

Federated Domain Generalization: A Survey

Ying Li, Xingwei Wang, Rongfei Zeng, Praveen Kumar Donta, *Senior Member, IEEE*,
Iilir Murturi, *Member, IEEE*, Min Huang, and Schahram Dustdar, *Fellow, IEEE*

Abstract—Machine learning typically relies on the assumption that training and testing distributions are identical and that data is centrally stored for training and testing. However, in real-world scenarios, distributions may differ significantly and data is often distributed across different devices, organizations, or edge nodes. Consequently, it is imperative to develop models that can effectively generalize to unseen distributions where data is distributed across different domains. In response to this challenge, there has been a surge of interest in federated domain generalization (FDG) in recent years. FDG combines the strengths of federated learning (FL) and domain generalization (DG) techniques to enable multiple source domains to collaboratively learn a model capable of directly generalizing to unseen domains while preserving data privacy. However, generalizing the federated model under domain shifts is a technically challenging problem that has received scant attention in the research area so far. This paper presents the first survey of recent advances in this area. Initially, we discuss the development process from traditional machine learning to domain adaptation and domain generalization, leading to FDG as well as provide the corresponding formal definition. Then, we categorize recent methodologies into four classes: federated domain alignment, data manipulation, learning strategies, and aggregation optimization, and present suitable algorithms in detail for each category. Next, we introduce commonly used datasets, applications, evaluations, and benchmarks. Finally, we conclude this survey by providing some potential research topics for the future.

Index Terms—Domain shift, domain generalization, privacy preserving, federated domain generalization, machine learning.

I. INTRODUCTION

HUMANS possess an extraordinary aptitude to transfer their knowledge and expertise to novel situations and environments they have not experienced before, while machines struggle to reproduce this proficiency, especially regarding *out-of-distribution* (OOD) data [1]. This raises questions such as whether a machine learning (ML) model trained to detect tumors in a specific type of medical image would be effective when tested on images from a different scanner or imaging protocol or whether a robot trained to perform a task in a simulated environment may be able to generalize to the real

world. Additionally, there is a concern about whether a speech recognition system trained on audio recordings from one speaker would perform well when tested on recordings from a different speaker or whether a language model trained on text from one domain (e.g., news articles) would be applicable to text from a different domain (e.g., social media). The answers to all these questions depend on how well the ML models can deal with the common problem of domain/dataset shift [2]. Domain shift refers to the discrepancy in distribution between the source dataset used for training and the target dataset used for testing [3]–[5]. It is a pervasive issue in ML, which poses significant challenges for the generalization and robustness of models to unseen domains in real-world scenarios.

Recently, deep learning has made revolutionary advances and emerged as a powerful tool in a variety of fields, including image recognition, natural language processing, speech recognition, and robotics, among others. However, most deep-learning algorithms are under the *independent and identically distributed* (IID) assumption that all the samples in the training and test datasets are independent of each other and identically distributed. Due to the OOD scenarios commonly encountered in many real-world applications, the model trained from some source domains may result in a significant performance drop on other unseen domains, leading to the phenomenon known as domain shift. To address the domain shift issue, *domain generalization* (DG) [1], [6], [7] technique is introduced for improving the ability of models to generalize across different domains, even if these domains are significantly different from the training domain.

There has been significant research in the area of DG, and numerous methods have been proposed to address the domain shift issue. Most existing DG methods are designed to achieve centralized domain generalization, where a central server with access to all source domain data is responsible for the DG task. For example, some domain alignment methods [8]–[14] aim to align the feature distributions of the source and target domains, some data augmentation methods [15]–[21] are used to increase the size and diversity of the training data without collecting new data, some meta-learning methods [22]–[28] aim to improve the generalization performance of a model by learning a good initialization or optimization strategy that can be adapted to new tasks or domains with minimal additional training, and some ensemble learning methods [29]–[35] combine multiple models trained on different domains, and then combine their predictions using an ensemble method to capture the common patterns across the different domains and improve the generalization performance of the model on unseen domains. However, these methods can seriously violate data protection regulations, such as the EU/UK *General Data Protection Regulation* (GDPR) [36], as tremendous privacy-

Ying Li is with the College of Computer Science and Engineering, Northeastern University, Shenyang 110819, China, and Distributed Systems Group, TU Wien, Vienna 1040, Austria. Email: liying1771@163.com.

Xingwei Wang is with the College of Computer Science and Engineering, Northeastern University, Shenyang 110819, China. Email: wangxw@mail.neu.edu.cn.

Rongfei Zeng is with the College of Software, Northeastern University, Shenyang 110819, China. Email: zengrf@swc.neu.edu.cn.

Min Huang is with the College of Information Science and Engineering, Northeastern University, Shenyang 110819, China. Email: mhuang@mail.neu.edu.cn.

P. K. Donta, I. Murturi, and S. Dustdar are with Distributed Systems Group, TU Wien, Vienna 1040, Austria. Email: {pdonta, imurturi, dustdar}@dsg.tuwien.ac.at.

*Corresponding author: Xingwei Wang.

TABLE I
COMPARISONS OF RELATIVE SURVEYS.

Ref.	Year	Domain	Scope
<i>Bonawitz et al.</i> [47]	2019	Federated Learning	A scalable production system for FL in the domain of mobile devices.
<i>Yang et al.</i> [44]	2019	Federated Learning	A comprehensive secure federated learning framework that includes horizontal, vertical, and transfer learning.
<i>Li et al.</i> [46]	2020	Federated Learning	Unique characteristics, challenges, current approaches, and future research areas.
<i>Kairouz et al.</i> [45]	2021	Federated Learning	Recent advancements, the challenges and open problems.
<i>Csurka et al.</i> [48]	2017	Domain Adaptation (DA)	An overview of DA and transfer learning in visual applications
<i>Kuw et al.</i> [49]	2019	Domain Adaptation	Sample-based, feature-based, and inference-based methods.
<i>Wilson et al.</i> [50]	2020	Domain Adaptation	A unsupervised domain adaptation methods in deep learning.
<i>Wang et al.</i> [51]	2018	Domain Generalization	Data manipulation, representation learning, and learning strategy.
<i>Zhou et al.</i> [1]	2022	Domain Generalization	Background, existing methods and theories, and insights and discussions on future research directions
<i>Sheth et al.</i> [52]	2022	Domain Generalization	Invariance via Causal Data Augmentation, Invariance via Causal representation learning, and Invariance via Transferring Causal mechanisms.
<i>Shen et al.</i> [53]	2021	Out-of-Distribution Generalization	Definition, methodology, evaluation, and implications of Out-of-Distribution generalization.
<i>Li et al.</i> [54]	2022	Out-of-Distribution Generalization	Existing methods, theories, and future research directions related to out-of-Distribution generalization.
Our work	2023	Federated Domain Generalization	Theories, methodologies, datasets, applications, evaluations, benchmarks, and future research directions.

containing data is stored in locally distributed locations due to the rise of ML and the proliferation of *Internet of Things* (IoT) devices [37]. For instance, public transportation, ride-sharing services, and logistics companies data of the transportation industries; banks, insurance companies, and credit rating agencies data of financial institutions; patient privacy data of healthcare institutions; schools, universities, and online learning platforms data of educational institutions. Therefore, developing a highly generalizable model in real-world scenarios that can ensure both privacy and high model performance is a dilemma [38]. On one hand, it requires careful consideration of privacy concerns across multiple domains. On the other hand, without simultaneous access to source domains, it would be difficult to accurately identify and learn domain-invariant features for improving model generalization.

The above dilemma can be addressed by the *federated domain generalization* (FDG) [38]–[41] technique, which leverages the power of *federated learning* (FL) [42]–[47] and DG techniques to enable multiple source domains to collaboratively learn a model such that it can directly generalize well to unseen domains while keeping their data privacy. However, generalizing the federated model under domain shifts is a technically challenging problem that has received little attention in the research area so far. Moreover, DG is an active research area with different methods being proposed recently, but the federated paradigm with distributed data domains has introduced new complexities to DG. As stated above, most conventional solutions achieve domain generalization by accessing multi-source domains in a centralized manner, while each client can only access its local data in FDG. In addition, local optimization of the FL model may result in a bias towards its own data distribution, which reduces its generalizability to new target domains. Therefore, the previous DG methods can not be applied to the FDG scenario.

This paper presents the first comprehensive survey on the topic of FDG, aiming to introduce recent advances in this field

with a special focus on its theories, methodologies, datasets, applications, evaluations, benchmarks, and future research directions. To our knowledge, there have been few recent attempts [1], [44]–[54] to explore this area, but their focus differs significantly from the scope of our survey. Table I provides a summary of the comparisons between these works. Specifically, Bonawitz *et al.* [47] present a system design for FL on mobile devices using TensorFlow and discuss the challenges and open problems in the domain of mobile phones. Yang *et al.* [44] propose a secure FL framework as a solution to the challenges faced by AI in industries where data exists in isolated islands and data privacy and security are strengthened. The framework includes horizontal federated learning, vertical federated learning, and federated transfer learning, and can be applied to various businesses successfully. Li *et al.* [46] discuss the concept of FL, which involves training statistical models over remote devices or siloed data centers while keeping data localized. The authors outline the unique challenges of federated learning and provide an overview of current approaches and future directions for research. Kairouz *et al.* [45] highlight recent advances, challenges, and open problems in FL research. Csurka *et al.* [48] provide an overview of DA and transfer learning in visual applications. They discuss the state-of-the-art methods for different scenarios, including shallow and deep methods, and relate DA to other ML solutions. Kouw *et al.* [49] review different approaches to solve the problem of DA in ML and the approaches are categorized into sample-based, feature-based, and inference-based methods. Wilson *et al.* [50] conducts a comprehensive survey that explores the methodologies employed in single-source unsupervised deep domain adaptation (DA), wherein deep learning techniques are combined with DA to alleviate the dependence on potentially expensive target data annotations. They undertake a meticulous comparative analysis of approaches in single-source unsupervised deep domain adaptation, encompassing the examination of alternative methods, identification of unique and common

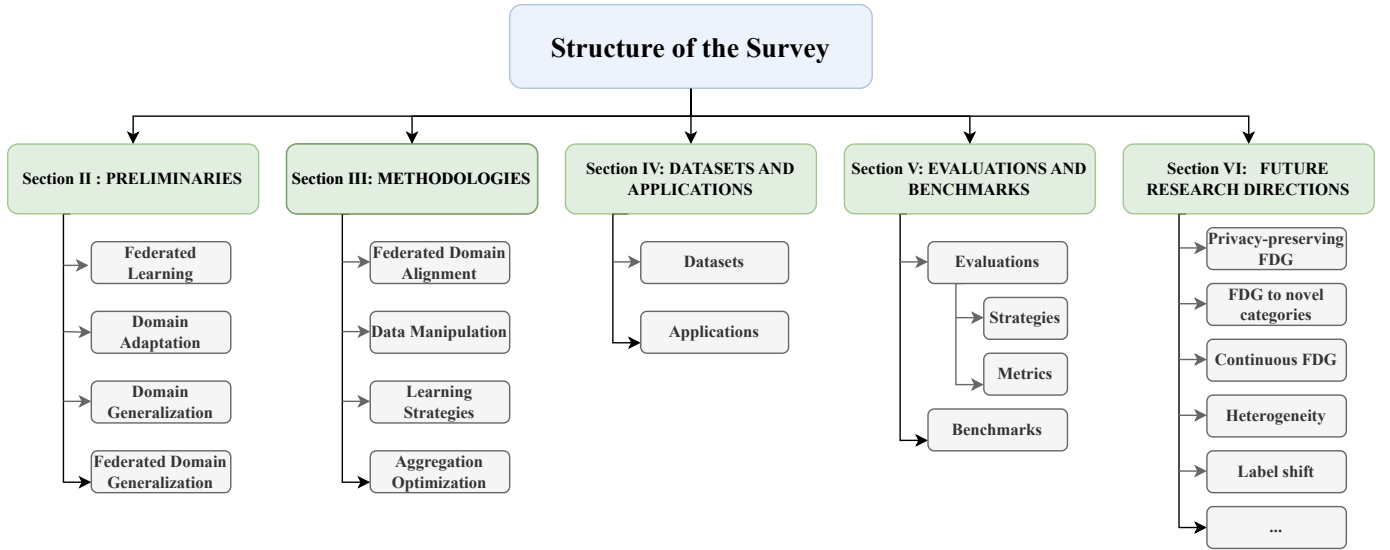


Fig. 1. The Structure of our proposed survey.

elements, assessment of results and theoretical insights, and an exploration of application areas and avenues for future research. Wang *et al.* [51] review recent advances in domain generalization, which deals with the challenge of developing models that can generalize to unseen test domains. The authors also discuss the theories, algorithms, datasets, applications, and potential research topics related to domain generalization. Zhou *et al.* [1] provide a comprehensive literature review of DG in ML, in which the review covers the background, existing methods and theories, and potential future research directions. Sheth *et al.* [52] discuss the problem of distribution shifts in ML models and how it affects their generalization capabilities. They survey various domain generalization methods that leverage causality to identify stable features or mechanisms that remain invariant across different distributions. Shen *et al.* [53] discuss the OOD generalization problem in ML and provide a comprehensive review of existing methods, evaluation metrics, and future directions for OOD generalization. Li *et al.* [54] review recent advances in OOD generalization on graphs, which goes beyond the in-distribution hypothesis, and presents a detailed categorization of existing methods into three classes based on their positions in the graph ML pipeline. With the rapid development of ML, there has been significant progress in federated learning and domain generalization in various fields in recent years. Despite notable advancements, a comprehensive survey encompassing the introduction and summarization of theories, methodologies, datasets, applications, evaluations, benchmarks, and future research directions in the field of FDG has yet to be undertaken. Our goal is to provide interested researchers with a comprehensive review of the current state of the field and inspire further research in this and related areas. To achieve this, we have conducted an extensive review of the existing literature and synthesized the main ideas and contributions of the most relevant works. We hope that this survey will serve as a valuable resource for researchers seeking to deepen their understanding of FDG and

contribute to the advancement of this important field.

The structure of this paper is as follows, as shown in Fig. 1. Frequently used acronyms are summarized using table II. We begin by providing an introduction to the field of federated learning, traditional ML, DA, DG, and FDG in Section II. Then, we introduce the existing methodologies for FDG, including federated domain alignment, data manipulation, learning strategies, and aggregation optimization algorithms in Section III. Next, we present the datasets and applications in Section IV and evaluations and benchmarks in Section V. Furthermore, in Section VI, we provide a concise summary of the insights derived from existing research and engage in a discussion concerning the future research directions in the field. Finally, a comprehensive conclusion of this paper is presented in Section VII, summarizing the key findings and insights derived throughout the study. On the analysis of the sections, we derive a generalized architecture and workflow of FDG, as depicted in Fig. 6.

II. PRELIMINARIES

In this section, we first present an overview of federated learning, DA, domain generalization, and FDG. Then, we introduce the integration of federated learning and domain generalization. For readers' convenience, frequently used notations are summarized using Table III.

A. Federated Learning

Artificial intelligence has rapidly become an indispensable aspect of our daily lives, fueled by notable advancements and breakthroughs in deep learning techniques across various domains. Moreover, the widespread adoption of IoT devices has resulted in enormous data being stored in locally distributed locations. In addition, with the proliferation of sensitive and private data, concerns over data privacy have become increasingly prominent. In response to these concerns, the concept of federated learning (FL) [42] has been proposed

TABLE II
FREQUENTLY USED ACRONYMS

Acronym	Meaning	Acronym	Meaning
ML	Machine Learning	FL	Federated Learning
DA	Domain Adaptation	FDA	Federated Domain Adaptation
DG	Domain Generalization	FDG	Federated Domain Generalization
DCCS	Distributed Computing Continuum Systems	MMD	Maximum Mean Discrepancy
DIR	Domain-Invariant Representation	COPA	Collaborative Optimization and Aggregation
CSAC	Collaborative Semantic Aggregation and Calibration	DANN	Domain Adversarial Neural Networks
DFKD	Data-free Knowledge Distillation	HD	Hausdorff Distance
UDA	Unsupervised Domain Adaptation	VAEs	Variational Autoencoders
VC	Vapnik-Chervonenkis	WD	Wasserstein Distance
SGD	Stochastic Gradient Descent	NLP	Natural language processing
MTSSL	Multi-Task Self-Supervised Learning	IID	Independent and Identically Distributed
FADH	Federated Adversarial Domain Hallucination	IoT	Internet of Things
OOD	Out of Distribution	RKHS	Reproducing Kernel Hilbert Space
RSC	Representation Self-Challenging	FID	Fréchet Inception Distance
LFRL	Lifelong Federated Reinforcement Learning	FZSL	Federated Zero-Shot Learning
GAN	Generative Adversarial Networks	GDPR	General Data Protection Regulation
GE	Group Effect	SD	Standard Deviation
IRM	Invariant Risk Minimization	KL	Kullback-Leibler Divergence
KD	Knowledge Distillation	MAML	Model-Agnostic Meta-Learning
MCKD	Mutually Collaborative Knowledge Distillation	mIoU	mean Intersection over Union
MoMP	Mixture of Motor Primitives	CCST	Cross-client Style Transfer
DRNN	Distributionally Robust Neural Networks	DTL	Distributed Ledger Technologies

TABLE III
FREQUENTLY USED ACRONYMS

Notation	Meaning
$\mathcal{N}$	The set of clients
N	The number of clients in FL.
$\mathcal{M}$	The set of end users in FL.
M	The number of end users in FL.
$\mathbb{S}_i$	The dataset of client i
$ \mathbb{S}_i $	The size of local dataset of client i
$\mathcal{H}$	Hypothesis space.
h	Labeling function from a $\mathcal{H}$.
$\mathcal{X}$	Non-empty input space.
$\mathcal{Y}$	Output space.
P_{XY}	Probability distribution of the joint input-output space.
$\mathcal{S}$	Domain.
M	The number of Domains.
$\mathcal{S}_{source}$	The set of the source domains.
$\mathcal{S}_{target}$	The set of the target domains.
M	The number of source domains.
n	The size of the domain.
$\mathcal{S}^i$	Domain i .
n_i	The size of the domain i .
$\mathcal{X}_i$	The selected subset from the local dataset $\mathbb{S}_i$ of client i for local model training.
$ \mathcal{X}_i $	The size of the chosen subset $\mathcal{X}_i$ from client i .
ω_t	Aggregated global model at round t .
ω_{t+1}^i	Local model from client i at round $t+1$.
ω^*	The optimal weights of the global model.
η	The learning rate of FL training.
$\theta(h, \hat{h})$	The difference in the source domain between two labeling function h and $\hat{h}$.
$\theta^s(h)$	The risk of labeling function h on the source domains.
$\theta^t(h)$	The risk of labeling function h on the target domains.
$\lambda_{\mathcal{H}}$	Complexity of $\mathcal{H}$ for prediction tasks.
d	Vapnik-Chervonenkis dimension of $\mathcal{H}$.
$\mathbb{E}$	Expectation operator.
$\mathcal{U}^s$	Unlabeled samples of size k from source domain.
$\mathcal{U}^t$	Unlabeled samples of size k from target domain.
$\hat{\psi}(h)$	The estimation of the average risk over all possible target domains.
π_i	π A normalized mixing weights.
$F^t((x_i^t))$	The feature representation of the target domain data.
$MMD^2(P, Q)$	The square of maximum mean discrepancy.

as a solution to address data privacy concerns while facilitating collaborative ML.

As illustrated in Fig. 2, we depict a scenario consisting of a model engineer, an admin, a central server, N clients, and M end users. The set of clients is denoted by $\mathcal{N} = \{1, 2, \dots, i, \dots, N\}$, while the set of end users is denoted by $\mathcal{M} = \{1, 2, \dots, j, \dots, M\}$. Each client $i \in \mathcal{N}$ has a local dataset $\mathbb{S}^i$, which may change over time. The size of the local dataset is denoted by $\mathbb{S}^i$ with the size of $|\mathbb{S}^i|$. Thus, clients

can iteratively perform FL processes due to the time-varying nature of their datasets. Each client can select a subset $\mathcal{X}^i \subseteq \mathbb{S}^i$ of its local dataset for local model training. The size of the chosen subset is denoted by $|\mathcal{X}^i|$. Next, we will introduce the lifecycle, key actors, and typical processes of a trained model in the FL system respectively.

1) The Lifecycle of a Trained Model in FL System.

The FL process is typically driven by a model engineer who develops a model for a particular application. Here's a high-level overview of a typical FL workflow:

Step 1: Problem Definition. The first step in the FL workflow is to define the problem that the model will solve. This involves identifying the data sources and the target task, as well as specifying the performance metrics that will be used to evaluate the model.

Step 2: Model Development. Once the problem is defined, the model engineer develops a model architecture that is suitable for the FL paradigm. This involves selecting the appropriate optimization algorithm, defining the model parameters, and deciding how to partition the data among the participating clients.

Step 3: Dataset Preparation. The dataset is prepared by the data owners, who collect or generate the data that will be used for training the model. The data is partitioned among the clients according to the protocol defined in Step 2.

Step 4: Training Initialization. The central server initializes the training process by sending the initial model weights to the participating clients. This model can be randomly initialized or pre-trained on a large dataset.

Step 5: Local Training. The participating clients perform local training using their own data and the initial model weights. This training is typically done using stochastic gradient descent (SGD) or a variant of it.

Step 6: Model Aggregation. After local training, the clients

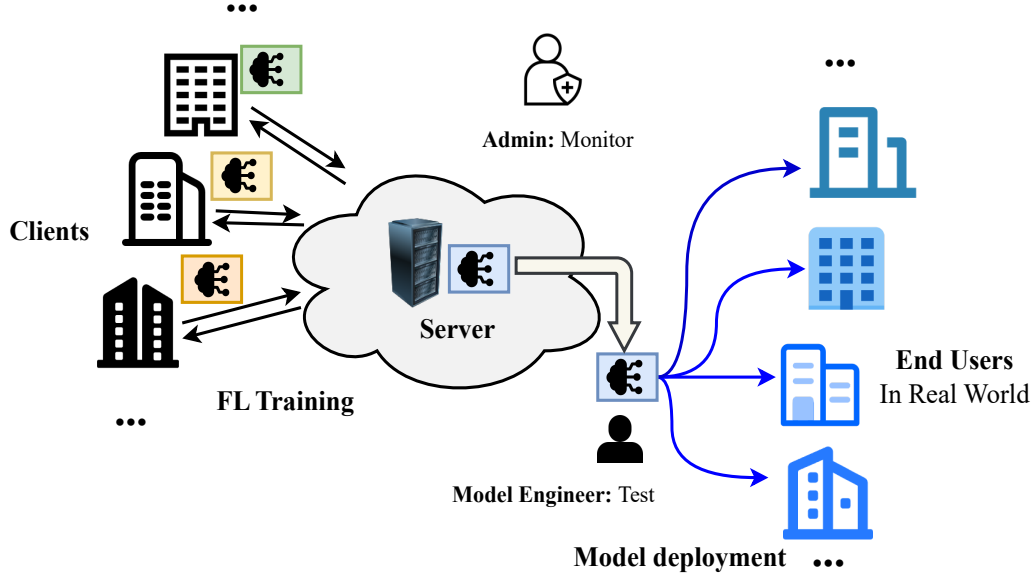


Fig. 2. The lifecycle of a trained model in FL System.

return their updated model weights to the central server, which aggregates them using an algorithm such as Federated Averaging (FedAvg) or other Optimized aggregation algorithms.

Step 7: Model Evaluation. The engineer evaluates the aggregated model on a validation dataset to determine its accuracy and generalization performance. If the performance is unsatisfactory, the training process can be repeated with a different set of initial weights or hyper-parameters.

Step 8: Model Deployment. Once the model is trained, it can be deployed to predict new and unseen domains. In some cases, the model may need to be fine-tuned on new data to adapt to changing conditions or to improve its performance.

Overall, the workflow involves multiple iterations of local training, model aggregation, and evaluation, and it requires close collaboration between the model engineer and the participating clients. The success of an FL project depends on careful planning, effective communication, and the ability to adapt to changing conditions and requirements.

2) The Key Actors in FL System.

FL involves multiple actors that play different roles in the training and deployment of a model. Here are the key actors involved in an FL system:

Model Engineer: The model engineer is responsible for developing the ML model and defining the training protocol for the FL system. This involves selecting the appropriate model architecture, optimization algorithm, and partitioning strategy for the distributed data. The model engineer also evaluates the performance of the trained model and fine-tunes it as needed.

Admin: In FL, an admin is a role that involves managing the overall operation of the FL system. An admin may be an individual or a team responsible for configuring and monitoring the system, as well as ensuring its security, privacy, and compliance with legal and ethical requirements.

Central Server: The central server coordinates the FL training process by collecting and aggregating the model parameters from the participating clients. The server performs the model aggregation using algorithms.

Participating Clients: Participating clients are the devices or organizations that contribute their local data and computational resources to the FL training process. Participating clients perform local training on their data and share their model parameters with the central server for aggregation.

End Users: End-users are the individuals or organizations that use the trained model to make predictions or decisions based on new or unseen domains. End-users may include consumers, businesses, or government agencies that benefit from the insights generated by the model. Depending on the application requirements, the trained model may be deployed on a mobile device, a cloud server, or an edge device.

