Efficiently Assemble Normalization Layers and Regularization for Federated Domain Generalization

Khiem Le[1], Long Ho[2], Cuong Do[2], Danh Le-Phuoc[3], Kok-Seng Wong[2,*]
[1] Department of Computer Science and Engineering, University of Notre Dame, IN, USA
[2] College of Engineering and Computer Science, VinUniversity, Hanoi, Vietnam
[3] Open Distributed Systems, Technical University Berlin, Berlin, Germany
kle3@nd.edu, danh.lephuoc@tu-berlin.de, {long.ht, cuong.dd, wong.ks}@vinuni.edu.vn

Abstract

*Domain shift is a formidable issue in Machine Learning that causes a model to suffer from performance degradation when tested on unseen domains. Federated Domain Generalization (FedDG) attempts to train a global model using collaborative clients in a privacy-preserving manner that can generalize well to unseen clients possibly with domain shift. However, most existing FedDG methods either cause additional privacy risks of data leakage or induce significant costs in client communication and computation, which are major concerns in the Federated Learning paradigm. To circumvent these challenges, here we introduce a novel architectural method for FedDG, namely gPerXAN [1], which relies on a normalization scheme working with a guiding regularizer. In particular, we carefully design **Per**sonalized e**X**plicitly **A**ssembled **N**ormalization to enforce client models selectively filtering domain-specific features that are biased towards local data while retaining discrimination of those features. Then, we incorporate a simple yet effective regularizer to **g**uide these models in directly capturing domain-invariant representations that the global model's classifier can leverage. Extensive experimental results on two benchmark datasets, i.e., PACS and Office-Home, and a real-world medical dataset, Camelyon17, indicate that our proposed method outperforms other existing methods in addressing this particular problem.*

1. Introduction

Over the past few decades, Machine Learning (ML) has demonstrated remarkable achievements across diverse areas such as Computer Vision, Natural Language and Speech Processing, or Robotics [5]. In general, most ML models rely on an over-simplified assumption, i.e., the training and testing data are independent and identically distributed, which does not always reflect real-world practices. In practical scenarios where the distribution of testing data diverges from that of training data, the performance of ML models often drops catastrophically due to the domain shift issue [28]. Additionally, obtaining or identifying the testing data before model deployment can be challenging in numerous applications. For instance, in biomedical applications where data characteristics vary across different equipment and institutions, gathering data from all potential domains in advance is impractical. Therefore, it is essential to have a solution that can improve the generalization capability of such ML models to adapt effectively to unseen domains.

Domain Generalization (DG) has been proposed to address the challenge of training ML models using data from single or multiple source domains with the expectation that these models will perform well on unseen domains [41]. The majority of existing DG methods fall under the category of ***domain-invariant representation learning*** approach [13, 20, 25, 29, 33]. This approach relies on a broadly acknowledged assumption that each domain contains its own domain-specific features, which are biased towards spurious relations in the data, and that all domains share domain-invariant features, which are general and robust to any unseen domains. From this assumption, previous works propose methods that remove domain-specific features and distill domain-invariant features to achieve the generalization ability. Alternative approaches for DG encompass *data augmentation* [23, 43, 44, 47], which involves exposing models to artificially generated domains, and *meta-learning* [1, 6, 12], an approach that emulates the domain shift during the meta-training phase. However, most methods require a centralized setting where all source domains are collected together. Consequently, these methods cannot be readily expanded to decentralized settings.

Federated Learning (FL) [21] is an emerging decentralized learning paradigm widely adopted in various applica-

*Corresponding author: wong.ks@vinuni.edu.vn
[1] https://github.com/lhkhiem28/gPerXAN

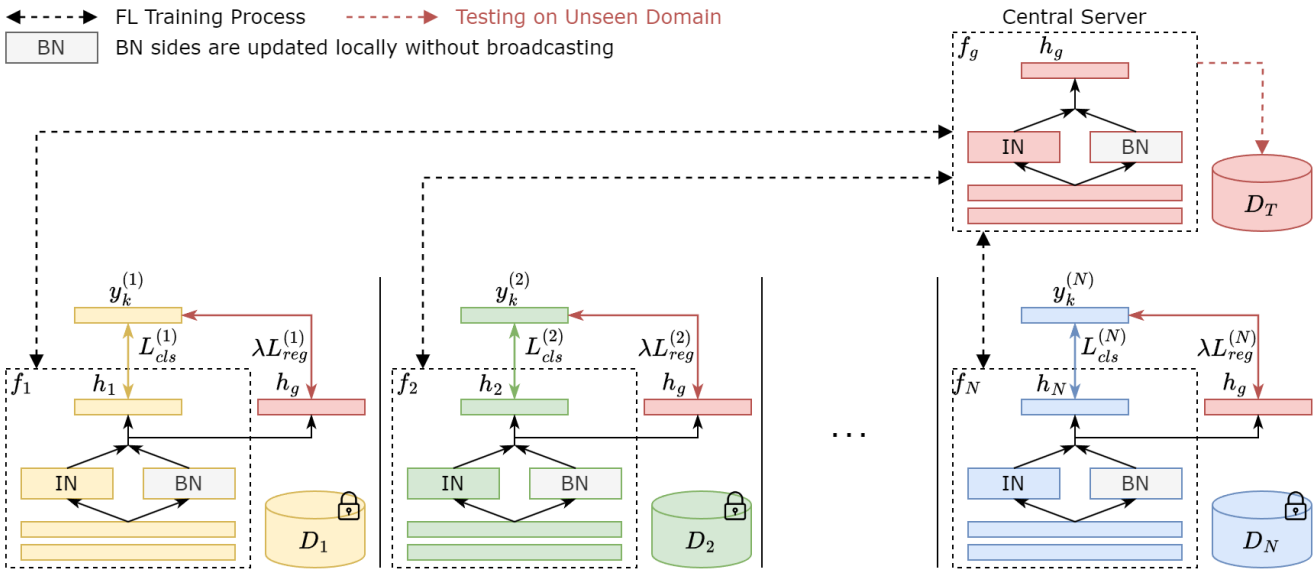


Figure 1. An overview of our proposed gPerXAN method for solving the FedDG problem.

