Exploring One-Shot Semi-supervised Federated Learning with Pre-trained Diffusion Models

Mingzhao Yang *, **Shangchao Su** *, **Bin Li** †, **Xiangyang Xue** †

Shanghai Key Laboratory of Intelligent Information Processing
School of Computer Science, Fudan University
{mzyang20,scsu20,libin,xyxue}@fudan.edu.cn

Abstract

Recently, semi-supervised federated learning (semi-FL) has been proposed to handle the commonly seen real-world scenarios with labeled data on the server and unlabeled data on the clients. However, existing methods face several challenges such as communication costs, data heterogeneity, and training pressure on client devices. To address these challenges, we introduce the powerful diffusion models (DM) into semi-FL and propose **FedDISC**, a **Fed**erated **D**iffusion-**I**nspired **S**emi-supervised **C**o-training method. Specifically, we first extract prototypes of the labeled server data and use these prototypes to predict pseudo-labels of the client data. For each category, we compute the cluster centroids and domain-specific representations to signify the semantic and stylistic information of their distributions. After adding noise, these representations are sent back to the server, which uses the pre-trained DM to generate synthetic datasets complying with the client distributions and train a global model on it. With the assistance of vast knowledge within DM, the synthetic datasets have comparable quality and diversity to the client images, subsequently enabling the training of global models that achieve performance equivalent to or even surpassing the ceiling of supervised centralized training. FedDISC works within one communication round, does not require any local training, and involves very minimal information uploading, greatly enhancing its practicality. Extensive experiments on three large-scale datasets demonstrate that FedDISC effectively addresses the semi-FL problem on non-IID clients and outperforms the compared SOTA methods. Sufficient visualization experiments also illustrate that the synthetic dataset generated by FedDISC exhibits comparable diversity and quality to the original client dataset, with a neglectable possibility of leaking privacy-sensitive information of the clients.

Introduction

Federated Learning (FL) (McMahan et al. 2017) is a new paradigm of machine learning that allows multiple clients to perform collaborative training without sharing private data. Realistic FL scenarios, such as mobile album classification and autonomous driving (Nguyen et al. 2022; Fantauzzo et al. 2022), often involve individual users who are unwilling or unable to provide reliable annotations. This often results in the client data being unlabeled in practice. Semi-supervised FL (Semi-FL) (Zhang et al. 2021b; Diao, Ding, and Tarokh 2021; Jeong et al. 2021) has been proposed to address this issue. In this setting, there are multiple clients with unlabeled data and a server with labeled data. The goal of semi-FL is to obtain a global model that adapts to all client distributions.

Due to its allowance for unlabeled client data, semi-FL should be the most practically valuable topic within FL. However, existing semi-FL methods are unable to be practically deployed in real-world scenarios due to the following reasons (Li et al. 2020a; Kairouz et al. 2021; Mammen 2021): Firstly, the primary challenge lies in communication. Currently, all semi-FL methods heavily rely on multi-round communication, which significantly increases the burden on the clients. Secondly, the challenge of data heterogeneity persists in semi-FL. When there are distribution differences between the server and the clients, the performance of the global model significantly decreases. The third point pertains to the diverse devices of clients in real-world scenarios. While many FL methods do not restrict the computing power of the clients, one of the greatest challenges in the practical implementation of FL methods is that most clients cannot support model training on their devices, such as a significant portion of mobile terminals in scenarios like mobile album classification and autonomous driving. Hence, to address the aforementioned challenges, it is essential to establish a **one-shot semi-FL method without any client training**.

Recently, the development of diffusion models (DM) (Radford et al. 2021; Rombach et al. 2022) offers fresh opportunities. These pre-trained DMs exhibit remarkable performance. With proper guidance, these DMs can generate data with sufficient variety in both categories and distributions. If there is a method to generate guidance about the personalized distributions of the clients, it becomes possible to limitlessly generate high-quality, large-scale realistic images that comply with various client distributions in semi-FL. With the synthetic datasets, one can achieve one-shot semi-FL without the need for any client training, even in scenarios with highly non-IID clients. This approach simultaneously addresses the aforementioned three challenges and significantly enhances the practicality of semi-FL.

Furthermore, an additional pivotal advantage of applying pre-trained DMs in semi-FL is the potential to surmount the

*These authors contributed equally.
†Corresponding author
Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

"performance ceiling" of traditional FL, which involves uploading all client images to the server for centralized training of the global model. This ceiling bypasses any performance losses caused by distributed training and privacy preservation, enabling the training of the optimal global model. In semi-FL, this ceiling becomes even more unattainable, as the data uploaded to the server can be labeled, introducing additional supervised information. But even this ceiling performance is constrained by the knowledge within the client samples. However, with the vast knowledge within the DMs, it becomes possible to generate samples with both higher diversity and quality than the original client data, with the great possibility of surpassing the ceiling performance of centralized training.