Communication Channel: The communication channel is the medium used for transmitting the model weights between the central server and the participating clients. It can be a wireless or wired network connection and must be designed to ensure the privacy and security of the data transmitted between the parties involved.

Overall, an FL system involves multiple actors collaborating to train and deploy a machine-learning model. The success of an FL project depends on the effective coordination and communication between these actors, as well as the careful

consideration of privacy, security, and performance concerns.

3) The Typical Process of FL Training.

The typical process of FL training involves several steps, which are outlined below:

Step1: Client Selection. Client selection is an important step as it determines which clients will participate in the training process. The goal is to select a representative set of clients that can contribute useful information to the global model without compromising data privacy or computational resources. The approaches to client selection can be a random selection or other well-designed client selection mechanisms. We select $\mathcal{N}^s$ clients out of $\mathcal{N}$ at this stage ($\forall \mathcal{N}^s \subseteq \mathcal{N}$) because some of the clients may be inappropriate.

Step2: Download. In this step, the clients first need to download the global model aggregated by the central server in the previous round t (the global model is randomly initialized in the first round).

$$\omega_t^s = \omega_t. \quad (1)$$

where ω_t denotes the global model in the round t .

Step3: Local Training. The participating clients perform local training on their respective domains using the downloaded model.

$$\omega_{t+1}^s = \omega_t^s - \eta \nabla L^i(\omega_t^s; \mathcal{X}^i). \quad (2)$$

where η denotes the learning rate, $L^i(\omega_t^s; \mathcal{X}^i)$ represents the local loss function for client i at round t , ω_{t+1}^s represents the updated local model parameters for client i at round $t+1$.

Step4: Upload. The trained local models are then sent back to the central server for aggregation.

Step5: Model Aggregation. The central server aggregates the model parameters from the participating clients using a suitable algorithm. The aggregation process aims to combine the local model updates into a single global model that reflects the collective knowledge of all clients.

$$\omega_{t+1} = \frac{\sum_{i \in \mathcal{N}^s} |\mathcal{X}^i| \omega_{t+1}^i}{\sum_{i \in \mathcal{N}^s} |\mathcal{X}^i|} \quad (3)$$

where ω_{t+1} denotes the aggregated global model in the round $t+1$.

Several rounds are required to achieve convergence for FL, in which one round consists of K iterations. Terminating the training process mainly depends on the specific objectives and requirements of accuracy and training time [55]. And the FL aims to get the optimal weights of the global model ω^* to minimize the global loss function $L(\omega)$:

$$L(\omega) = \frac{\sum_{i \in \mathcal{N}^s} |\mathcal{X}^i| L^i(\omega; \mathcal{X}^i)}{\sum_{i \in \mathcal{N}^s} |\mathcal{X}^i|}. \quad (4)$$

where $L^i(\omega; \mathcal{X}^i)$ is the loss for subset $\mathcal{X}^i$ given ω .

$$\omega^* = \arg \min_{\omega} L(\omega). \quad (5)$$

B. Domain Adaptation

Recent studies [56]–[58] have indicated that deep learning models are vulnerable to significant performance degradation when evaluated on out-of-distribution datasets, even with minor variations in the data generation process. Specifically, it demonstrated that deep learning models exhibit limited generalization capacity, which can significantly impact their performance when deployed in real-world scenarios. DA [59]–[65] emerged as a strategy to circumvent the out-of-distribution (OOD) data challenge by collecting a subset of data from the target domain, thereby enabling the adaptation of a model trained on a source domain with disparate distributions to the target domain. Fig. 3 compares traditional ML tasks and DA tasks. The basic idea behind DA is that even though the source and target domains may be different, they may still share some underlying structures or patterns that can be leveraged to improve the performance of a model in the target domain, summarized as **Definition 2**.

Definition 1 (Domain). A domain refers to a collection of data that is sampled from a joint distribution. Consider a nonempty input space denoted as $\mathcal{X}$ and an output space denoted as $\mathcal{Y}$. The domain, denoted as $\mathcal{S}$, comprises n samples $\mathcal{S} = \{(x_j, y_j)_{j=1}^n \sim P_{XY}\}$, where $x_j \in \mathcal{X} \subset \mathbb{R}^d$ represents a d -dimensional input sample, $y_j \in \mathcal{Y} \subset \mathbb{R}$ represents the corresponding output label, and P_{XY} represents the probability distribution of the joint input-output space, with X and Y as the corresponding random variables.

Definition 2 (Domain Adaptation). Let h represent any labeling function from the hypothesis space $\mathcal{H}$. Given M source domains denoted as $\mathcal{S}_{source} = \{\mathcal{S}^i | i = 1, 2, \dots, M\}$ where $\mathcal{S}^i = \{(x_j^i, y_j^i)_{j=1}^{n_i}\}$ represents the i -th domain with true labeling functions h^{*s} , it is observed that the joint distributions between each pair of domains differ ($P_{XY}^i \neq P_{XY}^j$, for $1 \leq i \neq j \leq M$). The objective of domain adaptation (DA) is to minimize the disparity between the source and target domains, enabling effective generalization on the target domain $\mathcal{S}_{target}$, where $\mathcal{S}_{target}$ is accessible during training but has a distinct distribution ($P_{XY}^{target} \neq P_{XY}^i$ for $i \in 1, 2, \dots, M$), with corresponding true labeling functions denoted as h^{*t} .

Thus, the difference in the source domain between two labeling functions h and $\hat{h}$ can be measured by:

$$\theta^s(h, \hat{h}) = \mathbb{E}_{x \sim P_X^s} [h(x) \neq \hat{h}(x)] = \mathbb{E}_{x \sim P_X^s} [|h(x) - \hat{h}(x)|]. \quad (6)$$

Similarly, we can get θ^t when $x \sim P_X^t$. Let $\theta^s(h) := \theta^s(h, h^{*s})$ and $\theta^t(h) := \theta^t(h, h^{*t})$ denote the risk of labeling function h on the source and target domains, respectively.

The aim of DA is to minimize the risk of the target domain $\theta^t(h)$. However, the target risk is often inaccessible because we do not have any information about the true underlying distribution of the target domain in practice. To overcome the challenge, Ben *et al.* [66] seek to bound the risk of target domain $\theta^t(h)$ using the tractable risk of source domain $\theta^s(h)$:

$$\theta^t(h) \leq \theta^s(h) + d_{\mathcal{H}\Delta\mathcal{H}}(P_X^s, P_X^t) + \lambda_{\mathcal{H}}. \quad (7)$$

where the discrepancy between cross-domain distributions is quantified by the $\mathcal{H}\Delta\mathcal{H}$ -divergence $d_{\mathcal{H}\Delta\mathcal{H}}(P_X^s, P_X^t)$, while

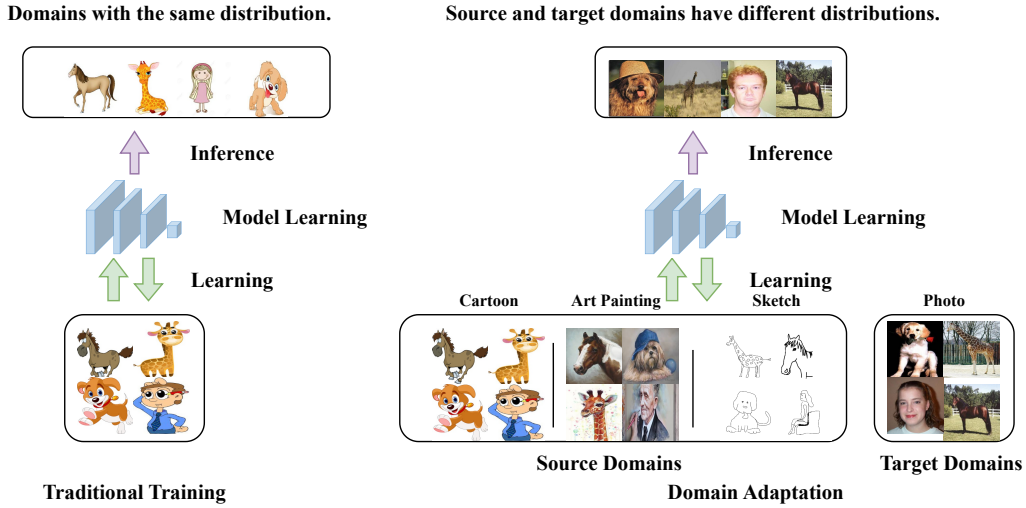


Fig. 3. Comparison of traditional ML tasks (left) and domain adaptation tasks (right).

the complexity of $\mathcal{H}$ for prediction tasks on the source and target domains is assessed by the ideal joint risk $\lambda_{\mathcal{H}}$. Notably, the $\mathcal{H}\Delta\mathcal{H}$ -divergence offers superior finite sample guarantees compared to other divergence measures and allows for non-asymptotic bounds. The non-asymptotic bound ensures that the bound holds with high probability, regardless of the sample size, which is particularly useful in applications where the sample size is small or limited. This means that with a finite sample size, the $\mathcal{H}\Delta\mathcal{H}$ -divergence can provide a more accurate estimate of the difference between two probability distributions than other measures.

Theorem 1 (DA error bound (non-asymptotic) [51], [66]). Let $\mathcal{U}^s$ and $\mathcal{U}^t$ represent unlabeled samples of size k from the two domains. The Vapnik-Chervonenkis (VC) dimension [67] of $\mathcal{H}$, denoted as d . For any $h \in \mathcal{H}$ and $\epsilon \in (0, 1)$, the following inequality holds with a probability of at least $1 - \epsilon$:

$$\theta^t(h) \leq \theta^s(h) + \hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}^s, \mathcal{U}^t) + \lambda_{\mathcal{H}} + 4\sqrt{\frac{2d \log(2k) + \log(\frac{2}{\epsilon})}{k}} \quad (8)$$

where $\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}^s, \mathcal{U}^t)$ denotes the estimated difference of $d_{\mathcal{H}\Delta\mathcal{H}}(P_X^s, P_X^t)$ on the two domains of finite data samples. DA error bound (non-asymptotic) refers to the concept of learning algorithms that can generalize to target domains by using a finite number of samples obtained from both the source and target domains during training. The error bound refers to the difference in risk between the target domain and the source domain. Specifically, it quantifies the distance between the average loss in the target and source domains. The bound is non-asymptotic, which means that it is valid for any size of the sample set, and thus enables effective control of the generalization performance of the algorithm in practical applications.

In DA, the difference in domain distributions between the source domain P_X^s and the target domain P_X^t is a significant challenge. The distribution difference in the above error bound, denoted as $d(P_X^s, P_X^t)$, can lead to poor generalization of

models trained on the source domain when applied to the target domain. Nonetheless, although direct control over the distribution difference can be challenging, a potential approach is to train a representation function $g: \mathcal{X} \rightarrow \mathcal{Z}$ that maps the original input data x to a representation space $\mathcal{Z}$, aiming to minimize the discrepancy between the distribution of representations from the two domains. This way, the representation distributions of the two domains can be brought closer together. By transforming the input data to a representation space where the domain difference is minimized, the learned model can generalize better across domains, known as DA based on domain-invariant representation (DA-DIR). In particular, the goal is to find a representation space $\mathcal{Z}$ where the distributions P_Z^s and P_Z^t of the source and target domains, respectively, are as close as possible. This can be achieved through various techniques, such as adversarial training, minimizing the discrepancy between the two distributions, domain confusion loss, and transfer learning.

Overall, DA is a powerful technique for improving the performance of ML models in situations where the source and target domains have different distributions, which is a common problem in many real-world scenarios.

C. Domain Generalization

DA assumes some labeled or unlabeled data from the target domain is accessible for model adaptation, but it may not always be the case in practice. In some real-world scenarios, collecting labeled data from the target domain can be time-consuming, expensive, or even impossible due to ethical or legal constraints. Additionally, in some cases, the distribution of the target domain data may be unknown, making it challenging to apply traditional DA techniques that rely on knowledge of the target domain distribution. To address the issue, domain generalization (DG) has emerged as a promising technique to improve the generalization performance of ML models to new and unseen domains by training them on diverse datasets that cover multiple domains. The process of

the domain generalization task is illustrated on the left of Fig. 4 and summarized in **Definition 3**.

Definition 3 (Domain Generalization). Given M source domains $\mathcal{S}_{source} = \{\mathcal{S}^i | i = 1, 2, \dots, M\}$ where $\mathcal{S}^i = \{(x_j^i, y_j^i)\}_{j=1}^{n_i}$ represents the i -th domain, it is observed that the joint distributions between each pair of domains differ ($P_{XY}^i \neq P_{XY}^{\hat{i}}$, for $1 \leq i \neq \hat{i} \leq M$). The objective of domain generalization (DG) is to develop a robust and generalizable labeling function $h : \mathcal{X} \rightarrow \mathcal{Y}$ solely utilizing the M source domain data, which minimizes the prediction error on an unseen target domain $\mathcal{S}_{target}$, where $P_{XY}^{target} \neq P_{XY}^i$ for $i \in \{1, 2, \dots, M\}$:

$$\min_h \mathbb{E}_{(x,y) \in \mathcal{S}_{target}} [\ell(h(x), y)]. \quad (9)$$

where $\mathbb{E}$ denotes the expectation operator and $\ell(\cdot)$ represents the loss function used to measure the difference between the predicted output $h(x)$ and the true output y . The ultimate aim is to find the optimal function $h(\cdot)$ that minimizes the expected loss on $\mathcal{S}_{target}$, which reflects the model's ability to generalize well to new and unseen domains.

In DG, underlying hyper-distribution is a key concept used to capture the relationship between the source and target domains. The probability distribution represents the joint distribution over all possible (x, y) for both source and target domains. Specifically, assume that all possible target distributions follow the hyper-distribution $\mathcal{P}$: $P_{XY}^t \sim \mathcal{P}$, and the same is true for the source distributions: $P_{XY}^1, P_{XY}^2, \dots, P_{XY}^M \sim \mathcal{P}$. This allows for generalization to any possible target domain, even if it is completely unknown, by learning a classifier that incorporates domain information P_X as part of its input. Thus, the prediction on the domain with the distribution P_{XY} takes the form $y = h(P_X; x)$. For such a labeling function $h(\cdot)$, its average risk over all possible target domains is given by the following formula:

$$\psi(h) = \min_h \mathbb{E}_{P_{XY} \sim \mathcal{P}} \mathbb{E}_{(x,y) \sim P_{XY}} [\ell(h(P_X, x), y)]. \quad (10)$$

However, it is often infeasible to evaluate expectations exactly. We can estimate them using finite domains/distributions that follow a given probability distribution $\mathcal{P}$, along with finite samples of (x, y) drawn from each distribution. Given that $P_{XY}^1, P_{XY}^2, \dots, P_{XY}^M \sim \mathcal{P}$, we can use the source domains and their corresponding supervised data to estimate the expectation. This is because the source domains and their associated data are assumed to be drawn from the same underlying distribution $\mathcal{P}$. As a result, the estimation of the average risk over all possible target domains based on the labeling function $h(\cdot)$ is as follows:

$$\hat{\psi}(h) = \frac{1}{M} \sum_{i=1}^M \frac{1}{n^i} \sum_{j=1}^{n^i} \ell(h(\mathcal{U}^i, x_j^i), y_j^i). \quad (11)$$

where M is the total number of domains, n^i is the number of samples in domain i , the supervised dataset $\mathcal{U}^i = \{x_j^i | (x_j^i, y_j^i) \in \mathcal{S}^i\}$ represents the empirical estimation for the marginal distribution P_X^i of the features in domain i .

Theorem 2 (Domain generalization error bound [51], [68]). Let $\gamma := \min_{\pi \in \Delta_M} d_{\mathcal{H}}(P_X^t, \sum_{i=1}^M \pi_i P_X^i)$ with minimizer

π^* represents the minimum distance between the target domain distribution P_X^t and the convex hull Λ formed by the source domain distributions P_X^i . Δ_M represents the $(M-1)$ -dimensional simplex that approximates the target domain distribution P_X^t with a convex hull Λ of source domain distributions P_X^i . The minimizer π^* corresponds to the weights that give the best approximation $P_X^* = \sum_{i=1}^M \pi_i^* P_X^i$ of P_X^t within Λ , where π denotes a normalized mixing weights. The parameter $\rho = \sup_{P_X', P_X'' \in \Lambda} d_{\mathcal{H}}(P_X', P_X'')$ represents the diameter of the convex hull Λ . Then, the domain generalization error bound can be expressed as:

$$\theta^t(h) \leq \sum_{i=1}^M \pi_i^* \theta^i(h) + \frac{\gamma + \rho}{2} + \lambda_{\mathcal{H}}(P_X^t, P_X^*). \quad (12)$$

where $\lambda_{\mathcal{H}}(P_X^t, P_X^*)$ represents the ideal joint risk across the target domain P_X^t and the domain with the best approximator distribution P_X^* . $(M-1)$ -dimensional simplex is often used to represent the mixing weights of source domains for approximating the target domain distribution. Specifically, for M source domain distributions, they can be viewed as M points in Euclidean space, and these points can be connected to form an $(M-1)$ -dimensional simplex. In this simplex, each point corresponds to a source domain distribution, and each edge corresponds to a mixing weight of a source domain distribution (i.e., a real number between 0 and 1). And the sum of mixing weights for each point is 1, representing their convex combination. Therefore, these mixing weights can be used to approximate the target domain distributions, thus achieving the purpose of domain generalization. In practical implementation, some optimization algorithms are usually needed to determine the values of these mixing weights so that the approximated target domain distributions can be as close as possible to the actual one.

Similar to DA, the bound encourages the use of DG methods that rely on the domain-invariant representation. These methods aim to minimize the risks across all source domains in the first term of the bound and reduce the differences in representation distribution between the source and target domains. The latter is measured by γ and ρ , which are used to quantify the distance between the representations of the source and target domains. By minimizing the error bound, the model can learn to generalize well across different domains and perform well on the target domain, even if it was an unseen domain.

To sum up, DG is a promising approach for developing a model that can effectively handle distribution differences across various source domains by minimizing the distribution discrepancy between the source and target domains and promoting domain-invariant representations, so that it can perform well on unseen target domains in the real-world scenarios.

D. Fedrated Domain Generalization

Recently, the proliferation of IoT devices has created enormous amounts of diverse data from multiple domains, providing ideal conditions for DG research. The fundamental principle of DG is to access the multi-source distributions in the learning process. However, the sensitive data generated

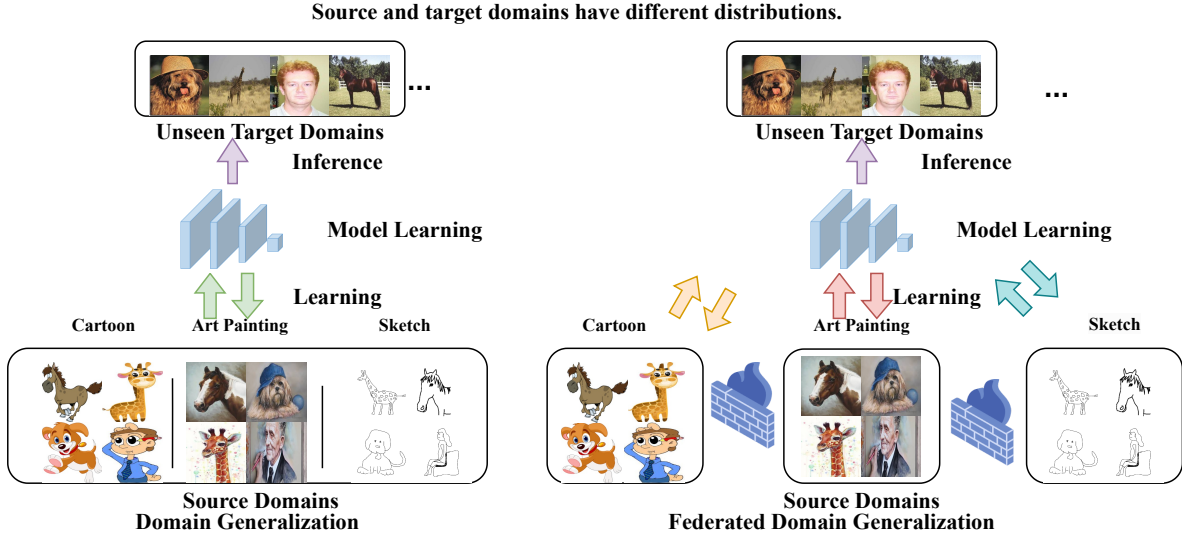


Fig. 4. Comparison of domain generalization task (left) and FDG task (right).

by IoT devices, such as healthcare and transportation data, are often subject to strict privacy regulations, making it challenging to implement traditional DG methods that rely on centralized data processing. To tackle the challenge of training models on data from multiple domains and obtaining a robust and generalizable predictive function while preserving data privacy, FDG has emerged as a promising solution [39]. FDG has many applications in areas such as healthcare, finance, and autonomous vehicles, where the data is often distributed across multiple devices or servers, and the privacy of the data is a critical concern. The process of the FDG task is illustrated on the right of Fig. 4 and summarized in **Definition 4**.

Definition 4 (FDG). Given M source domains $\mathcal{S}_{source} = \{\mathcal{S}^i | i = 1, 2, \dots, M\}$ involved in federated learning where $\mathcal{S}^i = \{(x_j^i, y_j^i)\}_{j=1}^{n_i}$ denotes the i -th domain. The joint distributions between each pair of domains are different: $P_{XY}^i \neq P_{XY}^{\hat{i}}, 1 \leq i \neq \hat{i} \leq M$. The objective of FDG is to develop a robust and generalizable labeling function $h : \mathcal{X} \rightarrow \mathcal{Y}$ using only M source domains data that can minimize the prediction error on unseen target domains $\mathcal{S}_{target}$ (where $P_{XY}^{target} \neq P_{XY}^i$ for $i \in \{1, 2, \dots, M\}$) while each domain i can only access its local data $\mathcal{S}^i$:

$$\min_h \mathbb{E}_{(x,y) \in \mathcal{S}_{target}} [l(h(x), y)]. \quad (13)$$

where $\mathbb{E}$ denotes the expectation operator and $l(\cdot)$ represents the loss function used to measure the difference between the predicted output $h(x)$ and the true output y . The ultimate aim is to find the optimal function $h(\cdot)$ that minimizes the expected loss on $\mathcal{S}_{target}$, which reflects the model's ability to generalize well to new and unseen domains.

It is important to note that the boundaries between these concepts can sometimes overlap, and the scope of them is shown in Fig. 5. The specific techniques and approaches used in each area can vary, but they all revolve around adapting or

generalizing models to different domains or decentralized data settings.

E. Challenges

FDG is a promising research area in ML that aims to develop models capable of generalizing well across multiple unseen domains while achieving privacy-preserving. However, the implementation of FDG encounters several challenges that need to be addressed to achieve this goal. One of the primary challenges faced in FDG is domain shift, which occurs when a model trained on data from one domain performs poorly on data from a different domain due to significant differences in statistical properties. Additionally, data heterogeneity, limited data availability, and privacy concerns pose significant challenges for FL. The distribution of data across multiple devices or organizations in FDG can result in significant heterogeneity, limited data for training, and privacy concerns due to the distribution of data. Moreover, high communication overhead can hinder efficient communication between domains during FL, which can further exacerbate these challenges.

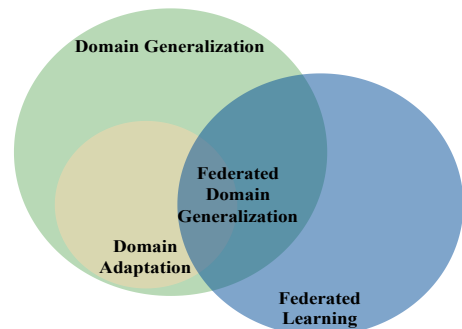


Fig. 5. The scope of proposed survey on federated domain generalization.