tions to cope with the increasing privacy concerns of data centralization [40]. Specifically, the paradigm works in a way that each client learns from their data and only aggregates local models' parameters at a certain frequency at the central server to generate a global model. Notably, all data samples are kept within each client during the FL training process. Due to the nature of data decentralization, where each client owns a single source domain, as illustrated in Figure 1, the FL paradigm poses further significant challenges for DG and limits the applicability of available DG methods. There have been some early attempts to address the DG problem in the FL scenario. For instance, Liu et al. [18] introduces a method that allows clients to share their image data in the frequency space with each other, thus relatively recovering the centralization process at each client. Similarly, Chen et al. [3] introduces another method that extracts and exchanges the style of local images among all clients. It is evident that these initial efforts employ a strategy that necessitates the sharing of partial client data, thereby compromising the data privacy constraints of FL to a certain extent. Although they show promising results, these methods can be overly complicated to implement in practice and lead to additional privacy risks during the FL training process.

To address the aforementioned challenges, this paper introduces a novel architectural method for domain-invariant representation learning within the FL framework. The proposed method enhances the generalization ability while upholding the fundamental privacy principles of FL. Based on the effectiveness of discarding domain-specific information from learned features [25, 30], we properly assemble Instance Normalization layers (IN) into Batch Normalization layers (BN) in well-known Convolutional Neural Networks (CNNs) using an explicit differential mixture as in Eqn (2). Moreover, thanks to the explicit property, the benefit of personalization in FL [27, 35] can be incorporated into the normalization scheme using local BN sides. Specifically, during the FL training process, while IN sides are globally aggregated along with other model parameters, BN sides are updated locally without broadcasting. In addition, we argue that only relying on the ability to filter domain-specific features of IN while lacking guidance to distill domain-invariant representations directly might lead to suboptimal performance. Based on this observation, we introduce a simple yet highly effective regularization term to *guide* client models to directly capture domain-invariant representations that can be used by the global model's classifier, which is aggregated from client models' classifiers.

To summarize, our main contributions in this paper are highlighted as follows:

- Different from existing methods for DG in the FL scenario, we propose a novel method that concentrates on a personalized normalization scheme, global IN while local BN, for filtering domain-specific features and fully respecting the privacy-preserving principles of FL.
- Furthermore, we propose a simple yet effective regularization term to introduce clear guidance to client models for directly capturing domain-invariant representations, further improving performance on unseen domains.
- Finally, we conduct extensive experiments on two benchmark datasets, i.e., PACS and Office-Home, and a real-world medical dataset, Camelyon17, where our proposed method outperforms existing relevant ones.

2. Related Work

2.1. Domain Generalization

Domain Generalization is a challenging task requiring models to perform well on unseen domains. One dominant approach is ***domain-invariant representation learning***, which aims to minimize the discrepancy among source domains, assuming that the resulting representation will be domain-invariant and generalize well on testing domains. Along this track, Muandet et al. [22] proposes to reduce the domain dissimilarity by using Maximum Mean Discrepancy. Pan et al. [25] combines IN and BN in popular CNNs to reap the benefits of removing domain-specific features while maintaining the ability to capture discriminative features. Meanwhile, Seo et al. [30] discloses that combining these normalization layers in a switchable mechanism can yield better performance. Another approach to DG is *data augmentation*, which augments source domains to a broader span of the training data space, enlarging the possibility of covering the span of the data in the testing domain. Zhou et al. [47] mixes styles of training instances, resulting in novel domains being synthesized implicitly. Recently, the *meta-learning* approach has drawn increasing attention from the DG community. Balaji et al. [1] proposes MetaReg that learns a regularization function, particularly for the network classification layer, while excluding the feature extractor. However, the majority of these methods rely on having access to a diverse set of source domains, making them not applicable to decentralized settings.

2.2. Federated Domain Generalization

While there is a growing interest in addressing the DG problem within the FL framework, particularly in scenarios where each client possesses a single source domain, the existing literature on this subject remains relatively sparse. This FedDG problem was initially introduced by Liu et al. [18], who then proposed ELCFS that involves exchanging amplitude information in the frequency space among clients and leveraging episodic learning to further enhance performance. Chen et al. [3] proposes a similar mechanism CCST that extracts and exchanges the overall domain style of local images among all clients. Unfortunately, these early works require sharing partial information about the local data, which can be viewed as a form of data leakage and is undesirable in the FL setting. Moreover, these methods include broadcasting partial information and interpolating this information into new training data, which adds significant extra cost to communication and computation during the FL training process. Besides, Wu and Gong [42] proposes an architectural method COPA in which the global model consists of a domain-invariant representation extractor and an ensemble of domain-specific classifiers. Sun et al. [36] introduces FedKA that employs a server-side voting mechanism that generates target domain pseudo-labels based on the consensus from clients to facilitate global model fine-tuning. Recently, Zhang et al. [46] proposes a novel global objective incorporating a variance reduction regularizer to encourage fairness, and then FedDG-GA is proposed to optimize this objective by dynamically calibrating the aggregation weights. Although relaxing from the data leakage issue, these methods also cause a large additional consumption of resources when the number of source domains and output classes increases. In this paper, we introduce an alternative architectural method that shares only the model update information during training. This ensures maximal data privacy and circumvents the communication and computation overhead issue mentioned earlier while achieving competitive results. In Table 1, we highlight the advantages of our method in comparison to previous ones.

