy and Incentive*, pages 225–239, 2020.

[91] Q. Yang, Y. Liu, T. Chen, and Y. Tong. Federated machine learning: Concept and applications. *TIST*, 10(2):1–19, 2019.

[92] T. Yoon, S. Shin, S. J. Hwang, and E. Yang. Fedmix: Approximation of mixup under mean augmented federated learning. In *ICLR*, 2021.

[93] J. Yuan, X. Ma, D. Chen, K. Kuang, F. Wu, and L. Lin. Domain-specific bias filtering for single labeled domain generalization. *arXiv preprint arXiv:2110.00726*, 2021.

[94] J. Yuan, X. Ma, K. Kuang, R. Xiong, M. Gong, and L. Lin. Learning domain-invariant relationship with instrumental variable for domain generalization. *arXiv preprint arXiv:2110.01438*, 2021.

[95] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena. Self-attention generative adversarial networks. In *ICML*, pages 7354–7363. PMLR, 2019.

[96] L. Zhang, X. Lei, Y. Shi, H. Huang, and C. Chen. Federated learning with domain generalization. *arXiv preprint arXiv:2111.10487*, 2021.

[97] L. Zhang, T. Zhu, P. Xiong, W. Zhou, and P. Yu. A robust game-theoretical federated learning framework with joint differential privacy. *IEEE Transactions on Knowledge and Data Engineering*, 2022.

[98] Q. Zhang, Y. Yang, Y. Wu, and S. Zhu. Interpreting cnns via decision trees. *CVPR*, pages 6254–6263, 2019.

[99] S. Zhao, M. Gong, T. Liu, H. Fu, and D. Tao. Domain generalization via entropy regularization. In *NeurIPS*, 2020.

[100] K. Zhou, Z. Liu, Y. Qiao, T. Xiang, and C. Change Loy. Domain generalization: A survey. *arXiv e-prints*, pages arXiv–2103, 2021.

[101] K. Zhou, Y. Yang, T. Hospedales, and T. Xiang. Deep domain-adversarial image generation for domain generalisation. In *AAAI*, 2020.

[102] K. Zhou, Y. Yang, T. Hospedales, and T. Xiang. Learning to generate novel domains for domain generalization. In *ECCV*, pages 561–578, 2020.

[103] K. Zhou, Y. Yang, Y. Qiao, and T. Xiang. Domain generalization with mixstyle. In *ICLR*, 2021.

[104] P. Zhou, K. Wang, L. Guo, S. Gong, and B. Zheng. A privacy-preserving distributed contextual federated online learning framework with big data support in social recommender systems. *IEEE Transactions on Knowledge and Data Engineering*, 33(3):824–838, 2019.

Junkun Yuan received his B.S. degree from the College of Information Engineering at Zhejiang University of Technology in 2019. He is currently pursuing the Ph.D. degree with the College of Computer Science and Technology at Zhejiang University since 2019. His research interests include domain adaptation/generalization and causal inference.

Xu Ma received his B.S. degree from the College of Computer Science and Technology at Harbin Institute of Technology in 2020. He is currently pursuing Master degree at the College of Computer Science and Technology at Zhejiang University since 2020. His research interests include domain adaptation and domain generalization.

Defang Chen received the B.S. degree in computer science from Zhejiang University of Technology, China, in 2019. He is currently pursuing the Ph.D. degree with the College of Computer Science, Zhejiang University, China. His research interests mainly include deep learning, machine learning and computer vision.

Kun Kuang is an Associate Professor of the College of Computer Science and Technology at Zhejiang University. He got his PhD in the Department of Computer Science and Technology at Tsinghua University in 2019. His research interests include causal inference and machine learning. In paricular, he is interested in promoting the convergence of causal inference and machine learning, including improving the effectiveness of causal inference with machine learning technologies, and bringing stability and interpretability of machine learning with causality.

Fei Wu is a professor of the College of Computer Science and Technology at Zhejiang University. He received B.S. degree in Lanzhou University in 1996, M.S. degree in University of Macau in 1999, and PhD in Zhejiang University in 2002. His research interests include artificial intelligence, multi-media analysis, and statistical learning theory.

Lanfen Lin is a professor of the College of Computer Science and Technology at Zhejiang University. She got her B.S. degree in Northwest University of Technology in 1990, and PhD in Northwest University of Technology in 1995. Her research interests include medical image analysis, recommender system, and data mining.