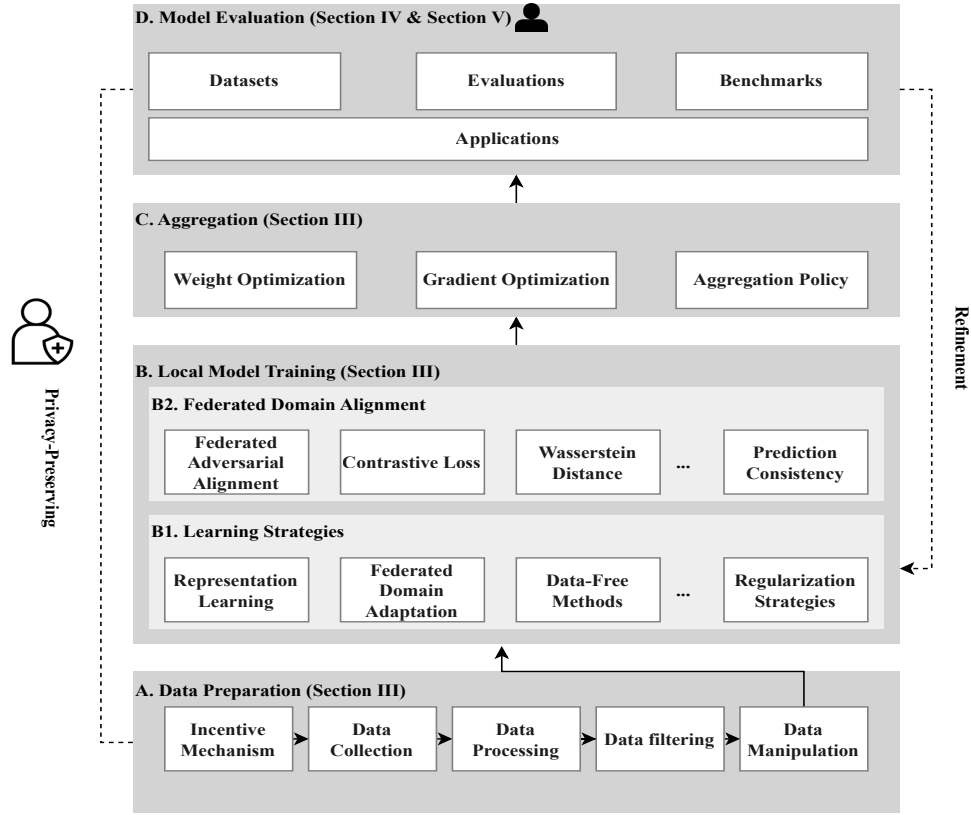


Fig. 6. The architecture and workflow of FDG.

Collectively, by addressing these challenges, FDG can advance the development of privacy-preserving, robust, and generalized ML models. This will not only benefit academic research but also have practical implications in real-world scenarios where the ability to generalize across multiple unseen domains is crucial.

III. METHODOLOGIES: A SURVEY

FDG methods have emerged as a promising approach for addressing the challenges of domain shift and heterogeneity in FL while preserving data privacy. Despite it has gained attention in recent years due to its potential to improve the generalization and robustness of models in federated settings, there has been limited research focused on this area. In this section, we aim to address this gap by providing a comprehensive overview of the existing literature on FDG and categorizing them into several distinct groups based on their methodologies and motivations, as illustrated in Fig. 7.

A. Federated Domain Alignment

Federated domain alignment [69]–[71] is of utmost importance in FDG as it enhances model generalization across multiple domains, promotes data efficiency, preserves data privacy, and enables the applicability of FL in real-world scenarios. The primary motivation of federated domain alignment is to enhance the generalization capability of models across multiple domains. By aligning the domains and harmonizing

their data representations, federated domain alignment aims to reduce the domain shift and improve model performance on unseen domains. This is particularly important in real-world applications where data privacy and security are paramount, as it allows for effective knowledge transfer while ensuring that sensitive data remains decentralized and protected. The objective for the joint distribution of alignment-based FDG methods can be expressed as:

$$\begin{aligned}
 P(X, Y) &= \sum_{i=1}^n P(D_i) P(X|D_i) P(Y|X, D_i) \\
 &= \sum_{i=1}^n P(D_i) P(Y|D_i) P(X|Y, D_i)
 \end{aligned} \tag{14}$$

where $P(X, Y)$ represents the joint distribution of input features X and corresponding labels Y across all domains. $P(D_i)$ represents the distribution of domain i , indicating the probability of selecting samples from domain i . $P(X|D_i)$ represents the conditional distribution of input features X given domain i , capturing the domain-specific characteristics of the data. $P(Y|X, D_i)$ represents the conditional distribution of labels Y given input features X and domain i , describing the relationship between the features and labels within each domain.

$P(X, Y)$ sums over all domains i to consider the contribution of each domain to the overall joint distribution $P(X, Y)$. The two equations represent two different decompositions of the joint distribution, highlighting different relationships

between the variables. In FDG, our main focus is on aligning the marginal distributions of the source domains. Specifically, we aim to align $P(D_i)P(X|D_i)$ to reduce the distribution shift between domains. This is because we assume that the posterior probability $P(Y|X, D_i)$ remains relatively stable and has a smaller impact on the model. Therefore, domain alignment methods primarily concentrate on adjusting the distribution of input features to mitigate differences across domains and improve the model's generalization performance on unseen domains.

a) Federated Adversarial Alignment

Federated adversarial alignment [69] is proposed to align the feature distributions between different domains in FL by dividing optimization into two independent steps: training domain-specific local feature extractors and training a global discriminator. The following steps are as follows. (1) Local Feature Extractors: A feature extractor G_s is trained for the source domain $\mathcal{S}^i$, and a feature extractor G_t is trained for the target domain $\mathcal{S}^t$. (2) Adversarial Domain Alignment: For each source-target domain pair $(\mathcal{S}_s, \mathcal{S}_t)$, an adversarial domain identifier DI is trained where the purpose of DI is to align the feature distributions between the domains in an adversarial manner. Specifically, DI is trained to distinguish between the features originating from the source and target domains. Initially, DI is trained to accurately identify the domain of origin for the features. Subsequently, the generator models (G_s, G_t) are trained to confuse DI by producing features that make it challenging for DI to differentiate between the source and target domains. The objective of DI_i is defined as follows:

$$L_{adv_{DI_i}}(\mathcal{S}_s, \mathcal{S}_t, G_s, G_t) = -\mathbb{E}_{\mathcal{S}_s \sim \mathcal{S}_s} [\log DI_i(G_s(\mathcal{S}_s))] - \mathbb{E}_{\mathcal{S}_t \sim \mathcal{S}_t} [\log(1 - DI_i(G_t(\mathcal{S}_t)))] \quad (15)$$

In the second step, the objective for updating the generators G_s and G_t (denoted as L_{adv_G}), while keeping L_{adv_D} unchanged, is defined as follows:

$$L_{adv_G}(\mathcal{S}_s, \mathcal{S}_t, DI_i) = -\mathbb{E}_{\mathcal{S}_s \sim \mathcal{S}_s} [\log DI_i(G_s(\mathcal{S}_s))] - \mathbb{E}_{\mathcal{S}_t \sim \mathcal{S}_t} [\log DI_i(G_t(\mathcal{S}_t))] \quad (16)$$

Wang *et al.* [72] introduced a multiclient feature alignment module using adversarial learning, where a generative adversarial network with class-wise information is utilized to generate a reference distribution and an adversarial training strategy is employed to learn domain generalization features from different clients, addressing distribution discrepancies.

The federated adversarial alignment approach aligns feature distributions across domains by training domain-specific local feature extractors and a global discriminator, facilitating knowledge transfer and adaptation while preserving privacy in the FL framework.

b) Minimizing Contrastive Loss

The contrastive loss [13], [14] encourages similar representations for samples from the same domain and dissimilar representations for samples from different domains by maximizing agreement within positive pairs and minimizing agreement within negative pairs. The goal of minimizing the contrastive loss is to learn a representation space where similar samples are pulled closer together, while dissimilar samples

are pushed farther apart. Mathematically, the contrastive loss can be defined as follows:

$$L_{Contrastive} = \frac{1}{2N} \sum_{i=1}^N y d^2 + (1 - y) \max(\text{margin} - d, 0)^2 \quad (17)$$

where N is the number of sample pairs, d is the distance between the representations of the paired samples, and margin is a hyperparameter that defines the minimum desired distance between representations of dissimilar samples.

Li *et al.* [73] proposed the MOON (model-contrastive federated learning) framework that leverages model-level contrastive learning to reduce the distance between global and local features and improve the performance of FL, surpassing other state-of-the-art algorithms. As an extension of [73], an improved contrastive loss [74] is utilized by considering the projected features of pseudo-data as positive pairs and incorporating the projected local feature of both pseudo-data and local data as negative pairs. To enhance model invariance across domains, an alignment loss is deployed between the original sample and its hallucinated counterpart using a negative-free contrastive loss [75] at the logit level. In response to the under-explored issues of data heterogeneity and class imbalance in federated learning, a novel privacy-preserving framework called FedIIC [76] is proposed, which effectively calibrates deep models to address imbalanced training using intra-client and inter-client contrastive learning techniques, achieving superior performance in both realistic and existing scenarios. Tan *et al.* [77] proposed a lightweight framework called FedPCL (Federated Prototype-wise Contrastive Learning), enabling clients to jointly learn by fusing representations from multiple pre-trained models, improving each client's ability to leverage class-relevant information in a personalized manner while maintaining compact shared knowledge for efficient communication.

To sum up, this approach helps to align the representations of different domains, reducing the domain shift and enhancing the model's ability to generalize across diverse domains.

c) Wasserstein Distance

The Wasserstein distance, also known as the Earth Mover's distance, is a measure of the dissimilarity between probability distributions. In FDG, the Wasserstein distance can be used to quantify the difference between the distributions of different domains and learn a model that can generalize well to unseen domains.

Chen *et al.* [78] introduced a Wasserstein-based feature critic with meta-optimization into FL, which effectively improved the robustness and generalization capability of the proposed method. Nguyen *et al.* [79] proposed a Wasserstein distributionally robust optimization scheme called WAFL for FL, which is more general than related approaches and is robust to all adversarial distributions inside the Wasserstein ball. Lin *et al.* [80] introduced a robust federated meta-learning framework that utilizes Wasserstein distance and Lagrangian relaxation to efficiently optimize the distance metric, enabling gradient-based methods to solve the problem effectively.

In summary, the integration of the Wasserstein distance within the objective function of FDG methods facilitates

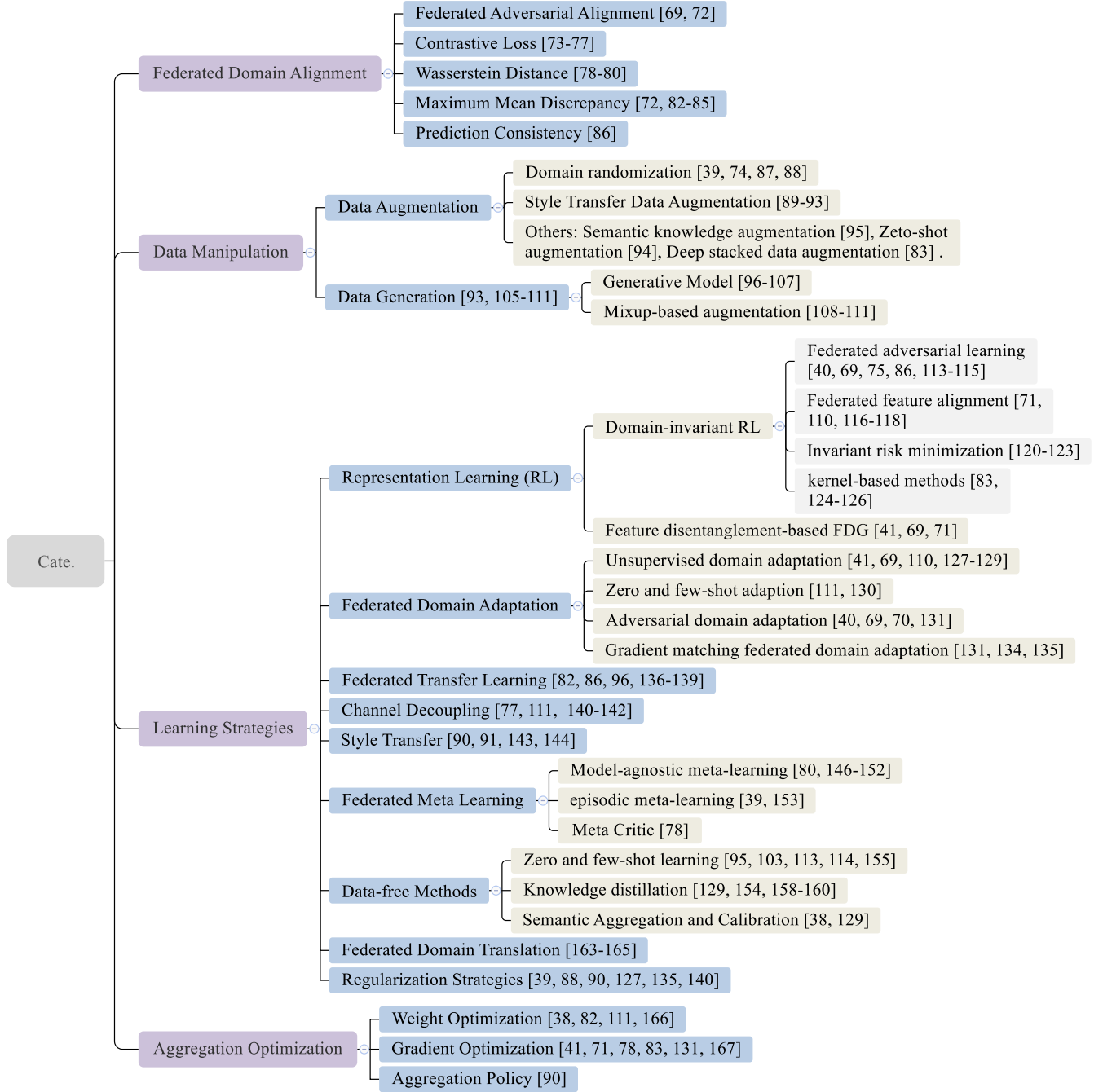


Fig. 7. Categorization of FDG Methods.

the learning of domain-invariant representations, enabling improved generalization across federated domains.

d) Minimizing Maximum Mean Discrepancy

Maximum Mean Discrepancy (MMD) [81] is a statistical metric used to measure the discrepancy between probability distributions of different domains. The goal of using MMD in FDG is to minimize the distribution shift between domains and promote domain alignment. By minimizing the MMD, FDG aims to learn domain-invariant representations that capture shared knowledge across domains while preserving domain-

specific characteristics, calculated as follows:

$$MMD^2(P, Q) = \left\| \frac{1}{n'} \sum_{j=1}^{n'} v(x_j) - \frac{1}{n''} \sum_{j=1}^{n''} v(y_j) \right\|^2 \quad (18)$$

where P and Q are probability distributions of the feature representations of two domains, n' and n'' are the number of samples in each domain, and $v(\cdot)$ is a feature mapping function that maps the samples to a high-dimensional space.

Chen *et al.* [82] implemented D-WFA using the MMD distance between the source and target clients, which helps

to reduce domain differences and improve the generalization capability. To sum up, we can effectively address the domain shift problem and enhance the generalization capabilities of FL models by incorporating MMD in FDG. The proposed method [83] enables the aggregated gradient to incorporate information from multiple domains by connecting gradient alignment with MMD, aiming for better generalization on unseen domains. Zhang *et al.* [84] employed MMD to measure and optimize the distance between different domains by calculating the squared distance between the kernel embeddings of the data representation distributions in the reproducing kernel Hilbert space (RKHS), with unbiased empirical estimation used to reduce computation load. Zhang *et al.* [85] proposed a blockchain-based decentralized federated transfer learning method for collaborative machinery fault diagnosis, addressing the limitations of data quality and quantity by ensuring data security and privacy, achieving high testing accuracies and offering a promising tool for real industrial applications. Wang *et al.* [72] presented a federated adversarial domain generalization network for machinery fault diagnosis, enabling collaborative training between a central server and multiple clients while ensuring data privacy, achieving significant performance improvements and demonstrating promise for data privacy-preserving generalization diagnosis tasks.

Overall, minimizing the MMD help FDG encourages the feature representations from different domains to have similar means in the high-dimensional space, thus reducing the distributional discrepancy. This promotes the alignment of domains and improves the generalization capability of models across different domains.

e) Prediction Consistency

Prediction consistency refers to the property of an ML model to produce consistent and reliable predictions across different instances or repetitions of the same input. Achieving prediction consistency in FDG enhances the generalization capability, reduces bias, promotes transferable knowledge, offers cost and privacy benefits, and enables scalability. Zhang *et al.* [86] leveraged prediction consistency to enhance the generalization capability of the model. The inference results for the target domains are obtained by majority voting on the predictions made by multiple source domains. Considering the fact that these source domains have different parameters, prediction consistency generally improves the model's generalization capability while reducing the risk of overfitting. Thus, it is desirable to minimize variations in the inference results among the source domains. To achieve this, the loss function L_p is defined as follows:

$$\min L_p = \frac{1}{n^t M} \sum_{j=1}^{n^t} \sum_{i=1}^M \|I^{s,i}(F^t(x_j^t)) - \bar{I}(F^t(x_j^t))\|, \quad (19)$$

$$\bar{I}(F^t(x_j^t)) = \frac{1}{M} \sum_{i=1}^M I^{s,i}(F^t(x_j^t))$$

where n^t represents the number of target client data samples, M represents the number of source inference models, $F^t(x_j^t)$ represents the feature representation of the target domain data x_j^t in the model, $I^{s,i}$ represents the inference module at the

i -th source domain, $\bar{I}(F^t(x_j^t))$ denotes the mean value of predictions obtained from the source inference models.

Overall, the advantages of prediction consistency make it a desirable goal in various applications where robust and generalizable models are required across diverse and unseen domains.

B. Data Manipulation

The performance of ML models heavily relies on the quality and quantity of training data, especially in domain generalization where the models are expected to generalize well to unseen domains. However, acquiring a large and diverse set of training data that covers all possible domains is often infeasible or expensive. In addition, the privacy concerns associated with sensitive data make it difficult to collect and store data in a centralized manner. In the field of FDG, data manipulation techniques have been developed to address these challenges by creating more diverse and representative training data. One such technique is data augmentation, which can be used to create additional training data by modifying the existing data through techniques such as rotation, scaling, and adding noise. Furthermore, adversarial data augmentation can be used to generate challenging examples that can help improve the model's robustness against domain shifts and adversarial attacks.

The objective of data manipulation can be expressed mathematically as:

$$\min_{\omega} \sum_{i=1}^M \nu_i \mathcal{L}_i(\hat{\mathcal{S}}^i, \omega) + \xi R(\omega) \quad (20)$$

where M is the total number of domains, ω represents the model parameters to be learned, ν_i represent the importance of domain i in FL and is determined based on the quantity and quality of the domain i , $\mathcal{L}_i(\hat{\mathcal{S}}^i, \omega)$ measures the performance of the model on the domain i and includes both real and synthetic data generated by data generation techniques, the regularization term $R(\omega)$ encourages the model to learn domain-invariant features that are less sensitive to domain shifts and can improve generalization performance. Depending on the specific techniques used and the formula for manipulating the data $\hat{\mathcal{S}}^i$, existing work on data manipulation can be broadly categorized into two common types: data augmentation and data generation.

1) Data augmentation-based FDG

Data augmentation-based FDG involves applying various transformations to the existing training data, such as rotating, scaling, or adding noise, to create new training data that are similar to the original ones. This can help increase the diversity and robustness of the training data and improve the model's generalization performance to unseen domains. Here, the $\hat{\mathcal{S}}^i = \mathcal{S}^i \cup A^i$, where $\mathcal{S}^i$ is the local real dataset of domain i and A^i is the set of random perturbations used for data augmentation.

a) Domain randomization

In FDG, domain randomization can be used to simulate the variability of the data across different clients by randomizing certain aspects of the data during training. This can help the

model learn to be robust to domain shifts and better generalize to unseen clients.

Duan *et al.* [87] uses local domain randomization augmentation to create a globally balanced distribution. Guo *et al.* [74] proposed FedDebias, a unified algorithm that leverages mean-based random noise augmentation that can reduce the learning bias on local features and improve the performance of the model without introducing real data. Liu *et al.* [39] employed the data augmentation of random rotation, scaling, and flipping in FL tasks. Atwany *et al.* [88] applied data augmentations including resizing, horizontal and vertical flipping, random grayscale, ColorJitter, random rotation, translation, and Gaussian blur to fundus images to improve model performance and intra-domain generalizability.

b) Style Transfer Data Augmentation

Style transfer data augmentation [89] is a technique that leverages the principles of neural style transfer to generate new training examples by combining the content of one image with the style of another, enhancing the diversity and richness of the training dataset.

Shenaj *et al.* [90] took full advantage of the source dataset during the pre-training stage with style transfer data augmentation, randomly loading the target styles in the source images to mimic the target distributions. Chen *et al.* [91] proposed cross-client style transfer in FL involves each client computing and sharing their local styles with the central server, which can be either single-image styles or a local image style bank, to facilitate the data augmentation process and enhance diversity in the training data. Georgiadis *et al.* [92] proposed ST-FL (Style Transfer Federated Learning) for COVID-19 image segmentation, addressing the challenges of data variability and privacy regulations by incorporating a denoising CycleGAN at each client node to achieve robustness and comparable performance to a centrally-trained model. Lewy *et al.* [93] processed the augmentation involves dividing images within each node into batches, selecting a random image to compute statistics, and using those statistics to independently augment all images in the batch, resembling a style transfer augmentation within each node.

Apart from the data augmentation-based FDG mentioned above, there are other popular used techniques for data augmentation. *Zero-shot augmentation* generates diverse and realistic augmented samples by leveraging external knowledge or pre-trained models without requiring any additional labeled data, enhancing model generalization and performance. Hao *et al.* [94] introduced Fed-ZDA, a federated learning system that leverages zero-shot data augmentation to generate pseudo-exemplars of unseen classes while ensuring privacy, aiming to enhance fairness and accuracy performance uniformity across clients in federated networks. *Semantic knowledge augmentation* is a technique used in ML to improve model performance and generalization by leveraging external semantic knowledge to generate more accurate, diverse, and semantically consistent training data. Sun *et al.* [95] enrich the semantic space by combining text embedding, Gaussian noise, and attribute label during the training process, making the model better adapt to the instance-level visual space and improving the model performance and generalization. *Deep stacked Ddata*

augmentation-based FDG leverages the power of deep learning, data augmentation, and FL techniques to enhance the generalization capability of models across all domains. By augmenting the data and aligning the gradients, the models become more robust and capable of handling unseen domains, thereby improving the performance and applicability of FL in real-world scenarios. Tian *et al.* [83] proposed a deep stacked transformation data augmentation approach (called BigAug) for generalizing models to unseen domains, achieving good generalization on several unseen datasets. These methods have been demonstrated to generate diverse and domain-invariant training samples, facilitating effective knowledge transfer and adaptation across heterogeneous domains while preserving privacy within the FL framework.

2) Data generation-based FDG

Data generation-based FDG involves synthesizing new training examples from scratch using generative models, such as Variational Autoencoders (VAEs) or Generative Adversarial Networks (GANs). These models learn to generate new data samples that are similar to the training data in terms of their statistical properties and distribution but can be different in content and appearance. Data generation-based FDG can help overcome the limitations of limited and biased training data and improve the model's ability to generalize to new and unseen domains. Here, the $\hat{S}^i = S^i \cup C^i$, where S^i is the local real dataset of domain i and C^i is the virtual domain dataset used for data generation.

a) Generative model

The generative model can be used to generate synthetic data that captures the common characteristics of the different domains, enabling better domain generalization. The choice of generative model for FDG depends on factors such as the characteristics of the data from different domains, the availability of labeled data in each domain, and the desired level of DG.

Zhang *et al.* [96] proposed FedDA, a data augmentation method based on GAN for generating high-quality overlap samples by learning features from both overlap and non-overlap data, effectively expanding the available training data and enhancing the effect of data augmentation. Guo *et al.* [97] demonstrated the effectiveness of ACGAN in addressing data imbalance and augmentation tasks, and based on its mechanism, it can serve as a feasible data bridge for data privacy-preserving federated transfer learning, which is a novel application of ACGAN. To enable membership inference on other participants, Chen *et al.* [98] proposed a novel approach that utilizes locally deployed GANs to generate samples with all labels, thereby obtaining the data distribution necessary for the inference process. Zhang [99] leveraged GANs to generate synthetic data with the same underlying distribution as the original dataset, guided by the target federated learning model's discriminator, thereby enriching the training dataset and improving the imitative capability of the generated images. Zhang *et al.* [100] proposed DPGAN framework allows different hospitals to employ a privacy-preserving data augmentation method by utilizing distributed DPGAN models to generate high-quality training samples, addressing the issue of insufficient training data and subsequently achieving accurate

detection using the ResNet model in FL. Yan *et al.* [101] synthesized samples by a privacy-preserving generative adversarial network (PPWGAN-GP), which solves the cross-client variation problem and protects privacy with wide applicability. Li *et al.* [102] proposed *improved FL-GAN* (IFL-GAN) to learn globally shared GAN models by aggregating locally trained generator updates with the maximum mean difference (MMD). Sariyildiz *et al.* [103] built a generative model to generate training samples for unseen classes upon the WGAN [104], [105] that takes a combination of noise vectors and class embeddings as input, enabling the production of class-specific samples based on the information provided by the class embedding. Tang *et al.* [106] introduced virtual homogeneity learning (VHL), a method for handling data heterogeneity in federated learning (FL), which generates a separable and homogeneous virtual dataset by utilizing StyleGAN [107] from shared noise across clients to improve convergence speed and generalization performance while preserving privacy.

b) Mixup-based augmentation

The models mentioned above can serve as starting points, and further customization or hybrid approaches may be necessary based on the specific requirements of the federated learning setting. Mixup-based augmentation is forced to learn features that are common across the clients by mixing samples from different clients during training, thereby improving its ability to generalize to new, unseen clients.