Method	Privacy Risk	Additional Cost	
		Communication	Computation
ELCFS	✓	✓	✓
CCST	✓	✓	✓
COPA	×	✓	✓
FedDG-GA	×	×	✓
gPerXAN (Ours)	×	×	×

Table 1. An advantage comparison of different methods.

One related research field to FedDG is Federated Domain Adaptation (FedDA). While both fields aim to maximize the model performance on unseen domains using existing source domains, FedDA can access the target domains while FedDG cannot see those data during training. Leveraging this assumption, FedDA methods such as FADA [26] or FMTDA [45], are typically able to align target and source features. However, this specific assumption also makes FedDA methods are inapplicable to FedDG.

2.3. Personalized Federated Learning

Another related and orthogonal line of work is Personalized Federated Learning (pFL). pFL aims to learn personalized models for different clients to tackle distribution shifts across client data, known as data heterogeneity. Li et al. [14] utilizes a regularization term in the local loss function so that the clients' trained models will not significantly differ from the global model. On the other side, Sun et al. [35] and Pillutla et al. [27] mitigate the heterogeneous distributions by only loading a subset of the global model's parameters rather than loading the entire model at each training round. Although related in terms of overcoming distribution shifts across clients, pFL focuses on improving the performance of participating clients, whereas our considered field FedDG focuses on improving the performance of unseen clients with unseen domains.

3. Methodology

Problem Formulation. First, we denote X and Y as the input space and the label space, respectively, of a specific task $\mathcal{T}$. In the standard FL setting, N clients $\{c_i\}_{i=1}^{N}$ are involved in collaboratively constructing a global model f_g for solving the task, where each client c_i owns a dataset $D_i = \{(x_k, y_k)\}_{k=1,|D_i|}$ which is associated to a specific domain defined by a joint distribution $P_{X,Y}^{(i)}$. Importantly, scattered datasets $\{D_i\}_{i=1}^{N}$ across clients satisfy:

- $P_{X,Y}^{(i)} \neq P_{X,Y}^{(j)}$ with $1 \leq i, j \leq N$ and $i \neq j$
- $P_{Y|X}^{(i)} = P_{Y|X}^{(j)}$ with $1 \leq i, j \leq N$ and $i \neq j$

The objective of FedDG is leveraging scattered datasets to construct a model f_g that can directly generalize to the unseen dataset D_U with an unseen domain, which means:

- $P_{X,Y}^{(U)} \neq P_{X,Y}^{(i)}$ with $1 \leq i \leq N$
- $P_{Y|X}^{(U)} = P_{Y|X}^{(i)}$ with $1 \leq i \leq N$

To this end, N clients communicate with a central server for T rounds. At each round, every client c_i receives the same global model f_g from the server and updates f_g with their local dataset D_i for E epochs to establish its local model f_i. The server then collects all trained models and aggregates them to update the global model. This process repeats until the global model converges. In this work, we consider the most popular FL framework, FedAvg [21], which aggregates client models as:

$$f_g = \sum_{i=1}^{N} \frac{|D_i|}{\sum_{1 \leq j \leq N} |D_j|} f_i \tag{1}$$

Challenges. In the general spirit of DG, a model is expected to extensively explore multiple source domains to achieve domain-invariance in its learned latent representation. However, under the FL setting, each client is restricted to accessing only its own local data, which constrains to make full use of source domains and, consequently, limits learning of generalizable representation. Moreover, sharing data or even partial information about data among clients poses additional privacy risks and introduces communication and computation costs during the FL training process.

3.1. Normalization Scheme

eXplicitly Assembled Normalization. The first step towards solving the highlighted challenges, we introduce a novel normalization scheme called eXplicitly Assembled Normalization (XAN), which combines IN and BN as:

$$\hat{h} = w_{in}(\gamma_{in}\frac{h - \mu_{in}}{\sqrt{\sigma_{in}^2 + \epsilon}} + \beta_{in}) + w_{bn}(\gamma_{bn}\frac{h - \mu_{bn}}{\sqrt{\sigma_{bn}^2 + \epsilon}} + \beta_{bn}), \tag{2}$$

where $h, \hat{h} \in \mathrm{R}^{B\mathrm{x}C\mathrm{x}W\mathrm{x}H}$ are layer input and output activations, which are 4D tensors with dimensions of batch size B, number of channels C, width W and height H. Meanwhile, μ and σ^2 are means and variances captured by the normalization layers, respectively, γ and β are affine parameters, and ϵ is for numerical stability.

In essence, XAN is a sophisticated mixture mechanism of IN and BN to replace BN in the feature extractors of CNNs. In particular, w_{in} and w_{bn} are ratios to weight the mixture that allow the model to switch between IN and BN. These parameters are randomly initialized and optimized along with other model parameters during training in an end-to-end manner. Unlike previous works [19, 30], which uses an implicit mechanism to combine computed statistics of IN and BN, i.e., means and variances, XAN uses an explicit mechanism that combines the output activations of two normalization layers, this provides the model with a unique ability to separate IN from BN completely. Note that IN has shown great success in style transfer tasks [9] as it allows discarding the variability of visual styles such as object colors or textures from the content of images. This property makes IN beneficial for solving the DG problem. However, directly using IN to replace the conventional BN leads to the loss of discrimination in the learned features, resulting in performance degradation in classification tasks.