Motivated by these opportunities, in this paper, we introduce **FedDISC**, a **Fed**erated **D**iffusion-**I**nspired **S**emi-supervised **C**o-training method, to leverage powerful foundation models in one-shot semi-FL. In brief, FedDISC involves four key steps: Firstly, following the common approach in semi-FL, we obtain prototypes for each category at the server and then send these prototypes to the clients. Secondly, the clients extract the features of the unlabeled client images and employ the received prototypes to assign pseudo-labels to these images. Thirdly, we obtain some cluster centroids and a domain-specific representation by clustering and averaging the client features for each category. These selected features possess the capability to capture the semantic information and the style characteristics of the personalized client distribution, which is then transmitted to the server. Finally, guided by the received representations, the server utilizes the pre-trained DM to limitlessly generate various samples complying with the specified distributions, resulting in a high-quality synthetic dataset. With the powerful DM, the generated samples closely resemble both the distribution and quality of the client dataset, enabling training a global model that achieves performance comparable to the ceiling performance or even surpassing it in some cases.

Our experiments on DomainNet (Peng et al. 2019), OpenImage (Kuznetsova et al. 2020), and NICO++ (Zhang et al. 2022b) demonstrate that FedDISC can obtain a high-performance global model that adapts to various client distributions within one round communication. In some cases, it even outperforms the ceiling of centralized training. A large number of visualization experiments also demonstrate that we can generate synthetic datasets that exhibit quality and diversity comparable to the original client datasets, without the leaking of privacy-sensitive information. In certain cases, these synthetic datasets can even possess a more comprehensive knowledge than the original client datasets.

Our contributions are summarized as follows:

- We demonstrate the excellent performance of DMs when applied to FL, enabling us to obtain high-quality large-scale synthetic datasets complying with the various client distributions without any training on the clients, which has not been explored before.
- We propose the FedDISC method. With the help of cluster centroids and domain-specific representations, our method further improves both the quality and variety of the generated samples, resulting in a global model that has the potential to outperform the performance ceiling of centralized training with only one communication round.
- We conduct extensive experiments on multiple real-world large-scale image datasets to validate the effectiveness of FedDISC. The results demonstrate that FedDISC outperforms compared with all the baseline methods. In some cases, it even surpasses the performance ceiling of traditional FL. The sufficient visualization experiments also illustrate that our method can generate synthetic datasets with competitive quality and diversity compared to the original client images, without leaking the privacy-sensitive information of the clients.

Related Works

Federated Learning

Supervised Federated Learning. FedAvg (McMahan et al. 2017) proposes the FL problem setting. However, some studies (Li et al. 2019, 2020b) notice the problem in non-IID scenarios. To address this challenge, numerous works have attempted to establish stronger global models (Karimireddy et al. 2020; Li et al. 2020b; Wang et al. 2020; Reddi et al. 2020), or personalized FL that allows clients to obtain personalized parameters (Fallah, Mokhtari, and Ozdaglar 2020; Huang et al. 2021; Caruana 1997; Dinh, Tran, and Nguyen 2020; Zhang et al. 2021a). Some works (Li et al. 2007) involve aggregating distributed information from clients by uploading it to the server. In addition, to further reduce communication costs, some works (Zhang et al. 2022a; Heinbaugh, Luz-Ricca, and Shao 2022; Su, Li, and Xue 2023) propose one-shot FL, which performs one round of communication.

Semi-supervised Federated Learning. In realistic FL scenarios, clients own a large amount of unlabeled data. In response to this issue, FedMatch (Jeong et al. 2021) proposes semi-FL, which is mentioned in the introduction. (Zhang et al. 2021b) points out the importance of the gradient diversity problem and proposes several strategies. (Diao, Ding, and Tarokh 2021) supposes the additional auxiliary dataset to handle semi-FL. In the paper, an additional challenge about communication and client training is identified and a solution is proposed, enhancing the practicality of semi-FL.