Shin *et al.* [108] proposed a privacy-preserving XOR-based mixup data augmentation technique called XorMixup, which addresses the non-IID data problem in FL by collecting encoded data samples from other devices and decoding them only using each device's own data samples. Yoon *et al.* [109] proposed FedMix, a privacy-protected data augmentation technique, averaging local batches and subsequently applying Mixup in local iterations to generate augmented data. The proposed approach aims to protect the privacy of local data in FL scenarios while improving performance in non-iid settings. Yao *et al.* [110] generated a certain proxy to approximate the target distribution using for aligning features from different domains via mixup-based augmentation. [111] allows the generation of diverse domains by mixing local and global feature statistics (MixIG) while keeping data private, where MixIG is constructed by randomly interpolating instance and global statistics.

By incorporating data generation techniques in FDG, models can benefit from increased diversity and representativeness of the training data. This helps in capturing the shared features and reducing the domain-specific differences, ultimately leading to improved generalization performance across domains.

C. Learning Strategies

In addition to data manipulation, the field of FDG has garnered considerable attention within the broader ML paradigm. FDG research can be classified into several distinct approaches, including representation learning-based FDG, federated DA-based FDG, federated transfer learning-based FDG, federated adversarial learning-based FDG, style transfer-based FDG, federated meta learning-based FDG, and other strategies.

These approaches are devised to tackle the challenge of training models that can effectively generalize to unseen domains by employing various techniques and principles.

1) Representation Learning

Representation learning [112] has long been a focal point in ML, playing a pivotal role in the success of domain generalization. Representation learning in FDG follows the same principles as representation learning in general but is tailored to the FL setting. The basic principle involves training a neural network to learn a mapping function that transforms the input data from each domain into a common representation space. The objective function used in representation learning for FDG typically includes components that encourage domain-invariant representations while preserving the utility of the learned representation for downstream tasks. The goal is to learn domain-invariant representations that capture shared features across multiple federated domains, enabling models to generalize well to unseen domains. In FDG, we decompose the prediction function w as $w = u \circ z$, where z represents the representation learning function and u represents the classifier function. Mathematically, it can be formulated as:

$$\min_{u,z} \sum_{s \in \mathcal{S}} [\mathbb{E}_{(x,y) \sim P(s)} \mathbb{L}_t(u(z(x)), y) + \chi \mathbb{L}_d(z(x), D)] \quad (21)$$

where $\mathcal{S}$ represents the set of source domains, $\mathbb{E}_{(x,y) \sim P(s)}$ represents the expectation over the data samples (x, y) from a specific source domain $s \in \mathcal{S}$. $\mathbb{L}_t(\cdot)$ denotes the task-specific loss function, which measures the discrepancy between the predicted output of the representation function $u(z(x))$ and the true label y for a given input x . $\mathbb{L}_d(\cdot)$ represents the domain discrepancy loss function, which quantifies the difference between the representations $z(x)$ and the domain labels D . The domain discrepancy loss function can be defined using various approaches, such as maximum mean discrepancy, adversarial DA, or adversarial neural networks, depending on the specific method used for FDG. This loss encourages the learned representations to be domain-invariant, capturing shared knowledge across different source domains. The hyperparameter χ controls the trade-off between the task loss and the domain discrepancy loss, determining the relative importance of task performance and domain invariance in the learned representations.

a) Domain-invariant representation-based FDG

Domain-invariant representation-based FDG refers to the problem of learning models that can generalize well to unseen domains in the FL setting, while also capturing domain-invariant representations.

Federated adversarial learning: Federated domain adversarial learning combines the concepts of federated learning and adversarial training to train a global model on multiple source domains. By incorporating a domain classifier, the model learns domain-invariant representations that generalize well across domains.

Xu *et al.* [75] outlined a novel federated adversarial domain hallucination (FADH) learning framework for FDG, making the final model more robust to unseen domain shifts. Zhang *et al.* [40] proposed FedADG that aligns the distributions among different source domains using federated adversarial

learning, resulting in a universal feature representation with good generalization performance over unseen target domains while protecting local data privacy. Peng *et al.* [69] extends adversarial adaptation techniques to the constraints of the FL setting and leverages feature disentanglement to enhance knowledge transfer. Zhang *et al.* [86] introduces deep adversarial networks as a solution to effectively bridge the gap between domain distributions while maintaining data privacy. Micaelli *et al.* [113] developed a training approach in which a student network learns to align its predictions with those of a teacher network, solely relying on an adversarial generator to discover images where the student exhibits poor alignment with the teacher and utilizing them for student training, without relying on any data or metadata. Zhang *et al.* [114] developed a knowledge-agnostic approach for on-device knowledge transfer by adversarially training a generative model with the global model to obtain higher accuracy and better generalization performance. Dalmaz *et al.* [115] proposed a novel specificity-preserving federated learning method for MRI contrast translation based on an adversarial model that adaptively normalizes the feature maps across the generator based on site-specific latent variables.

In summary, federated adversarial learning contributes to improved generalization performance while ensuring privacy preservation and data security in the federated learning setting.

Federated feature alignment: Federated feature alignment is a technique used in FDG to address the challenge of domain shift across different domains in an FL setting. It aims to align the feature representations of data samples from different domains to improve the generalization capability of the federated model.

Nguyen *et al.* [116] presented FedSR, a novel representation learning framework for FL, which enables domain generalization while preserving privacy and decentralization, and incorporates CMI and L2R regularization techniques to improve generalization performance. To align features [110] from two domains in federated learning without direct access to target data, a proxy is constructed using re-weighted mixup of server data, which effectively improves model generalization, robustness, and stability, and further sampling is performed based on client data density using Gaussian Mixture Models (GMM) to create a proxy set. Zhang *et al.* [117] proposed FedUFO method, an adversary module is introduced to minimize feature divergence among different clients, while consensus losses are employed to mitigate inconsistencies in optimization objectives, enabling unified feature learning and alignment in non-IID FL. Sun *et al.* [71] proposed FedKA, a novel FDG method that utilizes feature distribution matching and a federated voting mechanism to align domain-specific features and generate pseudo-labels for global model fine-tuning, enabling domain-invariant learning in the presence of unknown client data. To learn feature representations from raw input data, Kang [118] utilized two neural networks for each party, where one network captures weakly shared representations and the other learns domain-specific representations, and three loss terms are proposed to enforce the desired learning objectives.

Invariant risk minimization (IRM): IRM [119] is an innovative learning paradigm that aims to estimate nonlinear,

invariant, and causal predictors from multiple training environments in order to achieve robust generalization to out-of-distribution (OOD) data. In deep learning, IRM consists of a feature encoder ($\mathbf{g}$) and a classifier ($\mathbf{h}$). Here, domain i $\mathcal{S}^i = \{(x_j^i, y_j^i)\}_{j=1}^{n_i}$ is collected under multiple source domains $i \in I$. The formulation of IRM is to minimize the loss function:

$$\begin{aligned} & \min_{\mathbf{h}, \mathbf{g}} \sum_i \mathcal{R}^i(\mathbf{h} \circ \mathbf{g}), \\ & \text{subject to: } \mathbf{h} \in \arg \min_{\mathbf{h}'} \mathcal{R}^i(\mathbf{h}' \circ \mathbf{g}), \text{ for all } i \in I. \end{aligned} \quad (22)$$

where $\mathcal{R}^i(f) := \mathbb{E}_{x^i, y^i}[\ell(f(x^i), y^i)]$

The authors in [119] treat the classifier $\mathbf{h}$ as a fixed scalar, resulting in the optimization problem formulation as follows.

$$\min_{\mathbf{g}} \sum_i \mathcal{R}^i(\mathbf{g}) + \lambda \|\nabla_{\mathbf{h}}|_{\mathbf{h}=1.0} \mathcal{R}^i(\mathbf{h} \cdot \mathbf{g})\|^2 \quad (23)$$

Francis *et al.* [120] utilized IRM to learn invariant predictors across client domains, achieving optimal empirical risk on all participating clients. The central objective of the FedGen [121] framework is to learn invariant predictors by aggregating local models and masks from participating clients, utilizing a penalty term similar to IRM as a regularizer to discourage overfitting to specific data distributions and encourage generalization across distributions. Zhang *et al.* [122] introduced distributionally robust neural networks (DRNN), domain adversarial neural networks (DANN), and maximum mean discrepancy (MMD) feature learning, as well as referencing previous work on invariance methods such as correlation alignment (CORAL) and invariant risk minimization (IRM), providing a comprehensive evaluation of these approaches. Luca *et al.* [123] proposed a federated version of Invariant Risk Minimization (IRM) that follows a similar procedure to FedAvg, but differs from other federated IRM approaches (such as CausalFed and CausalFedGSD) that involve data communication either through intermediate representations or a shared subset of data, making our Fed-IRM distinct in terms of data sharing.

Overall, IRM could be used in FDG to address the problem of domain shift and ensure model robustness across different domains, which aims to learn representations that are invariant to domain-specific variations while still being informative for the task at hand.

kernel-based methods: Kernel-based methods provide a powerful framework to address domain shifts and improve generalization performance for FDG. These methods can capture complex relationships and align distributions between different domains by leveraging kernel functions and operating in high-dimensional feature spaces, leading to more robust and transferable models in FL settings.

Tian *et al.* [83] developed a novel domain generalization method based on gradient aggregation, leveraging the gradient as a kernel mean embedding to align distributions across domains and improve model generalization capability under privacy constraints, achieving better performance and generalization capability under the privacy-preserving condition. Different distributions of observed data from multiple

sources are modeled separately, and a kernel-based method is used to adaptively learn their similarities [124], resulting in an adaptive factor that measures the similarity between the distributions. In the collaborative learning setting among distributed clients facilitated by a central server, Salgia *et al.* [125] introduced a kernel-based bandit algorithm using surrogate Gaussian process models, achieving near-optimal regret performance, and demonstrate the effectiveness of using sparse approximations to reduce communication overhead among clients. Hong *et al.* [126] proposed eM-KOFL and pM-KOFL algorithms for Online Federated Learning achieve near-optimal performance while minimizing communication costs, with pM-KOFL demonstrating comparable performance to vM-KOFL (or eM-KOFL) across different online learning tasks.

b) Feature disentanglement-based FDG

Feature disentanglement is a widely adopted strategy for mitigating challenges related to domain shift and negative transfer when dealing with different domains, aiming to separate the domain-invariant features and domain-specific features from training samples [49], [50]. In FDG, feature disentanglement-based methods extend the principles of feature disentanglement to the FL setting and aim to learn disentangled representations by leveraging the distributed data from multiple source domains while ensuring privacy preservation and data locality.

Sun *et al.* [71] proposed Federated Knowledge Alignment (FedKA) that can learn domain-invariant client features under the constraint of unknown client data and improve the transferability of learned domain knowledge and alleviates negative transfer in Federated Learning (FL). Peng *et al.* [69] made significant contributions to the field by extending adversarial adaptation techniques to the unique constraints of FL, which leveraged feature disentanglement, effectively addressed the challenges posed by domain shift in a privacy-preserving manner. Wu *et al.* [41] optimized domain-invariant feature extractors for central aggregation and domain-specific classifiers for central ensembling, allowing for selective knowledge disentanglement between domain-invariant feature representations and domain-specific classification information.

By disentangling the features and capturing the domain-invariant factors, feature disentanglement-based FDG methods enable the trained model to generalize well to unseen domains by leveraging the shared knowledge across domains. These approaches contribute to addressing the challenges of domain shift and improving the transferability of models in FL scenarios.

2) Federated Domain Adaptation

Federated domain adaptation (FDA) is an approach that synergistically combines FL and DA techniques to mitigate the challenge of domain shift in a decentralized setting, thereby improving model performance when source and target domains exhibit dissimilar distributions or come from distinct domains.

a) Unsupervised domain adaptation

Unsupervised domain adaptation (UDA) is an ML technique that addresses the challenge of adapting models from a source domain with labeled data to a target domain with unlabeled data. In FDG, UDA techniques aim to learn domain-invariant

representations that capture shared knowledge across domains while reducing domain-specific differences. The goal is to bridge the domain gap and enable models to generalize well to unseen domains without access to labeled data in those domains.

Peng *et al.* [69] put forward a method for unsupervised FDA that aligns the learned representations among different domains with the data distribution of the target domains to solve the domain shift. Wu *et al.* [41] proposed a new approach called Collaborative Optimization and Aggregation (COPA) by employing hybrid batch-instance normalization and collaboration of frozen classifiers to optimize a generalized target model for decentralized DG and UDA. Zhuang *et al.* [127] introduced FedFR, a method for UDA in face recognition, which combines clustering-based DA and federated learning to improve performance on a target domain with different data distributions from the source domain. Zhou *et al.* [128] introduced STU-KD, a scheme for privacy-preserving adaptation of a compact model to edge devices, which combines DA techniques with knowledge distillation to efficiently transfer knowledge from a large target model to the compact model. The proposed DualAdapt method [110] is a FDA technique that enables the model to adapt to the target domains while preserving knowledge learned from the source domains, achieving high accuracy while minimizing communication costs and computational resources required on client devices. Niu *et al.* [129] introduced a novel federated UDA strategy termed knowledge filter to adapt the central model to the target data when unlabeled target data is available.

Overall, UDA in FDG plays a crucial role in addressing the domain shift problem and enhancing the transferability of models across federated domains, allowing them to perform well in unseen domains with limited labeled data.

b) Zero and few-shot adaption

Zero-shot and few-shot adaptation techniques address data scarcity and enhance model generalization, enabling models to perform effectively on tasks or domains with limited or no explicit training data, and improving adaptability to new scenarios. In FDG, zero-shot and few-shot adaptation techniques aim to enhance model generalization across distributed domains,

Yang *et al.* [111] introduced zero-shot adaptation with estimated statistics to solve the challenges of FDG, where the zero-shot adapter helps the learned global model to directly bridge a large domain gap between seen and unseen clients at inference time. Huang *et al.* [130] tackled the challenge of few-shot model-agnostic federated learning by introducing a framework that leverages public datasets to improve performance despite limited private data while addressing problems such as inconsistent labels and domain gaps through model-agnostic federated learning and latent embedding adaptation.

Zero-shot and few-shot adaptation techniques in FDG address domain shifts, and privacy concerns, improving the performance of FL in real-world scenarios with limited data access and distribution constraints.

c) Adversarial domain adaptation

Adversarial domain adaptation (ADA) can also be applied in FDG, where the goal is to learn models that can generalize

well across multiple domains without explicitly sharing data. In this scenario, each domain has its own data distribution and potentially different labeling conventions or data biases.

Zeng *et al.* [131] introduced a novel strategy called one-common-source ADA (OCS-ADA) to address domain shifts in each target domain (private data) by leveraging a common source domain (public data), which involves utilizing ADA with gradient matching loss to pre-train encoders. Zhao *et al.* [70] proposed a federated multi-source DA method combining transfer learning and federated learning for machinery fault diagnosis with data privacy, utilizing a federated feature alignment approach to minimize feature distribution differences and incorporating a joint voting scheme for fine-tuning the global model with pseudo-labeled samples, achieving accurate target data identification while protecting data privacy. Peng *et al.* [69] introduced a principled approach to FDA by aligning representations learned among distributed nodes with the target node's data distribution, incorporating adversarial adaptation techniques, dynamic attention mechanism, and feature disentanglement to enhance knowledge transfer and address the challenge of domain shift in federated learning. Zhang *et al.* [40] propose FedADG, a federated adversarial domain generalization approach that aligns distributions among different domains using a reference distribution and ensures a universal feature representation with good generalization while protecting local data privacy.

By incorporating ADA into FDG, models can learn domain-invariant representations that capture common features across multiple domains, enabling the final global model to generalize well to unseen domains, even in the absence of explicit data sharing.

d) Gradient matching FDA

The aim of Gradient matching [132], [133] in FDA is to align the gradients of the local models from different clients with the gradients of a reference model trained on a source domain. Through gradient matching, the proposed method incentivizes local models to acquire domain-invariant representations, promoting effective generalization to the target domain. The gradient matching loss L_{GM} is expressed as the expected cosine distance between $\dot{g}_s(\vartheta)$ and $\dot{g}_t(\vartheta)$, considering all possible values of ϑ :

$$L_{GM} = \mathbb{E}_{\vartheta} \left[1 - \frac{\dot{g}_s(\vartheta)^T \dot{g}_t(\vartheta)}{\|\dot{g}_s(\vartheta)\|_2 \times \|\dot{g}_t(\vartheta)\|_2} \right] \quad (24)$$

where $\dot{g}_s(\vartheta)$ represents the expected gradient vector from the source domain, $\dot{g}_t(\vartheta)$ represents the expected gradient vector from the target domain. and the gradients are computed with respect to the compatibility model parameters ϑ . $\|\cdot\|$ represents the Euclidean norm and $\mathbb{E}_{\vartheta}$ denotes the expectation over the compatibility model parameters ϑ . Minimizing the gradient matching loss L_{GM} aims to align the gradient vectors between the source and target domains, facilitating domain-invariant representation learning and improving the generalization capability across federated domains.

Zeng *et al.* [131] proposed a method called gradient matching FDA (GM-FDA), which aims to minimize domain discrepancy by utilizing a public image dataset and training resilient local federated models specifically for target domains.

Jeong *et al.* [134] proposed a novel algorithmic framework called Federated Gradient Matching Pursuit (FedGradMP) to solve the sparsity-constrained minimization problem in the FL setting. The proposed solution [135] aims to address the domain shift problem in FL by leveraging gradient covariance and the geometric mean of Hessians to capture both inter-silo and intra-silo consistencies of environments.

This technique improves the adaptation capability of FL models by explicitly considering the domain shift during the training process.

3) Federated Transfer Learning

Federated transfer learning [44] (FTL) is a valuable approach that can be utilized in scenarios where two datasets exhibit differences not only in terms of samples but also in their feature spaces. It allows models to leverage knowledge learned from one dataset to improve the performance on a different dataset with distinct features while preserving data privacy and security. FTL can also be applied to FDG, where the goal is to improve the generalization performance of models across multiple federated datasets that exhibit domain shifts and variations. The objective of FTL for FDG is as follows:

$$\Theta^* = \arg \min_{\Theta} \sum_{i=1}^n \delta_i * \mathcal{L}(\Theta, D_i) + \beta * F(\Theta) + \sigma * \Omega(\Theta) \quad (25)$$

where Θ represents the model parameters to be optimized, Θ^* represents the optimal model parameters, n is the number of clients in the FL system, $\mathcal{L}(\Theta, D_i)$ is the loss function that measures the discrepancy between the predicted outputs of the model and the ground truth labels on client i 's data D_i , δ_i is the weight assigned to the loss function $\mathcal{L}(\Theta, D_i)$ for client i , β is a regularization hyperparameter, $F(\Theta)$ is the regularization term that encourages model parameter smoothness or complexity control, σ is a hyperparameter that controls the trade-off between the regularization term and the domain generalization term, $\Omega(\Theta)$ is the domain generalization term that aims to minimize the distribution discrepancy between different clients' data distributions, promoting generalization across domains.

Chen *et al.* [82] developed a federated transfer learning framework D-WFA, incorporating discrepancy-based weighted federated averaging, to collaboratively train a global diagnosis model using local labeled source domain data and unlabeled target domain data, while ensuring data privacy. Zhang *et al.* [86] proposed a federated transfer learning method for fault diagnosis, addressing the issue of data privacy and domain shift phenomenon by using different models for different users and implementing a federal initialization stage and a federated communication stage using deep adversarial learning. Zhang *et al.* [96] introduced a federated transfer learning method for machinery fault diagnostics, utilizing deep learning-based models trained locally at each client while ensuring data privacy, and proposing a novel DA approach to bridge domain gaps without sharing local data, showing promising results for real industrial scenarios. Li *et al.* [136] proposed the FedSWP framework, a federated transfer learning-enabled smart work packaging (SWP) system for protecting the personal image information of construction workers in occupational health

and safety (OHS) management, which achieves accurate and personalized safety alerts and healthcare while addressing privacy concerns and paving the way for broader applications in construction OHS. Zhou *et al.* [137] introduced a novel bearing faulty prediction method based on federated transfer learning and knowledge distillation, utilizing signal-to-image conversion, multi-source federated transfer learning, and multi-teacher-based knowledge distillation to enhance generalization capability and achieve higher accuracy with lower parameter size in bearing fault prediction tasks. Fan *et al.* [138] presented IoTDefender, a personalized and distributed intrusion detection framework for 5G IoT that leverages federated transfer learning and 5G edge computing, addressing challenges of heterogeneity, data isolation, and limited data, achieving high detection accuracy, generalization capability and lower false positive rates while preserving privacy. Chen *et al.* [139] presented FedHealth, an FTL framework for wearable healthcare, addressing challenges of data isolation and lack of personalization by performing data aggregation through federated learning and building personalized models, demonstrating accurate and privacy-preserving healthcare in wearable activity recognition and Parkinson's disease diagnosis.

By leveraging transfer learning techniques within a federated learning setting, models can learn from multiple domains while preserving data privacy, enabling improved generalization capabilities and robustness in diverse federated environments.

4) Channel Decoupling

Channel decoupling is a technique used in ML, particularly in DG, to disentangle domain-specific information from shared information in feature representations. It can be used in FDG to disentangle domain-specific knowledge from shared knowledge by decoupling channel-wise statistics, enabling better generalization across unseen domains by reducing domain-specific variations.

Yang *et al.* [111] proposed client-agnostic learning with mixed instance-global statistics, in which local models learn client-invariant representations by applying client-agnostic objectives with the augmented data. Feng *et al.* [140] put forward the FedMRI algorithm by dividing the model into a shared encoder and client-specific decoder, enabling the preservation of domain-specific properties while maintaining a generalized representation, thereby addressing the issue of domain shifts. Shen *et al.* [141] proposed CD2-pFed framework, which utilizes cyclic distillation-guided channel decoupling to personalize the global model in federated learning, addressing the challenges of heterogeneous data distributions and achieving improved generalization across participating clients. Tan *et al.* [77] FedStar, a federated graph learning framework that addresses the non-IID problem in real-world graph data by extracting and sharing common underlying structure information through structure embeddings and independent structure encoders, enabling the capture of structure-based domain-invariant knowledge while avoiding feature misalignment issues. Weng *et al.* [142] proposed FedUCC, a federated unsupervised cluster-contrastive learning method for Person ReID, which utilizes a three-stage modeling strategy

to discover generic knowledge, specialized knowledge, and patch knowledge while ensuring cross-domain consistency and retaining local domain-specific knowledge.

By assigning specific channels to each domain, the global model can learn domain-specific representations, effectively addressing the challenges posed by heterogeneous data distributions across participating domains. Channel decoupling allows for fine-grained adaptation and enhances the model's ability to generalize across diverse domains in FL.

5) Style Transfer

Style transfer in FDG refers to the process of transferring the style or visual characteristics of a source domain to a target domain in the FL setting. The goal is to generalize the learned style representations across domains while preserving the content of the target domain.

Nergiz *et al.* [143] introduced a federated adaptation of the neural style transfer algorithm as a data augmentation technique to address the issues of data collaboration in computational pathology, specifically focusing on the highly class-imbalanced Chaoyang colorectal cancer imaging dataset, while ensuring privacy preservation without any data leakage. Chen *et al.* [91] introduced a domain generalization method called cross-client style transfer (CCST) for image recognition in FL, which surpasses state-of-the-art methods on DG while being compatible with existing DG techniques. Shenaj *et al.* [90] introduced FFREEDA, a task for Semantic Segmentation in autonomous driving, proposing the LADD algorithm that utilizes pre-trained model knowledge through self-supervision and a federated clustered aggregation scheme based on clients' style, demonstrating superior performance compared to existing approaches. Fantauzzo *et al.* [144] presented FedDrive, a benchmark for FL in Semantic Segmentation for autonomous driving, addressing challenges of statistical heterogeneity and domain generalization, evaluating state-of-the-art algorithms with style transfer methods to enhance generalization capabilities.

It's important to note that style transfer for FDG is an evolving research area, and the specific techniques and methodologies may vary depending on the latest advancements.

6) Federated Meta Learning

Model-agnostic meta-learning (MAML) [145] is a meta-learning approach that aims to learn an initial model that can quickly adapt to new tasks with limited labeled data. In FDG, MAML can be applied to address the challenge of generalizing across different federated domains with varying data distributions. The goal of MAML is to address domain shift by learning a meta-level optimization process that can quickly adapt a model to new tasks with limited data. Instead of training a model from scratch for each new task, MAML algorithms aim to learn a good initialization or set of initial parameters that can be fine-tuned or adapted to new tasks with minimal data. The specific form of the objective function may vary depending on the MAML algorithm used, but in general, it involves a two-level optimization process: inner-loop optimization and meta-optimization. In the inner-loop optimization, for each task or episode, the model is trained on a support set of labeled examples specific to that task. The goal is to update the model parameters to minimize task-

specific loss or error on the support set. This is typically done through gradient descent or other optimization techniques. In the meta-optimization, the meta-parameters of the model are updated based on the performance of the model on the query sets of multiple tasks. The objective is to find meta-parameters that can generalize well across tasks and enable fast adaptation. The meta-parameters are updated using the gradients computed from the meta-objective, which is often defined as the average loss or error across the query sets of multiple tasks.

a) Inner-loop optimization:

The base learner is fine-tuned using support samples D_{S_i} from the learnable initialization Φ for a fixed number of weight updates via gradient descent. The task adaptation objective is to minimize the loss function $L(D_{S_i}, \Phi_{i,j})$ using gradient descent. At the j -th step of inner-loop optimization, the base learner parameters are updated as:

$$\Phi_{i,j+1} = \Phi_{i,j} - \alpha \nabla_{\Phi_{i,j}} L(D_{S_i}, \Phi_{i,j}) \quad (26)$$

where α is the learning rate.

b) Inner-loop optimization iterations:

After performing J number of inner-loop update steps, the task-specific base learner parameters become $\Phi_{i,J}$.

c) Meta optimization:

The meta-learned initialization Φ is evaluated based on the generalization performance of the task-specific base learner with parameters Φ_i (or $\Phi_{i,J}$) on unseen query examples D_Q . The objective of MAML is to minimize the meta-learning algorithm's objective function, denoted as $L(D_Q, \Phi_i)$, as follows:

$$\Phi = \Phi - \zeta \nabla_{\Phi} L(D_Q, \Phi_i) \quad (27)$$

where ζ is the meta-learning rate.