Personalized eXplicitly Assembled Normalization. To accomplish our proposed normalization scheme, Personalized eXplicitly Assembled Normalization (PerXAN), we make a subtle modification to the conventional FedAvg [21] framework based on the explicit property of the above XAN. Specifically, during the FL training process, while IN sides of XAN layers are globally aggregated along with other model parameters, BN sides are updated locally, which means that parameters of BN sides are excluded from the broadcasting steps from the server. Notably, in inference time, the global model is generated by averaging all model parameters of clients. Our modification is motivated by a common observation that FedAvg normally results in poor convergence and performance in the presence of domain heterogeneity across clients [15, 48] with the major reason being that client models forget the acquired knowledge from previous rounds after aggregated [31]. The leading solution is to personalize a subset of the model, which benefits clients in learning better from their local data [27, 35]. Moreover, based on another finding that BN plays a crucial role in dealing with the domain shift issue in the centralized paradigm [16, 17], BN sides in XAN layers are finally made to be personalized. PerXAN is illustrated in Figure 1.

3.2. Regularization as Guidance

With the proposed normalization scheme PerXAN in place, now, we are ready to introduce our *"Regularization as Guidance"*. It targets to induce clear guidance to client

models through a regularizer, which brings an effect of domain alignment [33, 34], and guides these models to capture domain-invariant representations directly. We hypothesize that this will address a drawback of current DG methods that are later verified by our experiments in Section 4. The drawback is from DG methods that take advantage of the IN's function [25, 30] but do not actually equip the model with the capability of capturing domain-invariant features even if they demonstrated promising performance. Instead, it is only hoped that domain-invariant features would be distilled through achieving the goal of removing domain-specific features. This indirect learning purpose might affect the model's learning efficacy, especially in the FL setting, where each client only owns a separate single source domain. Specifically, we assume a classification model f, either the global model f_g or a client model f_i, comprises a feature extractor g and a classifier head h. At each client c_i, during local training, the client model f_i is optimized on the local dataset D_i using the following loss function:

$$L_i = L_{cls}^{(i)}(f_i; D_i) + \lambda L_{reg}^{(i)}(g_i, h_g; D_i) \quad (3)$$

$$= \sum_{k=1}^{|D_i|} \ell(f_i(x_k^{(i)}), y_k^{(i)}) + \lambda \ell(h_g(g_i(x_k^{(i)})), y_k^{(i)}), \quad (4)$$

where ℓ is the base loss function, which is usually cross-entropy in classification tasks, and λ is a hyper-parameter to control the significance of the regularizer. In a deeper look, by freezing h_g during local training, the auxiliary loss term L_{reg} forces client models to arrive at representations that can be made use for classification by the same global classifier, hence, producing an alignment effect on these representations [33, 34]. Moreover, our proposed regularizer can also be interpreted as an implicit form of matching global knowledge to clients' knowledge, which has been demonstrated to bring performance gain [3, 18] for FedDG.

Due to its orthogonality, our proposed regularizer can be easily integrated into various methods, as shown in Algorithm 1. Specifically, our experiments in Section 4 verify its compatibility with the normalization scheme that justifies our motivation. Moreover, by using only the global model's classifier to regularize feature extractors, gPerXAN saves major communication and computation resources, as well as memory usage at clients compared to others, which use an ensemble of classifiers [42] or entire global model [46].

Algorithm 1 gPerXAN

Input: A model f uses PerXAN to replace BN layers in the feature extractor. N clients with their local datasets $\{D_i\}_{i=1}^N$. Notably, $f^{(t)}$ and $f^{(l;t)}$ is the model f and its l^{th} layer at the communication round t, respectively.
Initialization: $f_g^{(0)} \leftarrow f$
for each round $t = 1, 2, 3, \ldots, T$ **do**
  for each client $i = 1, 2, 3, \ldots, N$ **in parallel do**
    for each layer in $f_i^{(t)}$ **do**
      if $f_i^{(l;t)}$ is not BN **then**
        $f_i^{(l;t)} \leftarrow f_g^{(l;t)}$
      end if
    end for
    $f_i^{(t)} \leftarrow$ **LocalTraining**$(f_i^{(t)}, h_g; D_i)$
  end for
  $f_g^{(t)} = \sum_{i=1}^N \frac{|D_i|}{\sum_{1 \le j \le N} |D_j|} f_i^{(t)}$ // Eqn (2)
end for
return: $f_g^{(T)}$
LocalTraining$(f_i, h_g; D_i)$:
for each epoch $e = 1, 2, 3, \ldots, E$ **do**
  $L_i = L_{cls}^{(i)}(f_i; D_i) + \lambda L_{reg}^{(i)}(g_i, h_g; D_i)$ // Eqn (3)
  $f_i \leftarrow f_i - \eta \nabla L_i$
end for
return: f_i

4. Experiments and Results

4.1. Datasets

To evaluate the proposed method, we perform experiments on two standard DG benchmark datasets, i.e., PACS [11] and Office-Home [39], and a real-world medical image dataset, Camelyon17 [2], consisting of various sub-datasets that are considered as domains. Specifically, PACS is composed of 4 domains with large discrepancies from diverse image colors and textures, *Photo* (P), *Art painting* (A), *Cartoon* (C), and *Sketch* (S). Each domain contains 7 categories, with 9,991 images in total. Office-Home is also composed of 4 domains but with smaller discrepancies from various backgrounds and camera viewpoints, *Product* (P), *Art* (A), *Clipart* (C), and *Real-world* (R). Each domain contains a more extensive label set of 65 categories with 15,588 images. Camelyon17 is a binary tumor classification dataset containing 455,964 histology images with stains from 5 *different hospitals* worldwide. These datasets are at different difficulty levels and are commonly used in the literature.

4.2. Experimental Settings

Evaluation. For a fair comparison purpose, we follow the common leave-one-domain-out evaluation protocol as considered in [3, 18, 46]. In particular, we sequentially choose one domain as the unseen domain, train the model on all remaining domains where a single domain is treated as a client, and evaluate the trained model on the chosen domain. For PACS and Office-Home datasets, we split 90% of the data of each source client as the training set and 10%

Method	PACS					Office-Home					
	P	A	C	S	Avg	P	A	C	R	Avg	Avg
FedAvg (Baseline)	95.21	82.23	78.20	73.56	82.30	76.53	65.97	55.40	78.01	68.98	75.64
FedAvg w/ MixStyle	95.93	85.99	80.03	75.46	84.35	75.87	62.09	57.92	77.48	68.34	76.35
FedAvg w/ RSC	95.21	83.15	78.24	74.62	82.81	75.26	62.34	50.79	77.46	66.46	74.63
ELCFS	96.23	83.94	79.27	73.30	83.19	76.83	66.32	55.63	78.12	69.23	76.21
CCST	96.65	88.33	78.20	82.90	86.52	76.61	66.35	52.39	80.01	68.84	77.68
COPA	95.62	84.80	80.28	82.86	85.89	75.82	62.27	56.04	78.72	68.21	77.05
FedDG-GA	96.80	**86.91**	81.23	82.74	86.92	77.23	65.10	**58.29**	78.80	69.86	78.42
gPerXAN (Ours)	**97.27**	86.52	**84.68**	**83.28**	**87.94**	**78.91**	**67.24**	57.75	**80.15**	**71.01**	**79.48**

Table 2. Accuracy comparison on the PACS and Office-Home datasets in the leave-one-domain-out setting.

of that as the validation set, while for unseen clients, the entire data is used for testing. For the large-scale Camelyon17 dataset, the ratios of training set and validation set at each source client are 80% and 20%, respectively, while the entire data is used for testing at unseen clients. In all experiments, we report the test accuracy on each unseen client by using the best validation model selected based on the average of accuracies on validation sets of source clients. The reported numerical values are averaged over 3 runs. Finally, we straightforwardly compare the proposed method with the vanilla FedAvg [21]. Two centralized DG methods, which are free from the requirement of data centralization, MixStyle [47] and RSC [10], are also evaluated under the integration into the FedAvg framework. Furthermore, we directly compare our method with state-of-the-art relevant ones that address the problem of DG in the FL setting as discussed in Section 2, including ELCFS [18], CCST [3], COPA [42], and FedDG-GA [46].

Implementation Details. We present architectural details and hyper-parameter values used for experiments in the paper. Following [3, 46], for PACS and Office-Home datasets, we use the image size of 224x224 pixels and ResNet-50 [7] pre-trained on ImageNet as the backbone for the feature extractor and a linear layer as the classifier. BN layers in the first four and first two blocks in feature extractors are replaced with our PerXAN, respectively. For the Camelyon17 dataset, input images are resized to 96x96 pixels, and DenseNet-121 [8] is used instead of ResNet-50, BN layers in the first dense block and the following transition block are replaced by PerXAN. Notably, parameters related to BN layers are initialized with ImageNet pre-trained weights. Client models in all experiments are optimized by an SGD optimizer with a learning rate of 2e-3 for 100 communication rounds (i.e. $T = 100$) with one local update epoch at clients (i.e. $E = 1$). During the training, simple data augmentation techniques are applied including random horizontal flipping and color jittering. The hyper-parameter of our regularization term λ is searched in the range [0, 1] with a step of 0.25. The MixStyle [47] and RSC [10] can be directly integrated into the FedAvg without further modifications. All settings of other compared methods, i.e., ELCFS [18], CCST [3], COPA [42], and FedDG-GA [46] are chosen based on corresponding papers.

Method	Camelyon17					
	H1	H2	H3	H4	H5	Avg
FedAvg (Baseline)	97.0	91.8	89.9	94.2	81.0	90.8
FedAvg w/ MixStyle	91.1	85.5	86.2	93.3	87.9	88.8
FedAvg w/ RSC	90.6	90.6	88.3	94.5	**93.3**	91.5
ELCFS	92.9	90.6	89.9	93.2	89.9	91.3
CCST	91.5	90.2	87.3	94.6	91.6	91.0
COPA	93.2	90.9	92.2	93.6	90.2	92.0
FedDG-GA	**97.2**	90.7	91.0	92.3	90.5	92.3
gPerXAN (Ours)	96.5	**92.2**	**95.1**	**94.7**	91.9	**94.1**

Table 3. Accuracy comparison on the Camelyon17 dataset.

4.3. Main Results

Table 2 presents the quantitative results for different testing domains on the PACS and Office-Home datasets. Particularly, each result column shows the test accuracy of the global model on the domain of the column name. First of all, we can empirically verify that centralized DG methods such as MixStyle [47] and RSC [10] are inconsistent under the FedAvg framework and even harmful in certain cases compared to the baseline. Since these methods are

designed for the centralized setting, they might need access to inter-domain knowledge to perform well, which is unavailable in the FL setting. In comparing state-of-the-art FedDG methods designed for the FL scenario, methods that share partial client data information, i.e., ELCFS [18] and CCST [3], show impressive results on the PACS dataset. However, they do not significantly affect the Office-Home dataset. A similar pattern is found with the architectural method COPA [42] and FedDG-GA [46]. Meanwhile, on the PACS and Office-Home datasets, our proposed method, gPerXAN achieves average accuracies across unseen clients of 87.94% and 71.01%, which are 1.02% and 1.15% better than the second-best ones, respectively. Although most of the considered methods can perform better than the baseline FedAvg in most cases, our method demonstrates a significant boost over others on both two standard benchmarks.

In addition to standard PACS and Office-Home datasets, we further evaluate our proposed method on a medical imaging dataset Camelyon17 as presented in Table 3. On this real-world benchmark, we can observe that ELCFS [18] and CCST [3] show considerably inferior performance compared to architectural ones, i.e., COPA [42] and gPerXAN. This might be due to sophisticated and sensitive features in medical images that are more challenging to extract and interpolate than conventional datasets [38], leading to inefficiency in these information-sharing-based methods. Notably, although it yields impressive performance, COPA [42] involves several other advanced techniques, such as RandAugment [4], making the performance gain hard to justify. Meanwhile, our method achieves an average accuracy across unseen clients of 94.1%, outperforming FedDG-GA [46] by approximately 2%. In general, the above experimental results demonstrate the effectiveness of the proposed method across various applications.