Chen *et al.* [146] presented FedMeta, a federated meta-learning framework that effectively addresses statistical and systematic challenges in federated learning by reducing communication cost, improving convergence speed, increasing accuracy, and preserving user privacy, outperforming existing optimization algorithms. Wang *et al.* [147] drew inspiration from MAML and proposed two GraphFL methods to address the non-IID issue in graph data and tackle tasks involving new label domains by splitting each task into a labeled training set and a disjoint query set. Jiang *et al.* [148] presented a personalized FL approach using MAML algorithms, demonstrated the meta-learning interpretation of the widely used FedAvg algorithm, achieved a modified version to optimize the objectives of FL personalization to improve the generalization capability of the final global model. Lin *et al.* [80] introduced a platform-aided collaborative learning framework for real-time edge intelligence, utilizing a federated meta-learning approach to train a model across multiple source edge nodes and efficiently adapt it to new tasks at a target edge node with limited data samples. Fallah *et al.* [149] investigated the generalization properties of MAML algorithms for supervised learning and introduced a novel definition of stability that incorporates the number of tasks and samples per task, enabling a comprehensive analysis of MAML's generalization error in both seen and unseen test tasks. Li *et al.* [150]

presented a novel pancreas segmentation model utilizing a meta-learning strategy and latent-space feature flow generation, which effectively mitigates interference from background clutter and appearance-style discrepancies through a coarse-to-fine workflow, resulting in superior performance on three pancreas datasets and surpassing state-of-the-art generalization methods. Chen *et al.* [151] introduced FedMeta-FFD, a pioneering method that combines federated learning and meta-learning to address the few-shot fault diagnosis challenge in IIoT, enabling clients to leverage external datasets while training a global meta-learner, resulting in improved convergence speed and accuracy compared to existing techniques. Khodak *et al.* [152] presented a theoretical framework that combines task-similarity, online convex optimization, and sequential prediction algorithms to design and analyze practical meta-learning methods, leading to improved performance in few-shot learning and federated learning through adaptive learning of task-similarity, enhanced transfer-risk bounds, and average-case regret bounds in dynamic task-environments or tasks.

By utilizing MAML, the model parameters can be updated on meta-training data, which consists of raw input data from different federated domains, and then evaluated on held-out meta-test data generated from frequency space with different distributions. In addition to the above MAML methods, there are other approaches being explored in the field of federated meta-learning for FDG. One is episodic learning, which can be applied in FDG to improve the transferability of models across different domains. Liu *et al.* [39], [153] introduced episodic meta-learning for training models in FL with varying data distributions by simulating domain shift explicitly and utilizing a meta-learning scheme to update model parameters with the objective of achieving desired model generalization in unseen domains. Another one is the meta-critic network, where the meta-critic network provides feedback on the performance of primary models across different domains, allowing for guided adaptation and improvement of their generalization capabilities. Chen *et al.* [78] proposed a personalized Federated optimization framework with Meta Critic (FedMC) to address the heterogeneity issues in federated networks, which efficiently captures robust and generalizable domain-invariant knowledge across clients.

7) Data-free Methods

Data-free [154] methods aim to perform learning tasks without using any labeled data from the target domains by leveraging auxiliary information, prior knowledge, or transfer learning techniques to generalize to unseen domains.

Zero and few-shot learning: Zero-shot learning and few-shot learning aim to recognize or classify objects from unseen domains for which no labeled samples are available during training.

Sun *et al.* [95] proposed Federated Zero-Shot Learning (FZSL) to transfer mid-level semantic knowledge and optimize a generalizable FL model. Sariyildiz *et al.* [103] proposed a generative model combined with Gradient Matching loss to address the challenges of zero-shot learning by transforming it into a supervised classification task, leading to substantial improvements in generalized zero-shot classification accuracy compared to previous methods. Micaelli *et al.* [113]

demonstrated the possibility of achieving zero-shot knowledge transfer through an adversarial approach, where the adversarial generator is trained to produce challenging images for a student network to match the predictions of a teacher network. Zhang *et al.* [114] introduced FedZKT, a data-free FL framework that enables devices to autonomously select on-device models based on their local resources, facilitating knowledge transfer among these heterogeneous on-device models through a zero-shot distillation technique. Li *et al.* [155] developed a federated zero-shot fault diagnosis framework that utilizes a semantic knowledge base, bidirectional alignment network, and cloud-edge collaboration to effectively diagnose both local and global unseen fault categories. Yuan *et al.* proposed a few shot

By incorporating zero-shot learning and few-shot learning techniques into FDG, models can improve their ability to generalize across multiple domains with limited labeled data.

Knowledge distillation: Data-free knowledge distillation (DFKD) [113], [156], [157] is a technique used to transfer knowledge from a large, well-trained model to a smaller, more compact model without the need for labeled training data.

Frikha *et al.* [154] presented DEKAN, an approach named Domain Entanglement via Knowledge Amalgamation from Domain-Specific Networks, which effectively extracts and combines domain-specific knowledge from existing teacher models to create a domain-shift-robust student model. Niu *et al.* [129] introduced a Mutually Collaborative Knowledge Distillation (MCKD) framework for both federated UDA and DG, in which domain-invariant representations are learned in a data-free manner by performing mutually collaborative knowledge distillation between separated local models. Zhu *et al.* [158] proposed a data-free knowledge distillation method for heterogeneous federated learning, which utilizes a lightweight generator to aggregate user information and improve the generalization performance of the global model with fewer communication rounds. Jeong *et al.* [159] developed data-free federated distillation (FD), a distributed model training algorithm for on-device ML that minimizes inter-device communication overhead and addresses the issue of non-IID data by introducing federated augmentation (FAug) to generate an IID dataset for improved performance. Zhang *et al.* [160] proposed FedFTG, a data-free knowledge distillation method that fine-tunes the global model in the server, effectively addressing data heterogeneity in FL and serving as a plugin to improve existing local optimizers.

DFKD can be leveraged successfully in FDG, wherein the student model is trained using multiple teachers that have been trained on diverse source domains, facilitating the generalization capability to previously unseen target domains.

Semantic Aggregation and Calibration: In FDG, semantic aggregation and calibration refer to techniques used to consolidate and refine semantic representations across multiple federated nodes in a privacy-preserving manner.

Yuan *et al.* [38] introduced Collaborative Semantic Aggregation and Calibration (CSAC), a novel privacy-preserving method that enables separated domain generalization by unifying multi-source semantic learning and alignment through an iterative process of data-free semantic aggregation and

cross-layer semantic calibration. Niu *et al.* [129] proposed a data-free semantic collaborative distillation strategy to learn domain-invariant representation for both federated UDA and DG.

By employing semantic aggregation and calibration techniques, FDG methods can effectively leverage shared knowledge across domains while adapting to the specific characteristics of individual domains, leading to improved generalization performance and robustness in FL scenarios.

8) Federated Domain Translation

Domain translation [161], [162] refers to the task of converting or adapting content from one domain to another domain while preserving its meaning and quality. Federated domain translation is a method that aims to perform translation tasks across multiple domains in a federated learning framework while preserving data privacy and security.

Iterative naive barycenter [163] for FDG aims to find a common representation or domain-invariant subspace by iteratively averaging the distributions or representations from multiple source domains. Zhou *et al.* [164] proposed a federated domain translation method that generates pseudo-data for each client which could be useful for multiple downstream learning tasks. Wang *et al.* [165] introduced FedMed-GAN, a benchmark for federated domain translation, which effectively addresses mode collapse, maintains generator performance, and demonstrates adaptability to different data proportions and variation.

By translating data or knowledge across domains in the FL setup, models can benefit from a more diverse and comprehensive training dataset, leading to improved performance and generalization in real-world applications.

9) Regularization Strategies

Regularization strategies in FDG aim to improve the generalization performance of models across multiple domains by mitigating the negative effects of domain shift and improving domain-invariant representations.

To combat overconfidence and dropping learning curves in pseudo-labeling for FDG, Shenaj *et al.* [90] proposed using Knowledge Distillation (KD) loss based on soft predictions and Stochastic Weight Averaging (SWA) applied to clients' teachers to stabilize the learning curve and prevent the model from forgetting pre-training knowledge. Zhuang *et al.* [127] presented a novel domain constraint loss (DCL) in FL to regularize source domain training where DCL serves as a regularization term during the training of the source domain model, promoting alignment towards the target domain within the aggregated global model. Zhu *et al.* [135] proposed a new regularization methodology called Federated Invariant Learning Consistency (FedILC) to promote both inter-silo and intra-silo consistencies in federated networks by leveraging the domain-level gradient covariance and the geometric averaging of Hessians. In addition, the authors also introduced a novel weighted geometric mean mechanism to calculate geometric mean with inconsistent signs on gradients. Atwany *et al.* [88] employed flatness in the training of separated domain generalization, incorporating iteration-wise weight averaging and domain-level gradient variance regularization as effective techniques. Feng *et al.* [140] developed a weighted contrastive

regularization method for the globally shared part of the model, which helps alleviate domain shifts among clients during training and improves convergence. Liu *et al.* [39] employed the InfoNCE objective to impose regularization, using positive and negative feature pairs, to enhance the domain-invariance and discriminability of the features in their federated learning model, ultimately improving its performance.

Overall, regularization strategies play a crucial role in FDG by mitigating overfitting and improving the generalization performance of models across different unseen domains.

D. Aggregation Optimization Algorithms

a) Weight Optimization

The FedAvg algorithm in FL aggregates global model parameters based on the number of local model parameters contributed by each client, disregarding the quality and importance of the local model parameters, which can adversely impact the performance of the global model. In FDG, optimizing aggregation weights is crucial to address the challenges posed by heterogeneous data distributions and non-uniform network architectures across participating clients.

Chen *et al.* [82] proposed D-WFA, a dynamic weighting strategy based on MMD, adjusting the weights of clients and updating multiple local models with generalization capability. Zhang *et al.* [166] presented Generalization Adjustment (GA) method in FL dynamically calibrates aggregation weights to optimize the objective, achieving a tighter generalization bound by explicitly re-weighting aggregation instead of relying on implicit multi-domain data sharing as in conventional DG settings. Yuan *et al.* [38] introduced a semantic similarity-based attention mechanism that dynamically assigns weights to cross-layer pairs, prioritizing pairs with higher semantic similarity for precise alignment while reducing the emphasis on less similar pairs to prevent semantic dislocation. Yang *et al.* [111] proposed to dynamically generate instance-wise interpolation parameters for mixing instance and global statistics with a learning-based network to build the global model deployed to unseen clients.

Optimizing aggregation weights in FDG enables effective integration of client contributions, accounting for their importance and relevance, resulting in a more adaptive and robust global model capable of performing well across diverse data distributions and network architectures.

b) Gradient Optimization

Gradient optimization in FL refers to the process of optimizing the model parameters by leveraging the gradients computed on local participant data. Gradient optimization plays a crucial role in FDG by enabling efficient parameter updates across decentralized domains, ensuring convergence to an optimal global model while considering the variations in data distributions and network architectures, thereby enhancing the model's generalization capability and performance across different domains and scenarios.

Federated voting aims to fine-tune the global model in federated learning without using the ground labels of the target domain samples, thus improving the effectiveness of feature distribution matching [71]. In the global mutual optimization

[167], each domain undergoes model distillation to learn from its peers, effectively mitigating any potential negative impacts of domain-specific gradient updates on the overall global model. To enhance the robustness of the global feature encoder across clients, a domain critic [78] is introduced to regularize the feature representations learned by both the local and global feature encoders, utilizing the Wasserstein critic as the chosen distance metric due to its gradient superiority. Tian *et al.* [83] proposed to improve the gradient aggregation process on the centralized server side with a novel gradient alignment loss, expecting that the trained model can be better generalized to the unseen but related domains. Zeng *et al.* [131] developed the Gradient Matching Federated (GM-Fed) fine-tuning method, which updates pre-trained local federated models using the One-Common-Source Adversarial Domain Adaptation (OCS-ADA) strategy, minimizing the gradient matching loss between sites to drive the optimization direction towards specific local minima. Wu *et al.* [41] introduced a prediction agreement mechanism to overcome local disparities towards central model aggregation for achieving better-decentralized domain generalization.

In summary, optimizing aggregated gradients in FDG is essential to address domain shift, accommodate data heterogeneity, and improve model performance across different domains.

c) Aggregation Policy:

The aggregation policy in federated learning refers to the strategy or algorithm used to combine the model updates from multiple participating clients and generate an updated global model. In FDG, designing an effective aggregation policy is essential for achieving accurate and robust models.

Shenaj *et al.* [90] introduced a clustered and layer-aware aggregation policy where global parameters are aggregated over selected clients and cluster-specific parameters are averaged within clusters, where clients are clustered based on their transferred styles using H-Means clustering.

Optimizing the aggregation policy effectively helps to combine and utilize knowledge from diverse domains, ensuring the generation of a robust and adaptable global model capable of performing well across different domains and scenarios.

IV. DATASETS AND APPLICATIONS

In this section, we first provide a summary of the existing commonly used datasets in FDG. Following this, we discuss the popular tasks/applications for this research area.

A. Datasets

Table IV provides an overview of several commonly used datasets in the field of FDG. These datasets have been extensively utilized in various FDG applications, including handwritten digit recognition, object recognition, transportation imaging, medical imaging, person re-identification, industrial edge intelligence, and human activity recognition. They serve as valuable resources for evaluating and benchmarking the performance of FDG algorithms and methodologies. Here, we provide a detailed description of several widely used datasets commonly employed in research studies.

TABLE IV
DATASETS FOR FDG.

Area & Dataset	#Domain	#Class	#Sample	Description	Ref.
Handwritten digit recognition					
<i>Digit</i>	4	10	203,695	MNIST, MNIST-M, SVHN, USPS	[154]
<i>Digits-DG</i>	4	10	219,289	MNIST, MNIST-M, SVHN, SYN	[41]
<i>Rotated MNIST</i>	6	10	70,000	M0, M15, M30, M45, M60, M75	[38], [116], [123], [167]
<i>Digit-Five</i>	5	10	215,695	MNIST, MNIST-M, Synthetic Digits, SVHN, USPS	[71], [129]
Object recognition					
<i>Office-Home</i>	4	65	15,588	Art, Clipart, Product, Real-world	[40], [41], [91], [111], [116], [123], [128], [129], [167], [168]
<i>PACS</i>	4	7	9,991	Art, Cartoon, Photos, Sketches	[38], [40], [41], [91], [111], [116], [123], [129], [154], [167], [168]
<i>VLCS</i>	4	5	39,305	Ascal, LabelMe, Caltech, Sun	[38], [40], [111], [168]
<i>DomainNet</i>	6	345	586,575	clipart, infograph, painting, quickdraw, real, sketch	[116]
<i>Office-31</i>	3	31	4,652	Amazon, Webcam, DSLR	[128]
<i>Office-Caltech-Home</i>	7	65	17,964	Amazon, Webcam, Caltech, Art, Clipart, Product, Real-World	[38]
<i>Office-Caltech10</i>	4	10	2,533	Caltech, Amazon, Webcam, DSLR	[71], [129]
<i>ImageCLEF-DA</i>	3	12	1,800	ImageNet ILSVRC 2012, Caltech-256, Pascal VOC 2012	[128]
<i>Amazon Review</i>	4	2	7,671	Books, DVDs, Electronics, Kitchen & housewares	[71]
Transportation imaging					
<i>Cityscapes</i>	50	19	3,475	Real photos taken in the streets of 50 different cities	[90], [144]
<i>IDDA</i>	105	16		A synthetic dataset in the field of self-driving cars	[144]
<i>Mapillary Vistas</i>		66	25,000	The dataset covers diverse scenes from different domains and cities around the world	[90]
<i>IDDA</i>	105	16		A synthetic dataset in the field of self-driving cars	[144]
Medical imaging					
<i>Skin Lesion</i>	7	7		HAM10000, Dermofit, Derm7pt, MSK, PH2, SONIC, UDA	[83]
<i>ISIC 2018</i>	7	7	10,015	Melanocytic Nevus, Melanoma, Benign Keratosis, Basal Cell Carcinoma, Actinic Keratosis, Vascular Lesion, Dermatofibroma	[169]
<i>Dermofit Image Library</i>	10	10	1,300	A comprehensive collection of dermatology images for study, analysis, and diagnosis of skin conditions	[169]
<i>YawDD</i>	2	3		A dataset of videos, recorded by an in-car camera	[136]
Person Re-identification					
<i>Person Re-ID dataset</i>	8	7,270	93,796	DukeMTMCReID, Market1501, CUHK03-NP, PRID, CUHK01, VIPeR, 3DPeS, iLIDt.	[142]
Industrial Edge Intelligence					
<i>CWRU</i>	4	10	900	A widely used dataset for fault diagnosis and condition monitoring of rotating machinery	[72], [86], [96], [151]
<i>PU</i>	/	10	52,497	A 6203 bearing dataset, obtained from Paderborn University, includes artificially induced and real damages.	[151]
Human Activity Recognition					
<i>UCI Smartphone</i>	30	6	10,299	Walking, Walking_upstairs, walking_downstairs, Sitting, Standing, Laying	[139]

Rotated MNIST is a modified version of the MNIST dataset that introduces rotation transformations to the original handwritten digit images (M0, M15, M30, M45, M60, M75), serving as a benchmark dataset for evaluating the robustness and generalization ability of ML algorithms and models in the context of image recognition and computer vision tasks. *PACS*

is a demanding dataset for domain generalization, comprising seven distinct object categories from four diverse domains (Art Painting, Cartoon, Photo, and Sketch), challenging models to effectively generalize across significant domain discrepancies and variations in artistic styles, renders techniques, and visual representations. *Office-Home*, encompassing four distinct do-

mains (Artistic, Clipart, Product, and Real-World), has gained widespread usage in the fields of domain adaptation and transfer learning, serving as a valuable resource for evaluating models' ability to generalize across diverse visual domains that encompass stylized and real-world images. *Digits-DG* is a notable dataset specifically designed for domain generalization in digit recognition tasks incorporating four distinct digit datasets (MNIST, MNIST-M, SVHN, and SYN), introducing variations in font styles and backgrounds to simulate real-world scenarios and challenging models to generalize across diverse writing styles, backgrounds, and digit representations. *VLCS* is a comprehensive benchmark dataset designed to bridge the gap between vision and language understanding tasks by providing a rich collection of visual scenes accompanied by linguistic descriptions (Ascal, LabelMe, Caltech, Sun), enabling to explore and develop models that integrate visual perception and linguistic reasoning. *CWRU* is a widely recognized benchmark in ML and fault diagnosis, comprising vibration signals from diverse rotating machinery under different operating conditions and fault scenarios, providing researchers with a valuable resource for developing and evaluating fault diagnosis and condition monitoring algorithms.

B. Applications

FDG has broad applications in scenarios where data is distributed across multiple domains (organizations), and data sharing is limited due to privacy concerns or regulatory reasons. Here, we will discuss some of the popular tasks/applications for FDG.

1) Healthcare

In the field of healthcare, each hospital collects medical data from its patients, and each hospital has its own patient population. The data collected may differ in terms of demographics, medical conditions, and treatment protocols. However, the goal is to develop a model that can accurately predict disease outcomes for patients from all hospitals, regardless of the differences in data distribution. The FDG approach can be used to address this challenge by training models locally at each hospital while preserving data privacy [75], [170]–[172]. The locally trained models can then be combined to obtain a more generalized model that predicts well on any unseen domain.

2) Finance

The finance industry deals with sensitive and confidential data and the data is often distributed among multiple institutions, such as banks, insurance companies, and payment processors. Federated learning is of great importance in financial applications because it addresses data privacy concerns, enables cross-silo collaboration, improves the accuracy of the model, and provides better financial services. While federated learning in finance offers significant benefits in terms of data privacy and cross-institutional collaboration, it is true that the federated model may face challenges when applied to unseen finance domains. FDG plays a crucial role in finance to address the above issue by providing financial distress predictions [173], risk analysis [174], [175], and fault detection [176]. The adoption of FDG in the finance industry contributes to

the advancement of more precise, flexible, and comprehensive models tailored for a range of finance applications, thereby yielding mutual benefits for financial institutions and their clients.

3) Education

Educational institutions are increasingly engaging in international collaborations, allowing students to connect and learn from peers in different countries, promoting cultural understanding and global perspectives. However, educational institutions must adhere to privacy and data protection regulations when collecting, storing, and analyzing student data. FDG could be employed to achieve international collaborations in educational institutions while preserving data privacy, such as grades classification [177], digital educational environment [178]. Introducing FDG in educational institutions helps to understand data privacy protection, handle data distribution differences, strengthen transfer learning, foster innovation and research, integrate theory with practice, and develop the ability to tackle the challenges of training and generalizing models in distributed data environments.

4) Transportation

The transportation sector plays a vital role in facilitating the movement of people and goods within and between cities, regions, and countries. Transportation data is typically collected and managed by transportation authorities, government agencies, transportation service providers, and technology companies, which is crucial to ensure the security and privacy of information. FDG in the transportation domain can facilitate data collaboration between cities and regions, support the design and optimization of intelligent transportation systems [179]–[181], autonomous drive [90], optimal transport [182], and enhance capabilities for traffic safety [179]. By applying FDG methods and techniques, the transportation domain can leverage distributed data for modeling and decision-making, leading to more intelligent, efficient, and sustainable transportation systems.

5) Natural Language Processing

In NLP, data plays a crucial role as it provides the foundation for building and training various NLP models and systems. It's worth noting that NLP models often require large and diverse datasets to achieve high performance and generalization. However, there exists data quality, data bias, insufficient data volume, challenging data annotation, and data privacy and ethical concerns in NLP, addressing them effectively is essential for developing high-quality and reliable NLP models. The introduction of FDG in NLP addresses domain shift, data distribution, privacy, and adaptability challenges by enhancing generalization capability, facilitating collaboration while maintaining data privacy, and enabling robust models for unseen domains [183]–[185].

6) Robotics

In general, robotic systems often involve sensitive data (i.e., such as images or sensor readings) provided via robots deployed in various locations or from their surrounding environments, which may pose privacy and security risks for centralized data transfer. The convergence of FDG and robotics has substantial potential to foster knowledge transfer, and facilitate cross-robot collaboration, resulting in improved

performance, adaptability, and privacy protection of robotic systems. FDG has paid attention to the robotic areas, such as learning autonomously [186] to smart robotic wheelchairs [187], learning basic table tennis movements from human inputs [188], cloud robotic systems [189], [190], object recognition and grasp planning in surface decluttering [191]. The significance of FDG for robotics lies in its ability to enhance the generalization capability and robustness of robotic systems across different environments, tasks, and robots, thereby improving their adaptability, performance, and privacy protection in real-world applications.

7) Industrial Edge Intelligence

Intelligent instrumentation and measurement technology has advanced rapidly which has resulted in a significant increase in industrial data. This has led to the inevitable arrival of the "era of industrial big data" [192], [193]. However, the utilization of industrial big data poses challenges in terms of data management, analysis, and privacy. FDG is crucial in the era of industrial big data as it enables scalable processing and analysis, ensures data privacy and security, facilitates domain-specific knowledge sharing, and supports real-time analytics. For instance, [82], [86], [194] utilize FDG techniques to solve the faulty diagnosis issue in industrial edge intelligence. Plus, [151] allows clients to leverage indirect datasets from collaborators, training a global meta-learner capable of addressing few-shot problems and adapting to new clients or fault categories with minimal labeled examples and iterations. By leveraging FDG, industrial organizations can effectively utilize and derive value from the wealth of data generated in the industrial environment.

8) Distributed Computing Continuum Systems

The emergence of edge computing as an extension of cloud computing has led to the development of Distributed Computing Continuum Systems (DCCS) [195]–[199], which leverage the virtually limitless resources of the cloud and the diversity and proximity of the edge. The current research in edge computing and distributed computing continuum tends to address specific issues, resulting in solutions with limited applicability. The traditional computer system architecture is insufficient for representing the DCCS, which consists of heterogeneous devices and networks. Since the functional requirements of these systems may evolve over time, services may be dynamically added or changed, or unexpected events may occur, resulting in alterations to the underlying infrastructure configuration. This necessitates a new and innovative representation of DCCS that is not limited by outdated architectural models. Furthermore, several research works emphasize the need to distribute functionalities and processes among computing continuum entities in a decentralized manner [200]–[205]. Therefore, this has opened the possibility of developing FDG in DCCS for maintaining data privacy, enabling cross-domain model generalization, and enhancing system performance and robustness in distributed computing environments.

V. EVALUATIONS AND BENCHMARKS

A. Evaluation

1) Evaluation Strategies

When evaluating FDG methods, diverse strategies can be employed to comprehensively assess the model's capacity to generalize effectively to previously unseen domains. Here is a summary of evaluation strategies commonly used for FDG:

- **Leave-one-domain-out cross-validation (LODO)** [39], [40], [206]: Leave-one-domain-out cross-validation is an evaluation strategy where one domain is held out during training, allowing the model's performance to be assessed on unseen domains and providing insights into its generalization capabilities across federated settings.
- **Test-domain validation set (TEVS)** [206]: Test-domain validation set refers to a subset of data from a specific domain that is used for model evaluation during the validation phase.
- **Training-domain validation set (TRVS)** [206]: Training-domain validation set refers to a subset of data from the same domain as the training data that is used for model validation during the training phase.

2) Evaluation Metrics

Evaluation metrics in FDG aim to assess the performance and generalization ability of models across different domains within the FL setting.

- **Accuracy:** Accuracy is a prevalent evaluation metric employed to quantitatively assess the performance of machine learning models. Except for average accuracy, previous work [207] follow [208] using the average accuracy among all classes and the average accuracy of the unknown classes in federated open-set DA:

The average accuracy among all classes (OS):

$$Acc(OS) = \frac{1}{\mathbb{K} + 1} \sum_{k=1}^{\mathbb{K}+1} \frac{|x^t \in \mathcal{D}_k^T \wedge \hat{y}_k = y_k|}{\mathcal{D}_k^T} \quad (28)$$

The accuracy of the unknown classes (OS):*

$$Acc(OS*) = \frac{1}{\mathbb{K}} \sum_{k=1}^{\mathbb{K}} \frac{|x^t \in \mathcal{D}_k^T \wedge \hat{y}_k = y_k|}{\mathcal{D}_k^T} \quad (29)$$

The average accuracy among known classes (UNK):

$$Acc(UNK) = \frac{|x^t \in \mathcal{D}_{\mathbb{K}+1}^T \wedge \hat{y}_{\mathbb{K}+1} = y_{\mathbb{K}+1}|}{\mathcal{D}_{\mathbb{K}+1}^T} \quad (30)$$

where $\hat{y}$ denotes the predicted result, $\mathcal{D}_k^T$ represents the set of target samples with the label y_k , $\mathcal{K} := \{1, \dots, k, \dots, \mathbb{K}\}$ and the unknown class is defined as class $\mathbb{K}$. However, accuracy may not be suitable for imbalanced datasets (Non-IID in FL) where the class distribution is skewed, as it can be influenced by the majority class. Therefore, it is important to consider other evaluation metrics.

- **Standard deviation (SD)** [111]: In FDG, the standard deviation can be used to assess the distributional differences or variations across different domains.
- **Dice coefficient(Dice)** [39]: The Dice coefficient, also known as the Sørensen-Dice coefficient, is a similarity

metric commonly used in image segmentation tasks and medical image analysis. It measures the agreement or overlap between two domains. The Dice coefficient is calculated as the ratio of twice the intersection of the two domains to the sum of the sizes of the individual domains:

$$Dice(\mathcal{P}, \mathcal{Q}) = \frac{2 * Comm(\mathcal{P}, \mathcal{Q})}{|\mathcal{P}| + |\mathcal{Q}|} \quad (31)$$

where $Comm(\mathcal{P}, \mathcal{Q})$ represents the number of common samples between the two domains, while $|\cdot|$ represents the total number of samples in each domain. The Dice coefficient is commonly used as an evaluation metric in tasks such as image segmentation, where it quantifies the accuracy and similarity of the segmented domains compared to a ground truth or reference segmentation. It provides a measure of the model's performance in accurately delineating domains of interest and is particularly useful in cases where class imbalance or uneven distribution of domains is present.

- **Hausdorff distance (HD) [39]:** Hausdorff distance is a metric used to measure the dissimilarity or distance between two domains of points, contours, or shapes. It quantifies the maximum distance between any point in one domain and its nearest point in the other domain. The Hausdorff distance is defined as follows:

$$\begin{aligned} HD(\mathcal{P}, \mathcal{Q}) &= \max(h(\mathcal{P}, \mathcal{Q}), h(\mathcal{Q}, \mathcal{P})) \\ h(\mathcal{P}, \mathcal{Q}) &= \max_{p \in \mathcal{P}} \{ \min_{q \in \mathcal{Q}} \|p - q\| \} \\ h(\mathcal{Q}, \mathcal{P}) &= \max_{q \in \mathcal{Q}} \{ \min_{p \in \mathcal{P}} \|q - p\| \} \end{aligned} \quad (32)$$

where $\mathcal{P}$ and $\mathcal{Q}$ are the two domains being compared, and $h(\mathcal{P}, \mathcal{Q})$ represents the directed Hausdorff distance from domain $\mathcal{P}$ to domain $\mathcal{Q}$. It is calculated by finding the maximum distance from each point in $\mathcal{P}$ to its nearest point in $\mathcal{Q}$. Similarly, $h(\mathcal{Q}, \mathcal{P})$ represents the directed Hausdorff distance from domain $\mathcal{Q}$ to domain $\mathcal{P}$.

- **Wasserstein Distance (WD) [164]:** Wasserstein Distance is a mathematical metric that quantifies the dissimilarity between two probability distributions. In FDG, the WD measures the dissimilarity or discrepancy between the probability distributions of different domains in the FL setting. The WD is calculated as follows:

$$WD = \frac{1}{M^2} \sum_m \sum_{m'} \widehat{WD}(x_m, f_{m' \rightarrow m}(x_m)) \quad (33)$$

where M denotes the number of domains, each $\widehat{WD}$ [209] is computed with the Sinkhorn algorithm.

- **Fréchet Inception Distance (FID) score [164]:** The FID score is a metric used to evaluate the quality of generative models, particularly in GANs. In FDG, it can be used as an evaluation metric to provide a quantitative measure of the dissimilarity between the generated samples from different domains. The FID is defined as follows:

$$FID = \frac{1}{M^2} \sum_m \sum_{m'} \widehat{FID}(x_m, f_{m' \rightarrow m}(x_m)) \quad (34)$$

The FID [210] could be utilized to assess GAN performance and effectively quantifies disturbance levels.

- **A-distance [207]:** A-distance (also known as the adversarial distance) is a metric used in domain generalization and DA tasks, which measures the discrepancy between probability distributions of different domains or datasets.

$$\hat{d}_A = 2(1 - 2\varepsilon) \quad (35)$$

where ε represents the generalization error of a two-sample classifier trained on the binary task of distinguishing input samples between the source and target domains.

- **The standard deviation of the mean Intersection over Union (mIoU) [144]:** By computing the mean and standard deviation of the mean IoU, the average performance and the variability in performance across multiple evaluations can be accessed. The mean provides an overall measure of accuracy, while the standard deviation gives an indication of the consistency or instability of the results.

1) Mean (mIoU): $mIoU = \frac{1}{N} * \sum_{i=1}^N (IoU_i)$.

where N is the total number of evaluations and IoU_i represents the Intersection over Union value for the i -th evaluation.

2) Standard Deviation: $SD = \sqrt{\frac{\sum_{i=1}^N ((IoU_i - mIoU)^2)}{N}}$.

3) The standard deviation of mIoU: $mIoU \pm SD$.

- **Group Effect (GE) [71]:** Group effect evaluates negative transfer caused by inefficient model aggregation in FL, aiming to quantify the impact of client data differences on diverse local model updates that ultimately lead to negative transfer in the parameter space.

$$\begin{aligned} TTA_f(\mathcal{G}_t) &= \frac{\sum_{(x,y) \in D_T} \mathbb{1}\{\arg \max_j \mathcal{F}(x; \mathcal{G}_t)_j = y\}}{|D_T|} \\ GE_t &= \frac{1}{M} \sum_{i \in 1, 2, \dots, M} TTA_f(\mathcal{G}_t + \Delta_t^{(i)}) + TTA_f(\mathcal{G}_{t+1}) \end{aligned} \quad (36)$$

where $\mathcal{F}(x)$ is the neural network classifier and $\mathcal{F}(x)_j$ denotes the j -th element of $\mathcal{F}(x)$, D_T represents the target domain dataset, $|D_T|$ denotes the size of the target domain dataset, $\mathcal{G}_t$ is the global model at time step t in the target task, $\Delta_t^{(i)}$ is the update of domain i .

B. Benchmarks

We present a comprehensive summary of existing benchmarks for evaluating the performance of FDG algorithms. Our analysis covers benchmarks in three categories: centralized domain generalization (DG), federated learning (FL), and benchmarks specifically designed for FDG. Also, the detailed summary of these evaluations and benchmarks can be found below and in Table V.

- **DANN [211]** is a neural network architecture designed to accomplish precise classification of source data while simultaneously learning feature representations that exhibit invariance across multiple source domains.
- **JiGen [212]** is a supervised framework that leverages jigsaw puzzles as a training task to learn effective generalization across diverse visual domains.

TABLE V
SUMMARY OF EVALUATIONS AND BENCHMARKS.

Paradigm	Method	Strategy	Metrics
DG	DANN [211]	TEVS	Accuracy
	JiGen [212]	LODO, TRVS	Accuracy
	Epi-FCR [26]	TEVS	Accuracy
	MTSSL [213]	LODO	Accuracy
	EISNet [214]	LODO, TRVS	Accuracy
	L2A-OT [17]	LODO	Accuracy
	DSO [34]	LODO	Accuracy
	Mixstyle [16]	LODO	Accuracy
FL	RSC [215]	LODO	Accuracy
	FedAvg [43]	TRVS	Accuracy
	FedProx [216]	TRVS	Accuracy
	Scaffold [217]	TRVS	Accuracy
FDG	Moon [73]	TRVS	Accuracy
	FedDG [39]	LODO	Dice, HD
	CSAC [38]	LODO	Accuracy
	FedADG [40]	LODO	Accuracy
	COPA [41]	LODO	Accuracy
	FedHealth [139]	TEVS	Accuracy
	FedIG(-A) [111]	LODO	Accuracy, SD
	FADH [75]	LODO	Accuracy
	CCST [91]	LODO	Accuracy
	FedINB [164]	TEVS	WD, FID
	FOSDA [207]	TEVS	OS, OS*, UNK
	FedDrive [144]	TEVS	mIoU $\pm$ SD
	FedKA [71]	LODO	GE

- **Epi-FCR [26]** is a scheme that learns domain shift using episodic training.
- **MTSSL [213]** is a method that facilitates the learning of transferable features by employing a self-supervised task focused on predicting Gabor filter bank responses.
- **EISNet [214]** is an innovative network that synergistically integrates self-supervised learning and metric learning approaches, effectively enhancing classifier performance specifically in target domains.
- **L2A-OT [17]** is a method that leverages synthetic data augmentation to learn domain-invariant features, facilitating effective generalization across different domains.
- **DSO [34]** is a novel scheme that effectively enhances the generalization performance on target domains by integrating batch normalization and instance normalization methods.
- **Mixstyle [16]** is an innovative method that combines features from different source domains to generate synthetic source domains, enabling optimization of model generalization in a comprehensive manner.
- **RSC [215]** is an approach that selectively discards dominant features present in the training data to optimize the generalization capability of a model.
- **FedAvg [43]** (Federated Averaging) is a distributed learning algorithm commonly used in FL settings without any generalization technique.
- **FedProx [216]** is a novel framework that addresses heterogeneity in federated networks by serving as a generalized and re-parametrized version of FedAvg.
- **Scaffold [217]** is a novel algorithm that utilizes control variates to effectively address 'client-drift' and minimize communication rounds, demonstrating resilience to data

heterogeneity and client sampling while leveraging data similarity for faster convergence.

- **Moon [73]** is an innovative FL framework, employing model-level contrastive learning to address the challenge of heterogeneity in local data distribution.
- **FedDG [39]** (FDG) is a solution that aims to learn a federated model such that it can directly generalize to completely unseen domains.
- **CSAC [38]** is a novel privacy-preserving method for the separated domain generalization task.
- **FedADG [40]** (Federated Adversarial Domain Generalization) is a scheme that utilizes federated adversarial learning to address the domain generalization problem in FL for IoT devices.
- **COPA [41]** (Collaborative Optimization and Aggregation) is a decentralized approach for domain generalization and multisource unsupervised DA, utilizing a collaborative optimization and aggregation process to construct a generalized target model without sharing data across domains.
- **FedHealth [139]** is a pioneering FTL framework for personalized wearable healthcare, leveraging FL and TL to overcome cloud-based personalization limitations.
- **FedIG(-A) [111]** is a novel approach that combines client-agnostic learning with mixed instance-global statistics for local training and zero-shot adaptation with estimated statistics for inference.
- **FADH [75]** (Federated Adversarial Domain Hallucina) is an innovative FDG approach that prioritizes domain hallucination to generate samples, optimizing the global model's entropy while minimizing the cross-entropy of the local model.
- **CCST [91]** is a novel domain generalization method for in FL, enabling style transfer across clients without data exchange to promote uniform source client distributions and mitigate model biases by aligning local models with the image styles of all clients.
- **FedINB [164]** is an innovative federated domain translation method that generates pseudodata specific to each client, offering potential benefits for multiple downstream learning tasks.
- **FOSDA [207]** (Federated OSDA) is an advanced federated algorithm that incorporates an uncertainty-aware mechanism to generate a global model by prioritizing source clients with high uncertainty while preserving high consistency.
- **FedDrive [144]** is an innovative benchmark framework that includes real-world challenges of statistical heterogeneity and domain generalization.
- **FedKA [71]** (Federated Knowledge Alignment) is an innovative FDG method that utilizes feature distribution matching and a federated voting mechanism to enable the global model to learn domain-invariant client features and refine its performance with target domain pseudo-labels.

VI. FUTURE RESEARCH DIRECTIONS

This paper offers an extensive and meticulous survey of the existing literature on FDG, highlighting its inherent advantages

as a fusion of federated learning and domain generalization, resulting in privacy-preserving, scalable, and diverse models capable of effective generalization across multiple domains. However, it is important to emphasize that despite these benefits, there remain numerous unresolved challenges that hinder the realization of FDG. To address this, a comprehensive identification and analysis of these challenges is crucial, accompanied by the pursuit of novel theoretical and technical solutions. Within this context, this study addresses prominent challenges in FDG, including privacy-preserving FDG, communication-efficient FDG, computation-efficient FDG, heterogeneity, label shift, scalable FDG, continuous FDG, and FDG to novel categories.

A. Privacy-preserving FDG

Federated learning already addresses privacy concerns by keeping data on local domains, but additional privacy-preserving techniques need to be developed to ensure that sensitive information is not leaked during the model aggregation process. The existing FDG methods usually neglect the research on privacy-preserving mechanisms [38], [39], [91], [111], [128], [164], [167]. Therefore, how to design robust and privacy-respecting models that can generalize well across diverse domains while protecting the sensitive information of individual clients or domains is worth investigating. Here are some approaches related to privacy-preserving FDG: Secure aggregation [218], differential privacy [219], privacy-aware model selection [220], privacy-preserving data preprocessing [221], and privacy regulations [222].

B. Communication-efficient FDG

Communication is a major bottleneck for many real-world applications in FL, as transmitting large model updates from multiple clients to a central server can be time-consuming and resource-intensive. Some work [41], [91] has demonstrated that certain proposed schemes in FDG can lead to increased communication costs. On the other hand, [164] has highlighted that neglecting communication limitations can result in poor model performance in FDG. In this case, it is of utmost importance to achieve communication efficiency in FDG, enabling faster and more scalable learning across diverse domains while minimizing the communication overhead and associated costs. Here are some strategies and techniques to achieve communication efficiency: Model compression and quantization [223], [224], differential updates, selective model aggregation [225], local adaptation [226], and communication-efficient aggregation algorithms [227].

C. Computation-efficient FDG

Similar to communication efficiency, Computation efficiency is also a crucial aspect of FDG, as it affects the speed and scalability of the learning process. While there are some studies in the field of FDG have indeed highlighted the need for increased computational resources [41], [111] and certain proposed methods may not be suitable where the clients have limited computational resource [39], [91], there

are only very few investigations on reducing the requirement for computational resource [152]. Here are some strategies and techniques to achieve computation efficiency: model architecture optimization [228], adaptive learning algorithms [229], communication optimization [223], [224], collaborative learning [230], efficient data sampling and preprocessing [231].

D. Heterogeneity

Heterogeneity in FL leads to performance disparities, communication inefficiency, privacy and security risks, bias and fairness issues, scalability limitations, and challenges in generalization across domains. In FDG, the participating clients or domains may have significant variations in terms of data distributions, feature representations, or label spaces, leading to a crucial challenge. Only a limited number of recent studies paid attention to this challenge, such as data and model heterogeneity challenges [78], [164], [232] and Non-IID [147]. Here are some approaches to solving heterogeneity in FDG: data preprocessing (feature scaling [233], data augmentation [94], [123], or DA [90], [234]), model adaptation [235], [236], model personalization [237], adaptive aggregation [238].

E. Label shift

Label shift in FDG refers to the situation where there is a discrepancy in label distributions between different domains in the FL setting, which could lead to performance disparities and hinder the overall generalization capability of the FL system. Therefore, addressing label shifts is crucial to ensure the robustness and accuracy of ML models in real-world scenarios. Very few recent works [75], [239] paid attention to Label shift. Here are some potential approaches to solve label shifts in FDG: domain adaptation [238] and transfer learning [240].

F. Scalable FDG

Scalability is a crucial aspect of FDG, as it ensures that the approach can effectively handle large-scale and diverse datasets from multiple domains. The work [38] pointed out that exploring scalable methods to implement FSG for large-scale datasets is very important. Up to now, only FADH [75] is proven to be scalable to different numbers of source domains. Here are some key considerations for achieving scalability in FDG: distributed computation [241], [242], communication efficiency [243], [244], model compression [245], model aggregation [246], and parallel processing [247].

G. Continuous FDG

Continuous FDG refers to the ability of the FL system to continuously adapt and generalize across multiple domains over time. However, the predominant research on FDG assumes fixed source domains and a one-time model learning process, which may not fully capture the dynamic nature of real-world scenarios. Only a limited number of investigations have focused on continuous FDG, and a recent study [248] utilizes incremental learning to address this challenge, mitigating catastrophic forgetting and enhancing generalization performance. Here are some key aspects and approaches related to continuous FDG: incremental learning [249], adaptive learning [226], DA [250], and lifelong learning [251], [252].

H. FDG to novel categories

FDG to novel categories refers to the ability of FL systems to generalize well to unseen or novel categories that are not present in the training data of participating domains. The existing FDG algorithms typically make an assumption that the label spaces (i.e., the set of possible categories) for different domains are identical. In reality, a more practical and general scenario arises when the source and target domains have both shared label spaces and private label spaces, resulting in an additional category gap between them [253]. Here are some key techniques that could be used for FDG to novel categories: transfer learning [254], meta-learning [116], or DA [207].

VII. CONCLUSION

FDG is a significant research area within machine learning, focusing on generalization learning and privacy-preserving in distributed scenarios. This paper offers a comprehensive analysis of the field of FDG, encompassing theoretical foundations, existing methodologies, available datasets, practical applications, evaluation strategies, evaluation metrics, and benchmarks. Additionally, through a meticulous analysis of these methods, several prospective research challenges are identified, which can pave the way for future investigations in the field. This survey aims to offer valuable insights to researchers and serve as a source of inspiration for future advancements in the field.

ACKNOWLEDGMENT

This work is supported by the National Natural Science Foundation of China under Grant No.92267206 and No. 62032013, the Science and Technology Planning Project of Shenyang City under Grant No. 21-108-9-19.

REFERENCES

- [1] K. Zhou, Z. Liu, Y. Qiao, T. Xiang, and C. C. Loy, "Domain generalization: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.
- [2] J. Quinero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence, *Dataset shift in machine learning*. Mit Press, 2008.
- [3] B. Recht, R. Roelofs, L. Schmidt, and V. Shankar, "Do imagenet classifiers generalize to imagenet?" in *International conference on machine learning*. PMLR, 2019, pp. 5389–5400.
- [4] R. Taori, A. Dave, V. Shankar, N. Carlini, B. Recht, and L. Schmidt, "Measuring robustness to natural distribution shifts in image classification," *Advances in Neural Information Processing Systems*, vol. 33, pp. 18 583–18 599, 2020.
- [5] G. Blanchard, A. A. Deshmukh, Ü. Dogan, G. Lee, and C. Scott, "Domain generalization by marginal transfer learning," *The Journal of Machine Learning Research*, vol. 22, no. 1, pp. 46–100, 2021.
- [6] G. Blanchard, G. Lee, and C. Scott, "Generalizing from several related classification tasks to a new unlabeled sample," *Advances in neural information processing systems*, vol. 24, 2011.
- [7] Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni, "Generalizing from a few examples: A survey on few-shot learning," *ACM computing surveys (csur)*, vol. 53, no. 3, pp. 1–34, 2020.
- [8] K. Muandet, D. Balduzzi, and B. Schölkopf, "Domain generalization via invariant feature representation," in *International conference on machine learning*. PMLR, 2013, pp. 10–18.
- [9] H. Li, S. J. Pan, S. Wang, and A. C. Kot, "Domain generalization with adversarial feature learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5400–5409.
- [10] Y. Li, X. Tian, M. Gong, Y. Liu, T. Liu, K. Zhang, and D. Tao, "Deep domain generalization via conditional invariant adversarial networks," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 624–639.
- [11] R. Shao, X. Lan, J. Li, and P. C. Yuen, "Multi-adversarial discriminative deep domain generalization for face presentation attack detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 10023–10031.
- [12] S. Erfani, M. Baktashmotlagh, M. Moshtaghi, X. Nguyen, C. Leckie, J. Bailey, and R. Kotagiri, "Robust domain generalisation by enforcing distribution invariance," in *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-16)*. AAAI Press, 2016, pp. 1455–1461.
- [13] S. Motiian, M. Piccirilli, D. A. Adjeroh, and G. Doretto, "Unified deep supervised domain adaptation and generalization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5715–5725.
- [14] C. Yoon, G. Hamarneh, and R. Garbi, "Generalizable feature learning in the presence of data bias and domain class imbalance with application to skin lesion classification," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part IV 22*. Springer, 2019, pp. 365–373.
- [15] X. Yue, Y. Zhang, S. Zhao, A. Sangiovanni-Vincentelli, K. Keutzer, and B. Gong, "Domain randomization and pyramid consistency: Simulation-to-real generalization without accessing target domain data," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 2100–2110.
- [16] K. Zhou, Y. Yang, Y. Qiao, and T. Xiang, "Domain generalization with mixstyle," *arXiv preprint arXiv:2104.02008*, 2021.
- [17] K. Zhou, Y. Yang, T. Hospedales, and T. Xiang, "Learning to generate novel domains for domain generalization," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16*. Springer, 2020, pp. 561–578.
- [18] R. Volpi and V. Murino, "Addressing model vulnerability to distributional shifts over image transformation sets," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 7980–7989.
- [19] S. Shankar, V. Piratla, S. Chakrabarti, S. Chaudhuri, P. Jyothis, and S. Sarawagi, "Generalizing across domains via cross-gradient training," *arXiv preprint arXiv:1804.10745*, 2018.
- [20] R. Volpi, H. Namkoong, O. Sener, J. C. Duchi, V. Murino, and S. Savarese, "Generalizing to unseen domains via adversarial data augmentation," *Advances in neural information processing systems*, vol. 31, 2018.
- [21] Z. Xu, D. Liu, J. Yang, C. Raffel, and M. Niethammer, "Robust and generalizable visual representation learning via random convolutions," *arXiv preprint arXiv:2007.13003*, 2020.
- [22] D. Li, Y. Yang, Y.-Z. Song, and T. Hospedales, "Learning to generalize: Meta-learning for domain generalization," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.
- [23] Y. Balaji, S. Sankaranarayanan, and R. Chellappa, "Metareg: Towards domain generalization using meta-regularization," *Advances in neural information processing systems*, vol. 31, 2018.
- [24] Y. Li, Y. Yang, W. Zhou, and T. Hospedales, "Feature-critic networks for heterogeneous domain generalization," in *International Conference on Machine Learning*. PMLR, 2019, pp. 3915–3924.
- [25] Q. Liu, Q. Dou, and P.-A. Heng, "Shape-aware meta-learning for generalizing prostate mri segmentation to unseen domains," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part II 23*. Springer, 2020, pp. 475–485.
- [26] D. Li, J. Zhang, Y. Yang, C. Liu, Y.-Z. Song, and T. M. Hospedales, "Episodic training for domain generalization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1446–1455.
- [27] Y. Zhao, Z. Zhong, F. Yang, Z. Luo, Y. Lin, S. Li, and N. Sebe, "Learning to generalize unseen domains via memory-based multi-source meta-learning for person re-identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 6277–6286.
- [28] Q. Dou, D. Coelho de Castro, K. Kamnitsas, and B. Glocker, "Domain generalization via model-agnostic learning of semantic features," *Advances in Neural Information Processing Systems*, vol. 32, 2019.
- [29] Q. Liu, Q. Dou, L. Yu, and P. A. Heng, "Ms-net: multi-site network for improving prostate segmentation with heterogeneous mri data," *IEEE transactions on medical imaging*, vol. 39, no. 9, pp. 2713–2724, 2020.
- [30] K. Zhou, Y. Yang, Y. Qiao, and T. Xiang, "Domain adaptive ensemble learning," *IEEE Transactions on Image Processing*, vol. 30, pp. 8008–8018, 2021.