4.4. Ablation Studies

We conduct ablation studies on the PACS dataset to investigate the impact of each building component. Specifically, we first compare PerXAN with other centralized variants such as conventional BN, I-BN [25], and DSON [30] under the same implementation details, and then our guiding regularizer is applied to ELCFS [18] and CCST [3] to verify its compatibility with the rest of the method.

4.4.1 Impact of the Normalization Scheme

Table 4 provides a quantitative comparison among considered normalization schemes where gFedAvg represents the FedAvg framework implemented with the guiding regularizer. This means that we sequentially utilize the conventional BN, I-BN [25], and DSON [30] to replace the PerXAN scheme in our proposed method. From this table, we can observe that I-BN [25] can yield better performance than DSON [30] and the conventional BN, which are slightly comparable to each other. Meanwhile, PerXAN shows optimal performance with a significant margin compared to others and largely contributes to the whole method.

Method	PACS				
	P	A	C	S	Avg
gFedAvg w/ BN	95.85	84.39	78.04	76.42	83.68
gFedAvg w/ I-BN	94.67	82.08	80.46	80.07	84.32
gFedAvg w/ DSON	93.77	81.64	79.91	80.63	83.99
gPerXAN (Ours)	**97.27**	**86.52**	**84.68**	**83.28**	**87.94**

Table 4. Evaluation of different normalization schemes.

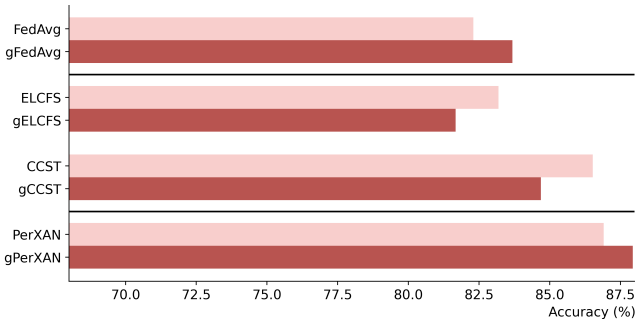


Figure 2. Contribution of the regularizer on different methods.

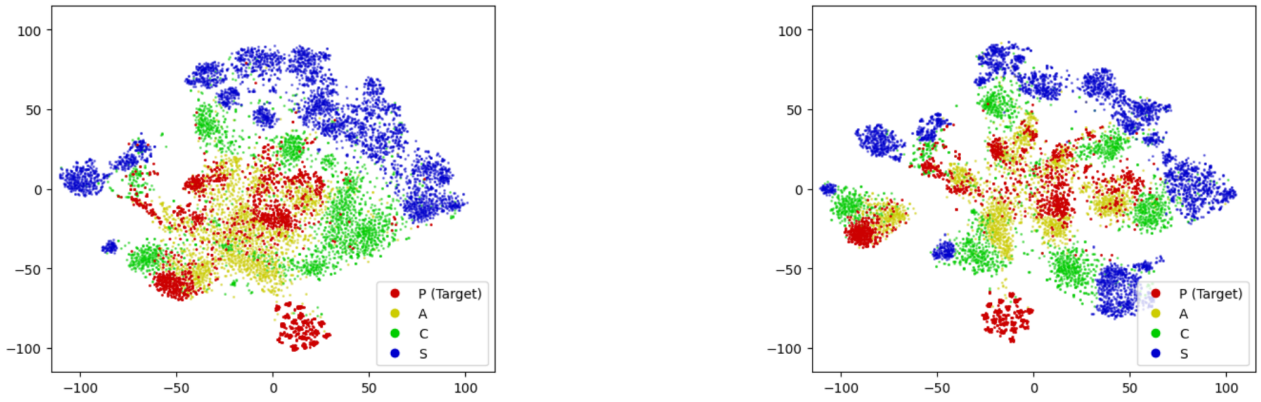
4.4.2 Impact of the Regularizer

To better understand the proposed regularization term's contribution, we sequentially employ it in FedAvg, ELCFS [18], and CCST [3] to observe performance changes. Figure 2 displays averaged accuracies across unseen clients of these methods without and with the involvement of our regularizer. It is straightforward to see that while regularizer improves the performance of FedAvg and PerXAN considerably, it does not enhance ELCFS [18] and CCST [3]. In information-sharing-based methods, i.e., ELCFS [18] and CCST [3], each client is exposed to other clients' data information, which means that clients can access global knowledge relatively, then making the effect of matching global knowledge to clients' knowledge redundant and harmful. Also, this ablation study empirically verifies the connection between the regularizer and normalization scheme.

5. Analysis and Discussion

5.1. Visualization

We now provide an in-depth ability analysis of the proposed method via representation visualization using t-SNE [37], a common dimensionality reduction technique. As shown in



(a) FedAvg (b) gPerXAN

Figure 3. Visualization of representation extracted from the global model on the PACS dataset.

Figure 3, we compare the representations extracted from the global model obtained from our gPerXAN and the baseline FedAvg on the PACS dataset with the testing domain, a.k.a. target domain *Photo* (P). From this figure, we can observe that features derived from our method are semantically separated according to 7 categories in the PACS dataset for both source domains and the target domain. Moreover, features of each category on all domains tend to be close and grouped, demonstrating the ability of our method to distill more discriminative and domain-invariant representation, leading to a significant improvement of more than 2% in classification accuracy on the testing domain *Photo* (P).

5.2. Privacy and Efficiency

Data privacy is a major concern in the field of FL, which is mitigated by only sharing client models' parameters instead of raw data. However, to resolve the FedDG problem, recently introduced methods sacrifice this principle by revealing partial information about client data. Specifically, leaking more information not only opens higher chances for attackers to perform an inversion attack [24], which aims to reconstruct the original data of clients, but also amplifies the risk of membership inference attack [32], which determines if a sample was in the model's training data. Furthermore, efficiency in communication and computation is critical in FL, especially in scenarios where resource-constrained clients are involved. Despite that, the common sharing mechanism in available methods introduces unacceptable extra costs to the FL training process. For COPA [42], using an ensemble of domain-specific classifiers in the model architecture results in an $O(N^2)$ increase in communication and computation complexities compared to $O(N)$ as conventional, according to the number of participating clients N. Meanwhile, FedDG-GA [46] consumes a double of memory at clients. Compared with existing ones, in addition to bypassing the limitations described above, our proposed method is more practical in terms of implementation yet provides competitive results in extensive evaluations.

6. Conclusion

In this paper, we introduce a novel architectural method, namely gPerXAN, to address the problem of FedDG. By explicitly assembling Instance Normalization layers into Batch Normalization layers in a personalized scheme and employing a simple yet effective guiding regularizer, our method allows the model to filter domain-specific features and actively distill domain-invariant representation for classification tasks. We conduct extensive experiments and in-depth analysis to quantitatively and qualitatively verify the effectiveness of our proposed method in solving this particular problem. Although our evaluation is more on cross-silo FL, our method can be easily extended to cross-device scenarios while Algorithm 1 remains unchanged. Unlike existing methods, due to the independence from imaging techniques, gPerXAN can be straightforwardly extended to diverse applications. A potential limitation of our method is its certain suitability with normalization-based models. In the future, investigating other forms of regularization terms is a promising research direction due to the available room for improvements. Moreover, the vulnerability of available methods under various attacks remains underexplored.

Acknowledgement

This work is supported by the Deutsche Forschungsgemeinschaft, German Research Foundation under grant number 453130567 (COSMO), by the Horizon Europe Research and Innovation Actions under grant number 101092908 (SmartEdge), and by the Federal Ministry for Education and Research, Germany under grant number 01IS18037A (BIFOLD).

References

[1] Yogesh Balaji, Swami Sankaranarayanan, and Rama Chellappa. MetaReg: Towards Domain Generalization Using Meta-Regularization. *Advances in Neural Information Processing Systems*, 31, 2018. 1, 3

[2] Peter Bandi, Oscar Geessink, Quirine Manson, Marcory Van Dijk, Maschenka Balkenhol, Meyke Hermsen, Babak Ehteshami Bejnordi, Byungjae Lee, Kyunghyun Paeng, Aoxiao Zhong, et al. From Detection of Individual Metastases to Classification of Lymph Node Status at the Patient Level: The Camelyon17 Challenge. *IEEE Transactions on Medical Imaging*, 38(2):550–560, 2018. 5

[3] Junming Chen, Meirui Jiang, Qi Dou, and Qifeng 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. 2, 3, 5, 6, 7

[4] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. RandAugment: Practical Automated Data Augmentation With a Reduced Search Space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 702–703, 2020. 7

[5] Shi Dong, Ping Wang, and Khushnood Abbas. A Survey on Deep Learning and Its Applications. *Computer Science Review*, 40:100379, 2021. 1

[6] Qi Dou, Daniel Coelho de Castro, Konstantinos Kamnitsas, and Ben Glocker. Domain Generalization via Model-Agnostic Learning of Semantic Features. *Advances in Neural Information Processing Systems*, 32, 2019. 1

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 6

[8] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely Connected Convolutional Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4700–4708, 2017. 6

[9] Xun Huang and Serge Belongie. Arbitrary Style Transfer in Real-Time With Adaptive Instance Normalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1501–1510, 2017. 4

[10] Zeyi Huang, Haohan Wang, Eric P Xing, and Dong Huang. Self-Challenging Improves Cross-Domain Generalization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 124–140. Springer, 2020. 6

[11] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, Broader and Artier Domain Generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5542–5550, 2017. 5

[12] Da Li, Jianshu Zhang, Yongxin Yang, Cong Liu, Yi-Zhe Song, and Timothy M Hospedales. Episodic Training for Domain Generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1446–1455, 2019. 1

[13] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain Generalization With Adversarial Feature Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5400–5409, 2018. 1

[14] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated Optimization in Heterogeneous Networks. *Proceedings of Machine Learning and Systems*, 2:429–450, 2020. 3

[15] Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the Convergence of Fedavg on Non-IID Data. In *International Conference on Learning Representations*, 2019. 4

[16] Yanghao Li, Naiyan Wang, Jianping Shi, Jiaying Liu, and Xiaodi Hou. Revisiting Batch Normalization for Practical Domain Adaptation. *arXiv preprint arXiv:1603.04779*, 2016. 4

[17] Quande Liu, Qi Dou, Lequan Yu, and Pheng Ann Heng. MS-Net: Multi-Site Network for Improving Prostate Segmentation With Heterogeneous MRI Data. *IEEE Transactions on Medical Imaging*, 39(9):2713–2724, 2020. 4

[18] Quande Liu, Cheng Chen, Jing Qin, Qi Dou, and Pheng-Ann Heng. FedDG: Federated Domain Generalization on Medical Image Segmenta- tion via Episodic Learning in Continuous Frequency Space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1013–1023, 2021. 2, 3, 5, 6, 7

[19] Ping Luo, Jiamin Ren, Zhanglin Peng, Ruimao Zhang, and Jingyu Li. Differentiable Learning-To-Normalize via Switchable Normalization. In *International Conference on Learning Representations*, 2018. 4