- [31] Z. Xu, W. Li, L. Niu, and D. Xu, "Exploiting low-rank structure from latent domains for domain generalization," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part III 13*. Springer, 2014, pp. 628–643.
- [32] L. Niu, W. Li, and D. Xu, "Multi-view domain generalization for visual recognition," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4193–4201.
- [33] Z. Ding and Y. Fu, "Deep domain generalization with structured low-rank constraint," *IEEE Transactions on Image Processing*, vol. 27, no. 1, pp. 304–313, 2017.
- [34] S. Seo, Y. Suh, D. Kim, G. Kim, J. Han, and B. Han, "Learning to optimize domain specific normalization for domain generalization," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16*. Springer, 2020, pp. 68–83.
- [35] J. Cha, H. Cho, K. Lee, S. Park, Y. Lee, and S. Park, "Domain generalization needs stochastic weight averaging for robustness on domain shifts," *arXiv preprint arXiv:2102.08604*, vol. 3, 2021.
- [36] N. Truong, K. Sun, S. Wang, F. Guitton, and Y. Guo, "Privacy preservation in federated learning: An insightful survey from the gdp perspective," *Computers & Security*, vol. 110, p. 102402, 2021.
- [37] Y. Li, X. Wang, R. Zeng, M. Yang, K. Li, M. Huang, and S. Dustdar, "Varf: An incentive mechanism of cross-silo federated learning in mec," *IEEE Internet of Things Journal*, 2023.
- [38] J. Yuan, X. Ma, D. Chen, K. Kuang, F. Wu, and L. Lin, "Collaborative semantic aggregation and calibration for separated domain generalization," *arXiv e-prints*, pp. arXiv–2110, 2021.
- [39] 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*, 2021, pp. 1013–1023.
- [40] L. Zhang, X. Lei, Y. Shi, H. Huang, and C. Chen, "Federated learning for IoT devices with domain generalization," *IEEE Internet of Things Journal*, 2023.
- [41] 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*, 2021, pp. 6484–6493.
- [42] H. B. McMahan, E. Moore, D. Ramage, and B. A. y Arcas, "Federated learning of deep networks using model averaging," *arXiv preprint arXiv:1602.05629*, vol. 2, 2016.
- [43] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial intelligence and statistics*. PMLR, 2017, pp. 1273–1282.
- [44] Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated machine learning: Concept and applications," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 10, no. 2, pp. 1–19, 2019.
- [45] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings *et al.*, "Advances and open problems in federated learning," *Foundations and Trends® in Machine Learning*, vol. 14, no. 1–2, pp. 1–210, 2021.
- [46] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," *IEEE signal processing magazine*, vol. 37, no. 3, pp. 50–60, 2020.
- [47] K. Bonawitz, H. Eichner, W. Grieskamp, D. Huba, A. Ingerman, V. Ivanov, C. Kiddon, J. Konečný, S. Mazzocchi, B. McMahan *et al.*, "Towards federated learning at scale: System design," *Proceedings of machine learning and systems*, vol. 1, pp. 374–388, 2019.
- [48] G. Csurka, "Domain adaptation for visual applications: A comprehensive survey," *arXiv preprint arXiv:1702.05374*, 2017.
- [49] W. M. Koww and M. Loog, "A review of domain adaptation without target labels," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 3, pp. 766–785, 2019.
- [50] G. Wilson and D. J. Cook, "A survey of unsupervised deep domain adaptation," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 11, no. 5, pp. 1–46, 2020.
- [51] 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.
- [52] P. Sheth, R. Moraffah, K. S. Candan, A. Raglin, and H. Liu, "Domain generalization—a causal perspective," *arXiv preprint arXiv:2209.15177*, 2022.
- [53] Z. Shen, J. Liu, Y. He, X. Zhang, R. Xu, H. Yu, and P. Cui, "Towards out-of-distribution generalization: A survey," *arXiv preprint arXiv:2108.13624*, 2021.
- [54] H. Li, X. Wang, Z. Zhang, and W. Zhu, "Out-of-distribution generalization on graphs: A survey," *arXiv preprint arXiv:2202.07987*, 2022.
- [55] R. Zeng, S. Zhang, J. Wang, and X. Chu, "Fmore: An incentive scheme of multi-dimensional auction for federated learning in mec," in *2020 IEEE 40th international conference on distributed computing systems (ICDCS)*. IEEE, 2020, pp. 278–288.
- [56] B. Recht, R. Roelofs, L. Schmidt, and V. Shankar, "Do imagenet classifiers generalize to imagenet?" in *International conference on machine learning*. PMLR, 2019, pp. 5389–5400.
- [57] D. Hendrycks and T. Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," *arXiv preprint arXiv:1903.12261*, 2019.
- [58] J. Yang, K. Zhou, Y. Li, and Z. Liu, "Generalized out-of-distribution detection: A survey," *arXiv preprint arXiv:2110.11334*, 2021.
- [59] K. Saenko, B. Kulis, M. Fritz, and T. Darrell, "Adapting visual category models to new domains," in *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5–11, 2010, Proceedings, Part IV 11*. Springer, 2010, pp. 213–226.
- [60] Z. Lu, Y. Yang, X. Zhu, C. Liu, Y.-Z. Song, and T. Xiang, "Stochastic classifiers for unsupervised domain adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9111–9120.
- [61] K. Saito, K. Watanabe, Y. Ushiku, and T. Harada, "Maximum classifier discrepancy for unsupervised domain adaptation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3723–3732.
- [62] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *International conference on machine learning*. PMLR, 2015, pp. 1180–1189.
- [63] M. Long, Y. Cao, J. Wang, and M. Jordan, "Learning transferable features with deep adaptation networks," in *International conference on machine learning*. PMLR, 2015, pp. 97–105.
- [64] Z. Liu, Z. Miao, X. Pan, X. Zhan, D. Lin, S. X. Yu, and B. Gong, "Open compound domain adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 12406–12415.
- [65] B. Li, Y. Wang, S. Zhang, D. Li, K. Keutzer, T. Darrell, and H. Zhao, "Learning invariant representations and risks for semi-supervised domain adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1104–1113.
- [66] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan, "A theory of learning from different domains," *Machine learning*, vol. 79, pp. 151–175, 2010.
- [67] V. Vapnik, E. Levin, and Y. Le Cun, "Measuring the vc-dimension of a learning machine," *Neural computation*, vol. 6, no. 5, pp. 851–876, 1994.
- [68] I. Albuquerque, J. Monteiro, T. H. Falk, and I. Mitliagkas, "Adversarial target-invariant representation learning for domain generalization," *arXiv preprint arXiv:1911.00804*, vol. 8, 2019.
- [69] X. Peng, Z. Huang, Y. Zhu, and K. Saenko, "Federated adversarial domain adaptation," *arXiv preprint arXiv:1911.02054*, 2019.
- [70] K. Zhao, J. Hu, H. Shao, and J. Hu, "Federated multi-source domain adversarial adaptation framework for machinery fault diagnosis with data privacy," *Reliability Engineering & System Safety*, vol. 236, p. 109246, 2023.
- [71] Y. Sun, N. Chong, and H. Ochiai, "Feature distribution matching for federated domain generalization," in *Asian Conference on Machine Learning*. PMLR, 2023, pp. 942–957.
- [72] 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*, vol. 256, p. 109880, 2022.
- [73] Q. Li, B. He, and D. Song, "Model-contrastive federated learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 10713–10722.
- [74] Y. Guo, X. Tang, and T. Lin, "Feddebias: Reducing the local learning bias improves federated learning on heterogeneous data," 2022.
- [75] Q. Xu, R. Zhang, Y. Zhang, Y.-Y. Wu, and Y. Wang, "Federated adversarial domain hallucination for privacy-preserving domain generalization," *IEEE Transactions on Multimedia*, 2023.
- [76] N. Wu, L. Yu, X. Yang, K.-T. Cheng, and Z. Yan, "Federated learning with imbalanced and agglomerated data distribution for medical image classification," *arXiv preprint arXiv:2206.13803*, 2022.

- [77] Y. Tan, G. Long, J. Ma, L. Liu, T. Zhou, and J. Jiang, "Federated learning from pre-trained models: A contrastive learning approach," *arXiv preprint arXiv:2209.10083*, 2022.
- [78] C. Chen, T. Ye, L. Wang, and M. Gao, "Learning to generalize in heterogeneous federated networks," in *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 2022, pp. 159–168.
- [79] T.-A. Nguyen, T. D. Nguyen, L. T. Le, C. T. Dinh, and N. H. Tran, "On the generalization of wasserstein robust federated learning," *arXiv preprint arXiv:2206.01432*, 2022.
- [80] S. Lin, G. Yang, and J. Zhang, "A collaborative learning framework via federated meta-learning," in *2020 IEEE 40th International Conference on Distributed Computing Systems (ICDCS)*. IEEE, 2020, pp. 289–299.
- [81] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *The Journal of Machine Learning Research*, vol. 13, no. 1, pp. 723–773, 2012.
- [82] J. Chen, J. Li, R. Huang, K. Yue, Z. Chen, and W. Li, "Federated transfer learning for bearing fault diagnosis with discrepancy-based weighted federated averaging," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–11, 2022.
- [83] 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.
- [84] W. Zhang and X. Li, "Data privacy preserving federated transfer learning in machinery fault diagnostics using prior distributions," *Structural Health Monitoring*, vol. 21, no. 4, pp. 1329–1344, 2022.
- [85] W. Zhang, Z. Wang, and X. Li, "Blockchain-based decentralized federated transfer learning methodology for collaborative machinery fault diagnosis," *Reliability Engineering & System Safety*, vol. 229, p. 108885, 2023.
- [86] W. Zhang and X. Li, "Federated transfer learning for intelligent fault diagnostics using deep adversarial networks with data privacy," *IEEE/ASME Transactions on Mechatronics*, vol. 27, no. 1, pp. 430–439, 2021.
- [87] M. Duan, D. Liu, X. Chen, Y. Tan, J. Ren, L. Qiao, and L. Liang, "Astraea: Self-balancing federated learning for improving classification accuracy of mobile deep learning applications," in *2019 IEEE 37th international conference on computer design (ICCD)*. IEEE, 2019, pp. 246–254.
- [88] M. Atwany and M. Yaqub, "DRGen: Domain generalization in diabetic retinopathy classification," in *Medical Image Computing and Computer Assisted Intervention—MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part II*. Springer, 2022, pp. 635–644.
- [89] Y. Yang and S. Soatto, "Fda: Fourier domain adaptation for semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4085–4095.
- [90] D. Shenaj, E. Fani, M. Toldo, D. Caldarola, A. Tavera, U. Michieli, M. Ciccone, P. Zanuttigh, and B. Caputo, "Learning across domains and devices: Style-driven source-free domain adaptation in clustered federated learning," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 444–454.
- [91] 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*, 2023, pp. 361–370.
- [92] A. Georgiadis, V. Babbar, F. Silavong, S. Moran, and R. Otter, "ST-FL: style transfer preprocessing in federated learning for COVID-19 segmentation," in *Medical Imaging 2022: Imaging Informatics for Healthcare, Research, and Applications*, vol. 12037. SPIE, 2022, pp. 13–27.
- [93] D. Lewy, J. Mańdziuk, M. Ganzha, and M. Paprzycki, "StatMix: Data augmentation method that relies on image statistics in federated learning," in *Neural Information Processing: 29th International Conference, ICONIP 2022, Virtual Event, November 22–26, 2022, Proceedings, Part IV*. Springer, 2023, pp. 574–585.
- [94] W. Hao, M. El-Khamy, J. Lee, J. Zhang, K. J. Liang, C. Chen, and L. C. Duke, "Towards fair federated learning with zero-shot data augmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3310–3319.
- [95] S. Sun, C. Si, S. Gong, and G. Wu, "Federated zero-shot learning with mid-level semantic knowledge transfer," *arXiv preprint arXiv:2208.13465*, 2022.
- [96] J. Zhang and Y. Jiang, "A data augmentation method for vertical federated learning," *Wireless Communications and Mobile Computing*, vol. 2022, pp. 1–16, 2022.
- [97] W. Guo, Y. Wang, X. Chen, and P. Jiang, "Federated transfer learning for auxiliary classifier generative adversarial networks: framework and industrial application," *Journal of Intelligent Manufacturing*, pp. 1–16, 2023.
- [98] J. Chen, J. Zhang, Y. Zhao, H. Han, K. Zhu, and B. Chen, "Beyond model-level membership privacy leakage: an adversarial approach in federated learning," in *2020 29th International Conference on Computer Communications and Networks (ICCCN)*. IEEE, 2020, pp. 1–9.
- [99] J. Zhang, J. Zhang, J. Chen, and S. Yu, "Gan enhanced membership inference: A passive local attack in federated learning," in *ICC 2020-2020 IEEE International Conference on Communications (ICC)*. IEEE, 2020, pp. 1–6.
- [100] L. Zhang, B. Shen, A. Barnawi, S. Xi, N. Kumar, and Y. Wu, "FedDP-GAN: federated differentially private generative adversarial networks framework for the detection of COVID-19 pneumonia," *Information Systems Frontiers*, vol. 23, no. 6, pp. 1403–1415, 2021.
- [101] Z. Yan, J. Wicaksana, Z. Wang, X. Yang, and K.-T. Cheng, "Variation-aware federated learning with multi-source decentralized medical image data," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 7, pp. 2615–2628, 2020.
- [102] W. Li, J. Chen, Z. Wang, Z. Shen, C. Ma, and X. Cui, "If-gan: Improved federated learning generative adversarial network with maximum mean discrepancy model aggregation," *IEEE Transactions on Neural Networks and Learning Systems*, 2022.
- [103] M. B. Sariyildiz and R. G. Cinbis, "Gradient matching generative networks for zero-shot learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 2168–2178.
- [104] Y. Xian, T. Lorenz, B. Schiele, and Z. Akata, "Feature generating networks for zero-shot learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5542–5551.
- [105] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *International conference on machine learning*. PMLR, 2017, pp. 214–223.
- [106] 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*. PMLR, 2022, pp. 21 111–21 132.
- [107] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4401–4410.
- [108] M. Shin, C. Hwang, J. Kim, J. Park, M. Bennis, and S.-L. Kim, "Xor mixup: Privacy-preserving data augmentation for one-shot federated learning," *arXiv preprint arXiv:2006.05148*, 2020.
- [109] T. Yoon, S. Shin, S. J. Hwang, and E. Yang, "Fedmix: Approximation of mixup under mean augmented federated learning," *arXiv preprint arXiv:2107.00233*, 2021.
- [110] C.-H. Yao, B. Gong, H. Qi, Y. Cui, Y. Zhu, and M.-H. Yang, "Federated multi-target domain adaptation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 1424–1433.
- [111] S. Yang, S. Choi, H. Park, S. Choi, S. Chang, and S. Yun, "Client-agnostic learning and zero-shot adaptation for federated domain generalization," 2023. [Online]. Available: <https://openreview.net/forum?id=S4PGxC1bznF>
- [112] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.
- [113] P. Micaelli and A. J. Storkey, "Zero-shot knowledge transfer via adversarial belief matching," *Advances in Neural Information Processing Systems*, vol. 32, 2019.
- [114] L. Zhang, D. Wu, and X. Yuan, "Fedzkt: Zero-shot knowledge transfer towards resource-constrained federated learning with heterogeneous on-device models," in *2022 IEEE 42nd International Conference on Distributed Computing Systems (ICDCS)*. IEEE, 2022, pp. 928–938.
- [115] O. Dalmaz, U. Mirza, G. Elmas, M. Özbey, S. U. Dar, and T. Çukur, "A specificity-preserving generative model for federated mri translation," in *Distributed, Collaborative, and Federated Learning, and Affordable AI and Healthcare for Resource Diverse Global Health: Third MICCAI Workshop, DeCaF 2022, and Second MICCAI Workshop, FAIR 2022, Held in Conjunction with MICCAI 2022, Singapore, September 18 and 22, 2022, Proceedings*. Springer, 2022, pp. 79–88.
- [116] A. T. Nguyen, P. Torr, and S. N. Lim, "Fedsr: A simple and effective domain generalization method for federated learning," *Advances in Neural Information Processing Systems*, vol. 35, pp. 38 831–38 843, 2022.

- [117] L. Zhang, Y. Luo, Y. Bai, B. Du, and L.-Y. Duan, "Federated learning for Non-IID data via unified feature learning and optimization objective alignment," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 4420–4428.
- [118] Y. Kang, Y. Liu, and T. Chen, "Fedmvt: Semi-supervised vertical federated learning with multiview training," *arXiv preprint arXiv:2008.10838*, vol. 3, 2020.
- [119] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz, "Invariant risk minimization," *arXiv preprint arXiv:1907.02893*, 2019.
- [120] S. Francis, I. Tenison, and I. Rish, "Towards causal federated learning for enhanced robustness and privacy," *arXiv preprint arXiv:2104.06557*, 2021.
- [121] P. Venkateswaran, V. Isahagian, V. Muthusamy, and N. Venkatasubramanian, "FedGen: Generalizable federated learning," *arXiv preprint arXiv:2211.01914*, 2022.
- [122] M. Zhang, H. Marklund, N. Dhawan, A. Gupta, S. Levine, and C. Finn, "Adaptive risk minimization: Learning to adapt to domain shift," *Advances in Neural Information Processing Systems*, vol. 34, pp. 23 664–23 678, 2021.
- [123] 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.
- [124] T. V. Vo, A. Bhattacharyya, Y. Lee, and T.-Y. Leong, "An adaptive kernel approach to federated learning of heterogeneous causal effects," *Advances in Neural Information Processing Systems*, vol. 35, pp. 24 459–24 473, 2022.
- [125] S. Salgia, S. Vakili, and Q. Zhao, "Collaborative learning in kernel-based bandits for distributed users," 2023.
- [126] S. Hong and J. Chae, "Communication-efficient randomized algorithm for multi-kernel online federated learning," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 12, pp. 9872–9886, 2021.
- [127] W. Zhuang, X. Gan, X. Zhang, Y. Wen, S. Zhang, and S. Yi, "Federated unsupervised domain adaptation for face recognition," in *2022 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2022, pp. 1–6.
- [128] X. Zhou, Y. Tian, and X. Wang, "Source-target unified knowledge distillation for memory-efficient federated domain adaptation on edge devices," 2022. [Online]. Available: <https://openreview.net/forum?id=8rCMq0yJMG>
- [129] Z. Niu, H. Wang, H. Sun, S. Ouyang, Y.-w. Chen, and L. Lin, "MCKD: Mutually collaborative knowledge distillation for federated domain adaptation and generalization," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.
- [130] W. Huang, M. Ye, B. Du, and X. Gao, "Few-shot model agnostic federated learning," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 7309–7316.
- [131] L.-L. Zeng, Z. Fan, J. Su, M. Gan, L. Peng, H. Shen, and D. Hu, "Gradient matching federated domain adaptation for brain image classification," *IEEE Transactions on Neural Networks and Learning Systems*, 2022.
- [132] Y. Shi, J. Seely, P. H. Torr, N. Siddharth, A. Hannun, N. Usunier, and G. Synnaeve, "Gradient matching for domain generalization," *arXiv preprint arXiv:2104.09937*, 2021.
- [133] Z. Jiang, J. Gu, M. Liu, and D. Z. Pan, "Delving into effective gradient matching for dataset condensation," *arXiv preprint arXiv:2208.00311*, 2022.
- [134] H. Jeong, D. Needell, and J. Qin, "Federated gradient matching pursuit," *arXiv preprint arXiv:2302.10755*, 2023.
- [135] M. H. Zhu, L. N. Ezzine, D. Liu, and Y. Bengio, "FedILC: Weighted geometric mean and invariant gradient covariance for federated learning on non-iid data," *arXiv preprint arXiv:2205.09305*, 2022.
- [136] X. Li, H.-I. Chi, W. Lu, F. Xue, J. Zeng, and C. Z. Li, "Federated transfer learning enabled smart work packaging for preserving personal image information of construction worker," *Automation in Construction*, vol. 128, p. 103738, 2021.
- [137] Y. Zhou, J. Wang, and Z. Wang, "Bearing faulty prediction method based on federated transfer learning and knowledge distillation," *Machines*, vol. 10, no. 5, p. 376, 2022.
- [138] Y. Fan, Y. Li, M. Zhan, H. Cui, and Y. Zhang, "Totdefender: A federated transfer learning intrusion detection framework for 5G IoT," in *2020 IEEE 14th international conference on big data science and engineering (BigDataSE)*. IEEE, 2020, pp. 88–95.
- [139] Y. Chen, X. Qin, J. Wang, C. Yu, and W. Gao, "Fedhealth: A federated transfer learning framework for wearable healthcare," *IEEE Intelligent Systems*, vol. 35, no. 4, pp. 83–93, 2020.
- [140] C.-M. Feng, Y. Yan, S. Wang, Y. Xu, L. Shao, and H. Fu, "Specificity-preserving federated learning for mr image reconstruction," *IEEE Transactions on Medical Imaging*, 2022.
- [141] Y. Shen, Y. Zhou, and L. Yu, "Cd2-pfed: Cyclic distillation-guided channel decoupling for model personalization in federated learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 041–10 050.
- [142] J. Weng, K. Hu, T. Yao, J. Wang, and Z. Wang, "Robust knowledge adaptation for federated unsupervised person ReID," in *2022 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*. IEEE, 2022, pp. 1–8.
- [143] M. Nergiz, "Collaborative colorectal cancer classification on highly class imbalanced data setting via federated neural style transfer based data augmentation," *Traitement du Signal*, vol. 39, no. 6, 2022.
- [144] L. Fantauzzo, E. Fani, D. Caldarola, A. Tavera, F. Cermelli, M. Ciccone, and B. Caputo, "Feddrive: generalizing federated learning to semantic segmentation in autonomous driving," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 11 504–11 511.
- [145] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *International conference on machine learning*. PMLR, 2017, pp. 1126–1135.
- [146] F. Chen, M. Luo, Z. Dong, Z. Li, and X. He, "Federated meta-learning with fast convergence and efficient communication," *arXiv preprint arXiv:1802.07876*, 2018.
- [147] B. Wang, A. Li, M. Pang, H. Li, and Y. Chen, "Graphfl: A federated learning framework for semi-supervised node classification on graphs," in *2022 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2022, pp. 498–507.
- [148] Y. Jiang, J. Konečný, K. Rush, and S. Kannan, "Improving federated learning personalization via model agnostic meta learning," *arXiv preprint arXiv:1909.12488*, 2019.
- [149] A. Fallah, A. Mokhtari, and A. Ozdaglar, "Generalization of model-agnostic meta-learning algorithms: Recurring and unseen tasks," *Advances in Neural Information Processing Systems*, vol. 34, pp. 5469–5480, 2021.
- [150] J. Li, T. Chen, and X. Qian, "Generalizable pancreas segmentation modeling in ct imaging via meta-learning and latent-space feature flow generation," *IEEE Journal of Biomedical and Health Informatics*, 2022.
- [151] J. Chen, J. Tang, and W. Li, "Industrial edge intelligence: Federated-meta learning framework for few-shot fault diagnosis," *IEEE Transactions on Network Science and Engineering*, 2023.
- [152] M. Khodak, M.-F. F. Balcan, and A. S. Talwalkar, "Adaptive gradient-based meta-learning methods," *Advances in Neural Information Processing Systems*, vol. 32, 2019.
- [153] Q. Liu, Q. Dou, C. Chen, and P.-A. Heng, "Domain generalization of deep networks for medical image segmentation via meta learning," in *Meta-Learning with Medical Imaging and Health Informatics Applications*. Elsevier, 2023, pp. 117–139.
- [154] A. Frikha, H. Chen, D. Krompaß, T. Runkler, and V. Tresp, "Towards data-free domain generalization," in *Asian Conference on Machine Learning*. PMLR, 2023, pp. 327–342.
- [155] B. Li and C. Zhao, "Federated zero-shot industrial fault diagnosis with cloud-shared semantic knowledge base," *IEEE Internet of Things Journal*, 2023.
- [156] G. K. Nayak, K. R. Mopuri, V. Shaj, V. B. Radhakrishnan, and A. Chakraborty, "Zero-shot knowledge distillation in deep networks," in *International Conference on Machine Learning*. PMLR, 2019, pp. 4743–4751.
- [157] H. Chen, Y. Wang, C. Xu, Z. Yang, C. Liu, B. Shi, C. Xu, C. Xu, and Q. Tian, "Data-free learning of student networks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 3514–3522.
- [158] Z. Zhu, J. Hong, and J. Zhou, "Data-free knowledge distillation for heterogeneous federated learning," in *International Conference on Machine Learning*. PMLR, 2021, pp. 12 878–12 889.
- [159] E. Jeong, S. Oh, H. Kim, J. Park, M. Bennis, and S.-L. Kim, "Communication-efficient on-device machine learning: Federated distillation and augmentation under non-iid private data," *arXiv preprint arXiv:1811.11479*, 2018.
- [160] L. Zhang, L. Shen, L. Ding, D. Tao, and L.-Y. Duan, "Fine-tuning global model via data-free knowledge distillation for non-iid federated learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 174–10 183.
- [161] C. Chu and R. Wang, "A survey of domain adaptation for neural machine translation," *arXiv preprint arXiv:1806.00258*, 2018.