[20] Divyat Mahajan, Shruti Tople, and Amit Sharma. Domain Generalization Using Causal Matching. In *International Conference on Machine Learning*, pages 7313–7324. PMLR, 2021. 1

[21] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-Efficient Learning of Deep Networks From Decentralized Data. In *Artificial Intelligence and Statistics*, pages 1273–1282. PMLR, 2017. 1, 4, 6

[22] Krikamol Muandet, David Balduzzi, and Bernhard Scholkopf. Domain Generalization via Invariant Feature Representation. In *International Conference on Machine Learning*, pages 10–18. PMLR, 2013. 3

[23] Hyeonseob Nam, HyunJae Lee, Jongchan Park, Wonjun Yoon, and Donggeun Yoo. Reducing Domain Gap by Reducing Style Bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8690–8699, 2021. 1

[24] Ngoc-Bao Nguyen, Keshigeyan Chandrasegaran, Milad Abdollahzadeh, and Ngai-Man Cheung. Re-thinking Model Inversion Attacks Against Deep Neural Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 16384–16393, 2023. 8

[25] Xingang Pan, Ping Luo, Jianping Shi, and Xiaoou Tang. Two at Once: Enhancing Learning and Generalization Capacities via IBN-Net. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 464–479, 2018. 1, 2, 3, 5, 7

[26] Xingchao Peng, Zijun Huang, Yizhe Zhu, and Kate Saenko. Federated Adversarial Domain Adaptation. In *International Conference on Learning Representations*, 2020. 3

[27] Krishna Pillutla, Kshitiz Malik, Abdel-Rahman Mohamed, Mike Rabbat, Maziar Sanjabi, and Lin Xiao. Federated Learning With Partial Model Personalization. In *International Conference on Machine Learning*, pages 17716–17758. PMLR, 2022. 2, 3, 4

[28] Joaquin Quinonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. *Dataset Shift in Machine Learning*. Mit Press, 2008. 1

[29] Alexandre Rame, Corentin Dancette, and Matthieu Cord. Fishr: Invariant Gradient Variances for Out-Of-Distribution Generalization. In *International Conference on Machine Learning*, pages 18347–18377. PMLR, 2022. 1

[30] Seonguk Seo, Yumin Suh, Dongwan Kim, Geeho Kim, Jongwoo Han, and Bohyung Han. Learning to Optimize Domain Specific Normalization for Domain Generalization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 68–83, 2020. 2, 3, 4, 5, 7

[31] Neta Shoham, Tomer Avidor, Aviv Keren, Nadav Israel, Daniel Benditkis, Liron Mor-Yosef, and Itai Zeitak. Overcoming Forgetting in Federated Learning on Non-IID Data. *arXiv preprint arXiv:1910.07796*, 2019. 4

[32] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 3–18. IEEE, 2017. 8

[33] Baochen Sun and Kate Saenko. Deep Coral: Correlation Alignment for Deep Domain Adaptation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 443–450, 2016. 1, 5

[34] Baochen Sun, Jiashi Feng, and Kate Saenko. Correlation Alignment for Unsupervised Domain Adaptation. *Domain Adaptation in Computer Vision Applications*, pages 153–171, 2017. 5

[35] Benyuan Sun, Hongxing Huo, Yi Yang, and Bo Bai. PartialFed: Cross-Domain Personalized Federated Learning via Partial Initialization. *Advances in Neural Information Processing Systems*, 34:23309–23320, 2021. 2, 3, 4

[36] Yuwei Sun, Ng Chong, and Hideya Ochiai. Feature Distribution Matching for Federated Domain Generalization. In *Asian Conference on Machine Learning*, pages 942–957. PMLR, 2023. 3

[37] Laurens Van der Maaten and Geoffrey Hinton. Visualizing Data Using t-SNE. *Journal of Machine Learning Research*, 9(11), 2008. 7

[38] Gaël Varoquaux and Veronika Cheplygina. Machine Learning for Medical Imaging: Methodological Failures and Recommendations for the Future. *Npj Digital Medicine*, 5(1):48, 2022. 7

[39] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep Hashing Network for Unsupervised Domain Adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5018–5027, 2017. 5

[40] Paul Voigt and Axel Von dem Bussche. The EU General Data Protection Regulation (GDPR). *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, 10(3152676):10–5555, 2017. 2

[41] Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and Philip Yu. Generalizing to Unseen Domains: A Survey on Domain Generalization. *IEEE Transactions on Knowledge and Data Engineering*, 2022. 1

[42] Guile Wu and Shaogang 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. 3, 5, 6, 7, 8

[43] Zhenlin Xu, Deyi Liu, Junlin Yang, Colin Raffel, and Marc Niethammer. Robust and Generalizable Visual Representation Learning via Random Convolutions. In *International Conference on Learning Representations*, 2020. 1

[44] Fu-En Yang, Yuan-Chia Cheng, Zu-Yun Shiau, and Yu-Chiang Frank Wang. Adversarial Teacher-Student Representation Learning for Domain Generalization. *Advances in Neural Information Processing Systems*, 34:19448–19460, 2021. 1

[45] Chun-Han Yao, Boqing Gong, Hang Qi, Yin Cui, Yukun Zhu, and Ming-Hsuan Yang. Federated Multi-Target Domain Adaptation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1424–1433, 2022. 3

[46] Ruipeng Zhang, Qinwei Xu, Jiangchao Yao, Ya Zhang, Qi Tian, and Yanfeng Wang. Federated Domain Generalization with Generalization Adjustment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3954–3963, 2023. 3, 5, 6, 7, 8

[47] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain Generalization With Mixstyle. In *International Conference on Learning Representations*, 2020. 1, 3, 6

[48] Weiming Zhuang, AI Sony, and Lingjuan Lyu. FedWon: Triumphing Multi-domain Federated Learning Without Normalization. In *International Conference on Learning Representations*, 2024. 4