- [162] D. Saunders, "Domain adaptation and multi-domain adaptation for neural machine translation: A survey," *Journal of Artificial Intelligence Research*, vol. 75, pp. 351–424, 2022.
- [163] Y. Zhao, M. Li, L. Lai, N. Suda, D. Civin, and V. Chandra, "Federated learning with non-iid data," *arXiv preprint arXiv:1806.00582*, 2018.
- [164] Z. Zhou, S. S. Azam, C. Brinton, and D. I. Inouye, "Efficient federated domain translation," in *The Eleventh International Conference on Learning Representations*, 2023.
- [165] G. Xie, J. Wang, Y. Huang, J. Lyu, F. Zheng, Y. Zheng, and Y. Jin, "Fedmed-gan: Federated domain translation on unsupervised cross-modality brain image synthesis," *Available at SSRN 4342071*, 2022.
- [166] R. Zhang, Q. Xu, J. Yao, Y. Zhang, Q. Tian, and Y. Wang, "Federated domain generalization with generalization adjustment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 3954–3963.
- [167] Y. Li, W. Zhou, H. Wang, H. Mi, and T. M. Hospedales, "Fedh2l: Federated learning with model and statistical heterogeneity," *arXiv preprint arXiv:2101.11296*, 2021.
- [168] L. Zhang, X. Lei, Y. Shi, H. Huang, and C. Chen, "Federated learning with domain generalization," *arXiv preprint arXiv:2111.10487*, 2021.
- [169] X. Li, L. Yu, Y. Jin, C.-W. Fu, L. Xing, and P.-A. Heng, "Rare disease classification via difficulty-aware meta learning," in *Meta-Learning with Medical Imaging and Health Informatics Applications*. Elsevier, 2023, pp. 331–347.
- [170] W. Ding, M. Abdel-Basset, H. Hawash, M. Pratama, and W. Pedrycz, "Generalizable segmentation of covid-19 infection from multi-site tomography scans: A federated learning framework," *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2023.
- [171] M. Abdel-Basset, H. Hawash, and M. Abouhawwash, "Collaborative screening of covid-19-like disease from multi-institutional radiographs: A federated learning approach," *Mathematics*, vol. 10, no. 24, p. 4766, 2022.
- [172] G. Luo, T. Liu, J. Lu, X. Chen, L. Yu, J. Wu, D. Z. Chen, and W. Cai, "Influence of data distribution on federated learning performance in tumor segmentation," *Radiology: Artificial Intelligence*, p. e220082, 2023.
- [173] A. Imteaj and M. H. Amini, "Leveraging asynchronous federated learning to predict customers financial distress," *Intelligent Systems with Applications*, vol. 14, p. 200064, 2022.
- [174] Y. Li and G. Wen, "Research and practice of financial credit risk management based on federated learning," *Engineering Letters*, vol. 31, no. 1, 2023.
- [175] K. Cheng, T. Fan, Y. Jin, Y. Liu, T. Chen, D. Papadopoulos, and Q. Yang, "Secureboost: A lossless federated learning framework," *IEEE Intelligent Systems*, vol. 36, no. 6, pp. 87–98, 2021.
- [176] W. Zheng, L. Yan, C. Gou, and F.-Y. Wang, "Federated meta-learning for fraudulent credit card detection," in *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, 2021, pp. 4654–4660.
- [177] B. Xu, S. Yan, S. Li, and Y. Du, "A federated transfer learning framework based on heterogeneous domain adaptation for students' grades classification," *Applied Sciences*, vol. 12, no. 21, p. 10711, 2022.
- [178] E. V. Soboleva and N. L. Karavaev, "Characteristics of the project-based teamwork in the case of developing a smart application in a digital educational environment," *European Journal of Contemporary Education*, vol. 9, no. 2, pp. 417–433, 2020.
- [179] X. Wang, W. Liu, H. Lin, J. Hu, K. Kaur, and M. S. Hossain, "AI-empowered trajectory anomaly detection for intelligent transportation systems: A hierarchical federated learning approach," *IEEE Transactions on Intelligent Transportation Systems*, 2022.
- [180] M. Abdel-Basset, N. Moustafa, H. Hawash, I. Razzak, K. M. Sallam, and O. M. Elkomy, "Federated intrusion detection in blockchain-based smart transportation systems," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 3, pp. 2523–2537, 2021.
- [181] C. Zhang, Y. Zhu, C. Markos, S. Yu, and J. James, "Toward crowd-sourced transportation mode identification: A semisupervised federated learning approach," *IEEE Internet of Things Journal*, vol. 9, no. 14, pp. 11 868–11 882, 2021.
- [182] F. Farnia, A. Reisizadeh, R. Pedarsani, and A. Jadbabaie, "An optimal transport approach to personalized federated learning," *IEEE Journal on Selected Areas in Information Theory*, vol. 3, no. 2, pp. 162–171, 2022.
- [183] R. Bai, S. Bagchi, and D. I. Inouye, "Benchmarking algorithms for domain generalization in federated learning," 2023. [Online]. Available: <https://openreview.net/forum?id=IsCg7qoy8i9>
- [184] 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.
- [185] B. Y. Lin, C. He, T. Li, N. Mehrabi, F. Mireshghallah, C. Xie, M. Soltanolkotabi, and X. Ren, "Workshop on federated learning for natural language processing (flnlp 2022)," 2022.
- [186] Y. Xianjia, J. P. Queralta, J. Heikkonen, and T. Westerlund, "Federated learning in robotic and autonomous systems," *Procedia Computer Science*, vol. 191, pp. 135–142, 2021.
- [187] F. E. Casado and Y. Demiris, "Federated learning from demonstration for active assistance to smart wheelchair users," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 9326–9331.
- [188] K. Mülling, J. Kober, O. Kroemer, and J. Peters, "Learning to select and generalize striking movements in robot table tennis," *The International Journal of Robotics Research*, vol. 32, no. 3, pp. 263–279, 2013.
- [189] B. Liu, L. Wang, M. Liu, and C.-Z. Xu, "Federated imitation learning: A novel framework for cloud robotic systems with heterogeneous sensor data," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 3509–3516, 2020.
- [190] B. Liu, L. Wang, and M. Liu, "Lifelong federated reinforcement learning: a learning architecture for navigation in cloud robotic systems," *IEEE Robotics and Automation Letters*, vol. 4, no. 4, pp. 4555–4562, 2019.
- [191] A. K. Tanwani, N. Mor, J. Kubiawicz, J. E. Gonzalez, and K. Goldberg, "A fog robotics approach to deep robot learning: Application to object recognition and grasp planning in surface decluttering," in *2019 international conference on robotics and automation (ICRA)*. IEEE, 2019, pp. 4559–4566.
- [192] Z. He, H. Shao, Z. Ding, H. Jiang, and J. Cheng, "Modified deep autoencoder driven by multisource parameters for fault transfer prognosis of aeroengine," *IEEE Transactions on Industrial Electronics*, vol. 69, no. 1, pp. 845–855, 2021.
- [193] L. Cui, Y. Sun, X. Wang, and H. Wang, "Spectrum-based, full-band preprocessing, and two-dimensional separation of bearing and gear compound faults diagnosis," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–16, 2021.
- [194] A. Hazra, P. K. Donta, T. Amgoth, and S. Dustdar, "Cooperative transmission scheduling and computation offloading with collaboration of fog and cloud for industrial IoT applications," *IEEE Internet of Things Journal*, vol. 10, no. 5, pp. 3944–3953, 2023.
- [195] V. Casamayor Pujol, A. Morichetta, I. Murturi, P. Kumar Donta, and S. Dustdar, "Fundamental research challenges for distributed computing continuum systems," *Information*, vol. 14, no. 3, p. 198, 2023.
- [196] V. Casamayor Pujol, P. K. Donta, A. Morichetta, I. Murturi, and S. Dustdar, "Distributed computing continuum systems—opportunities and research challenges," in *Service-Oriented Computing—ICSOC 2022 Workshops: ASOCA, AI-PA, FMCIoT, WESOACS 2022, Sevilla, Spain, November 29–December 2, 2022 Proceedings*. Springer, 2023, pp. 405–407.
- [197] P. K. Donta and S. Dustdar, "The promising role of representation learning for distributed computing continuum systems," in *2022 IEEE International Conference on Service-Oriented System Engineering (SOSE)*. IEEE, 2022, pp. 126–132.
- [198] P. K. Donta, B. Sedlak, V. Casamayor Pujol, and S. Dustdar, "Governance and sustainability of distributed continuum systems: a big data approach," *Journal of Big Data*, vol. 10, no. 1, pp. 1–31, 2023.
- [199] S. Dustdar, V. C. Pujol, and P. K. Donta, "On distributed computing continuum systems," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 4, pp. 4092–4105, 2022.
- [200] S. Dustdar and I. Murturi, "Towards iot processes on the edge," *Next-Gen Digital Services. A Retrospective and Roadmap for Service Computing of the Future: Essays Dedicated to Michael Papazoglou on the Occasion of His 65th Birthday and His Retirement*, pp. 167–178, 2021.
- [201] I. Murturi, A. Egyed, and S. Dustdar, "Utilizing ai planning on the edge," *IEEE Internet Computing*, vol. 26, no. 2, pp. 28–35, 2022.
- [202] S. Dustdar and I. Murturi, "Towards distributed edge-based systems," in *2020 IEEE Second International Conference on Cognitive Machine Intelligence (CogMI)*. IEEE, 2020, pp. 1–9.
- [203] I. Murturi and S. Dustdar, "A decentralized approach for resource discovery using metadata replication in edge networks," *IEEE Transactions on Services Computing*, vol. 15, no. 5, pp. 2526–2537, 2021.
- [204] I. Murturi, C. Avasalcai, C. Tsigkanos, and S. Dustdar, "Edge-to-edge resource discovery using metadata replication," in *2019 IEEE 3rd International Conference on Fog and Edge Computing (ICFEC)*. IEEE, 2019, pp. 1–6.

- [205] C. Tsigkanos, I. Murturi, and S. Dustdar, "Dependable resource coordination on the edge at runtime," *Proceedings of the IEEE*, vol. 107, no. 8, pp. 1520–1536, 2019.
- [206] I. Gulrajani and D. Lopez-Paz, "In search of lost domain generalization," in *ICLR*, 2021.
- [207] Z. Qin, L. Yang, F. Gao, Q. Hu, and C. Shen, "Uncertainty-aware aggregation for federated open set domain adaptation," *IEEE Transactions on Neural Networks and Learning Systems*, 2022.
- [208] K. Saito, S. Yamamoto, Y. Ushiku, and T. Harada, "Open set domain adaptation by backpropagation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 153–168.
- [209] M. Cuturi, "Sinkhorn distances: Lightspeed computation of optimal transport," *Advances in neural information processing systems*, vol. 26, 2013.
- [210] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in neural information processing systems*, vol. 30, 2017.
- [211] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *The journal of machine learning research*, vol. 17, no. 1, pp. 2096–2030, 2016.
- [212] F. M. Carlucci, A. D'Innocente, S. Bucci, B. Caputo, and T. Tommasi, "Domain generalization by solving jigsaw puzzles," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2229–2238.
- [213] I. Albuquerque, N. Naik, J. Li, N. Keskar, and R. Socher, "Improving out-of-distribution generalization via multi-task self-supervised pre-training," *arXiv preprint arXiv:2003.13525*, 2020.
- [214] S. Wang, L. Yu, C. Li, C.-W. Fu, and P.-A. Heng, "Learning from extrinsic and intrinsic supervisions for domain generalization," in *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX*. Springer, 2020, pp. 159–176.
- [215] Z. Huang, H. Wang, E. P. Xing, and D. Huang, "Self-challenging improves cross-domain generalization," in *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*. Springer, 2020, pp. 124–140.
- [216] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," *Proceedings of Machine learning and systems*, vol. 2, pp. 429–450, 2020.
- [217] S. P. Karimireddy, S. Kale, M. Mohri, S. Reddi, S. Stich, and A. T. Suresh, "Scaffold: Stochastic controlled averaging for federated learning," in *International Conference on Machine Learning*. PMLR, 2020, pp. 5132–5143.
- [218] K. Bonawitz, V. Ivanov, B. Kreuter, A. Marcedone, H. B. McMahan, S. Patel, D. Ramage, A. Segal, and K. Seth, "Practical secure aggregation for privacy-preserving machine learning," in *proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, 2017, pp. 1175–1191.
- [219] K. Wei, J. Li, M. Ding, C. Ma, H. H. Yang, F. Farokhi, S. Jin, T. Q. Quek, and H. V. Poor, "Federated learning with differential privacy: Algorithms and performance analysis," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 3454–3469, 2020.
- [220] T. Chang, L. Li, M. Wu, W. Yu, X. Wang, and C. Xu, "Pagroup: Privacy-aware grouping framework for high-performance federated learning," *Journal of Parallel and Distributed Computing*, 2023.
- [221] O. Choudhury, A. Gkoulalas-Divanis, T. Salonidis, I. Sylla, Y. Park, G. Hsu, and A. Das, "Anonymizing data for privacy-preserving federated learning," *arXiv preprint arXiv:2002.09096*, 2020.
- [222] Y. Cheng, Y. Liu, T. Chen, and Q. Yang, "Federated learning for privacy-preserving ai," *Communications of the ACM*, vol. 63, no. 12, pp. 33–36, 2020.
- [223] C. Wang, R. Grosse, S. Fidler, and G. Zhang, "Eigendamage: Structured pruning in the kronecker-factored eigenbasis," in *International conference on machine learning*. PMLR, 2019, pp. 6566–6575.
- [224] T. Li, J. Li, Z. Liu, and C. Zhang, "Few sample knowledge distillation for efficient network compression," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 14 639–14 647.
- [225] D. Ye, R. Yu, M. Pan, and Z. Han, "Federated learning in vehicular edge computing: A selective model aggregation approach," *IEEE Access*, vol. 8, pp. 23 920–23 935, 2020.
- [226] Y. Deng, M. M. Kamani, and M. Mahdavi, "Adaptive personalized federated learning," *arXiv preprint arXiv:2003.13461*, 2020.
- [227] X. Guo, Z. Liu, J. Li, J. Gao, B. Hou, C. Dong, and T. Baker, "V eri fi: Communication-efficient and fast verifiable aggregation for federated learning," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 1736–1751, 2020.
- [228] H. Zhu, H. Zhang, and Y. Jin, "From federated learning to federated neural architecture search: a survey," *Complex & Intelligent Systems*, vol. 7, pp. 639–657, 2021.
- [229] S. Wang, T. Tuor, T. Salonidis, K. K. Leung, C. Makaya, T. He, and K. Chan, "Adaptive federated learning in resource constrained edge computing systems," *IEEE journal on selected areas in communications*, vol. 37, no. 6, pp. 1205–1221, 2019.
- [230] R. Yu and P. Li, "Toward resource-efficient federated learning in mobile edge computing," *IEEE Network*, vol. 35, no. 1, pp. 148–155, 2021.
- [231] C. Bettini, G. Civitarese, and R. Presotto, "Personalized semi-supervised federated learning for human activity recognition," *arXiv preprint arXiv:2104.08094*, 2021.
- [232] W. Huang, M. Ye, and B. Du, "Learn from others and be yourself in heterogeneous federated learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 143–10 153.
- [233] F. Yu, W. Zhang, Z. Qin, Z. Xu, D. Wang, C. Liu, Z. Tian, and X. Chen, "Heterogeneous federated learning," *arXiv preprint arXiv:2008.06767*, 2020.
- [234] M. Andreux, J. O. du Terrail, C. Beguier, and E. W. Tramel, "Siloe federated learning for multi-centric histopathology datasets," in *Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning: Second MICCAI Workshop, DART 2020, and First MICCAI Workshop, DCL 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4–8, 2020, Proceedings 2*. Springer, 2020, pp. 129–139.
- [235] X. Ma, J. Zhu, Z. Lin, S. Chen, and Y. Qin, "A state-of-the-art survey on solving non-iid data in federated learning," *Future Generation Computer Systems*, vol. 135, pp. 244–258, 2022.
- [236] B. Luo, X. Li, S. Wang, J. Huang, and L. Tassiulas, "Cost-effective federated learning design," in *IEEE INFOCOM 2021-IEEE Conference on Computer Communications*. IEEE, 2021, pp. 1–10.
- [237] A. Z. Tan, H. Yu, L. Cui, and Q. Yang, "Towards personalized federated learning," *IEEE Transactions on Neural Networks and Learning Systems*, 2022.
- [238] S. Ji, T. Saravirta, S. Pan, G. Long, and A. Walid, "Emerging trends in federated learning: From model fusion to federated x learning," *arXiv preprint arXiv:2102.12920*, 2021.
- [239] E. Jiang, Y. J. Zhang, and O. Koyejo, "Federated domain adaptation via gradient projection," 2023.
- [240] D. Gao, Y. Liu, A. Huang, C. Ju, H. Yu, and Q. Yang, "Privacy-preserving heterogeneous federated transfer learning," in *2019 IEEE International Conference on Big Data (Big Data)*, 2019, pp. 2552–2559.
- [241] T. Zhang, L. Gao, C. He, M. Zhang, B. Krishnamachari, and A. S. Avestimehr, "Federated learning for the internet of things: applications, challenges, and opportunities," *IEEE Internet of Things Magazine*, vol. 5, no. 1, pp. 24–29, 2022.
- [242] E. Guberović, T. Lipić, and I. Čavrak, "Dew intelligence: Federated learning perspective," in *2021 IEEE 45th Annual Computers, Software, and Applications Conference (COMPSAC)*. IEEE, 2021, pp. 1819–1824.
- [243] A. Reisizadeh, A. Mokhtari, H. Hassani, A. Jadbabaie, and R. Pedarsani, "Fedpaq: A communication-efficient federated learning method with periodic averaging and quantization," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2020, pp. 2021–2031.
- [244] J. Kang, X. Li, J. Nie, Y. Liu, M. Xu, Z. Xiong, D. Niyato, and Q. Yan, "Communication-efficient and cross-chain empowered federated learning for artificial intelligence of things," *IEEE Transactions on Network Science and Engineering*, vol. 9, no. 5, pp. 2966–2977, 2022.
- [245] N. Lang, E. Sofer, N. Shlezinger, R. G. D'Oliveira, and S. El Rouayheb, "Cpa: Compressed private aggregation for scalable federated learning over massive networks," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.
- [246] S. Lee, T. Zhang, C. He, and S. Avestimehr, "Layer-wise adaptive model aggregation for scalable federated learning," *arXiv preprint arXiv:2110.10302*, 2021.
- [247] 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.
- [248] K. Li, L. Yu, and P.-A. Heng, "Domain-incremental cardiac image segmentation with style-oriented replay and domain-sensitive feature

whitening,” *IEEE Transactions on Medical Imaging*, vol. 42, no. 3, pp. 570–581, 2023.

- [249] C. You, J. Xiang, K. Su, X. Zhang, S. Dong, J. Onofrey, L. Staib, and J. S. Duncan, “Incremental learning meets transfer learning: Application to multi-site prostate mri segmentation,” in *Distributed, Collaborative, and Federated Learning, and Affordable AI and Healthcare for Resource Diverse Global Health: Third MICCAI Workshop, DeCaF 2022, and Second MICCAI Workshop, FAIR 2022, Held in Conjunction with MICCAI 2022, Singapore, September 18 and 22, 2022, Proceedings*. Springer, 2022, pp. 3–16.
- [250] B. Wang, G. Li, C. Wu, W. Zhang, J. Zhou, and Y. Wei, “A framework for self-supervised federated domain adaptation,” *EURASIP Journal on Wireless Communications and Networking*, vol. 2022, no. 1, p. 37, 2022.
- [251] K. Kopparapu and E. Lin, “Fedfmc: Sequential efficient federated learning on non-iid data,” *arXiv preprint arXiv:2006.10937*, 2020.
- [252] N. Shoham, T. Avidor, A. Keren, N. Israel, D. Benditkis, L. Mor-Yosef, and I. Zeitak, “Overcoming forgetting in federated learning on non-iid data,” *arXiv preprint arXiv:1910.07796*, 2019.
- [253] K. You, M. Long, Z. Cao, J. Wang, and M. I. Jordan, “Universal domain adaptation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
- [254] W. Zellinger, V. Wieser, M. Kumar, D. Brunner, N. Shepeleva, R. Gálvez, J. Langer, L. Fischer, and B. Moser, “Beyond federated learning: On confidentiality-critical machine learning applications in industry,” *Procedia Computer Science*, vol. 180, pp. 734–743, 2021.



Ying Li received the B.S. degree in the Internet of Things from Anyang Institute Of Technology, Anyang, China, in 2017, and the M.S. degree in computer technology from Northeastern University, Shenyang, China, in 2020, where she is currently pursuing the Ph.D. degree in computer science and technology. She is a visiting PhD at Distributed Systems Group, TU Wien, Austria from 2022 to 2024.

Her research interests include distributed machine learning, blockchain, and knowledge-defined net-

working.



Xingwei Wang received the B.S., M.S., and Ph.D. degrees in computer science from Northeastern University, Shenyang, China, in 1989, 1992, and 1998, respectively. He is currently a Professor with the College of Computer Science and Engineering, Northeastern University. He has published more than 100 journal articles, books and book chapters, and refereed conference papers.

His research interests include cloud computing and future Internet. Prof. Wang has received several best paper awards.



Rongfei Zeng is the Associate Professor of Software College at Northeastern University. He received his Ph.D. degree in Computer Science and Technology from Tsinghua University with honor in 2012. He has published several papers in top journals and conferences such as *IEEE Transactions on Parallel and Distributed Systems (TPDS)*, *Elsevier Computer Networks*, *IEEE ICDCS*.

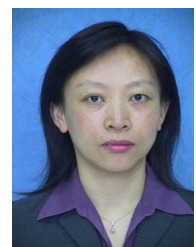
His research interests include network security and privacy, machine learning and its security, and industrial networks and IoT.



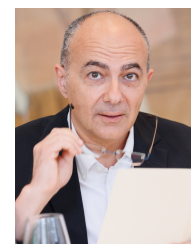
Praveen Kumar Donta (SM'22) is a Postdoctoral researcher in the Distributed Systems Group, TU Wien, Austria. He received his Ph.D. from the Department of Computer Science and Engineering in Indian Institute of Technology (Indian School of Mines), Dhanbad, India. He was a visiting Ph.D. student at the University of Tartu, Estonia. He received his Masters and Bachelor of Technology from JNTU Anantapur, India in 2014, and 2012. He is serving as Editorial board member for several journals including *Measurement*, *Measurement: Sensors*, *Computer Communications*, *Elsevier*, and *PLOS ONE*. His current research on Learning-driven distributed computing continuum systems, Edge Intelligence, and Intelligent data protocols.



Ilir Murturi (Member, IEEE) is a Postdoctoral Researcher in the Distributed Systems Group, TU Wien, Austria. He received a Ph.D. in the Distributed Systems Group, Technische Universität Wien (TU Wien), Vienna, Austria, and an MSc in Computer Engineering from the Faculty of Electrical and Computer Engineering, University of Prishtina, Prishtina, Kosova. He is a member of IEEE and a reviewer for several journals and conferences. His current research interests include the Internet of Things, Distributed Computing Continuum Systems, EdgeAI, and privacy in distributed, self-adaptive and cyber-physical systems.



Min Huang (Member, IEEE) received the B.S. degree in automatic instrument, the M.S. degree in systems engineering, and the Ph.D. degree in control theory from Northeastern University, Shenyang, China, in 1990, 1993, and 1999, respectively. She is currently a Professor with the College of Information Science and Engineering, Northeastern University. She has published more than 100 journal articles, books, and refereed conference papers. Her research interests include modeling and optimization for logistics and supply chain system.



Schahram Dustdar (Fellow, IEEE) is a full professor of computer science (informatics) with a focus on Internet Technologies heading the Distributed Systems Group at the TU Wien. He is member of the Academia Europaea. He is the recipient of the ACM Distinguished Scientist Award and Distinguished Speaker award, and the IBM Faculty Award. He is an associate editor of *IEEE Transactions on Services Computing*, *ACM Transactions on the Web*, and *ACM Transactions on Internet Technology*, and on the editorial board of *IEEE Internet Computing* and *IEEE Computer*. He is the editor-in-chief of *Computing* (an SCI-ranked journal of Springer).