Foundation Models

Recently, foundation models (Radford et al. 2021; Kirillov et al. 2023; Rombach et al. 2022; Yu et al. 2023) have achieved unprecedented success in computer vision. CLIP (Radford et al. 2021) has bridged the gap between text and vision. DMs (Sohl-Dickstein et al. 2015; Ho, Jain, and Abbeel 2020) have provided a new generative paradigm, with Stable Diffusion (Rombach et al. 2022) achieving remarkable performance. A major advantage of DMs is the ability to use various conditions to guide generation, such as a trained classifier (Dhariwal and Nichol 2021), text (Nichol et al. 2021; Saharia et al. 2022b; Kim, Kwon, and Ye 2022; Jin et al. 2023) and images (Saharia et al. 2022a; Zhang and Agrawala 2023; Su et al. 2022b; Wang et al. 2022; Preechakul et al. 2022). There are also some works (Liu et al. 2022; Huang et al. 2023; Du et al. 2023) studying compositional generation of DMs. From their performance in image generation,

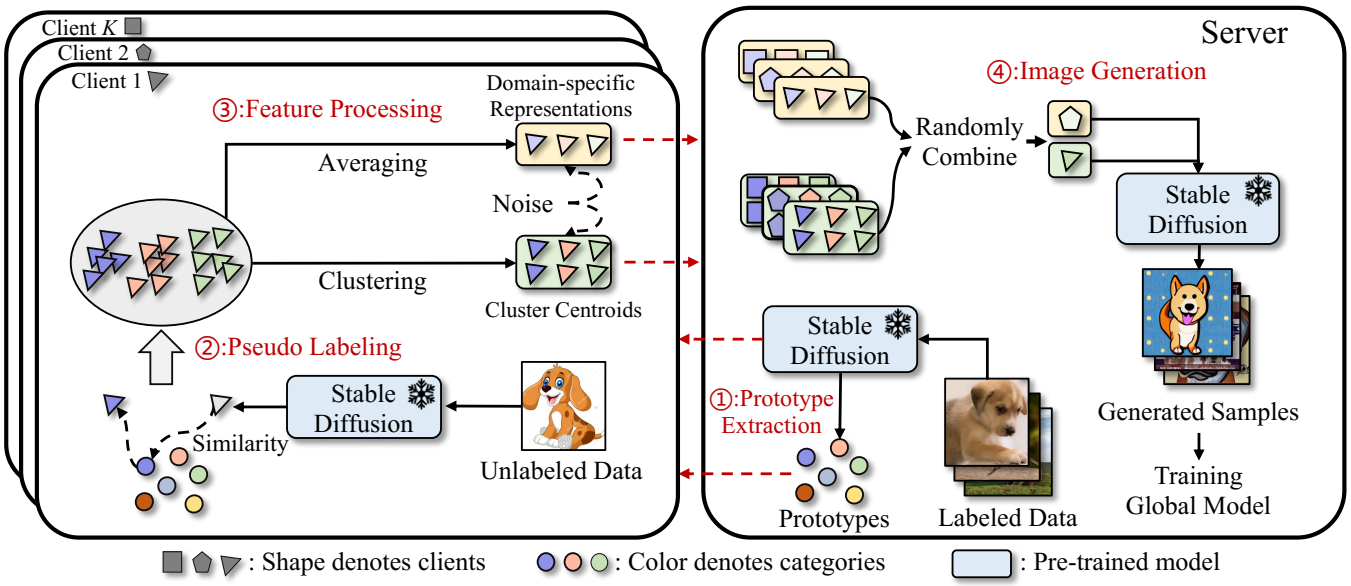

Figure 1: The framework of FedDISC. The overall method consists of four steps: Prototype Extraction, Pseudo Labeling, Feature Processing, and Image Generation.

a large-scale pre-trained DM can generate realistic images within an acceptable cost of time and computation on the server. This is the main reason why we apply pre-trained DMs in FL.

FL with Foundation Models

Leveraging the powerful performance of fundamental models, some works (Su et al. 2022a; Guo et al. 2022; Yang et al. 2023) in FL have explored the application of foundation models in federated image classification. However, to the best of our knowledge, no work has yet utilized pre-trained DMs, such as Stable Diffusion, in semi-FL. In this paper, we make a novel attempt and reveal the potential of DMs in semi-FL. Therefore, the method proposed in this paper, which does not require client training, has great potential for practical applications.

Method

In this section, we introduce the proposed FedDISC method in detail. Firstly, we provide some notations and background knowledge on DMs. Then we describe the proposed method in detail through four steps taken by the clients and the server.

Preliminaries

Diffusion Models. The DMs study the transformation from the Gaussian distribution to the realistic distribution by iterative denoising. In this paper, since only the pre-trained DMs are used and no training is conducted, the sampling process of the DMs is mainly introduced here. During sampling, the DM ϵ_θ samples $\mathbf{s}_T$ from the Gaussian distribution, where T is the predetermined maximum timestep. The DM takes $\mathbf{s}_T$ as the initial noise of the denoising process and uses the input text prompt p and the input image q as conditions. After T timesteps of denoising, $\mathbf{s}_T$ is restored to a real image $\mathbf{s}_0$ with specified semantics. For any given time step $t \in \{0, \ldots, T\}$, the sampling process is as follows:

$$\mathbf{s}_{t-1} = \sqrt{\alpha_{t-1}}\left(\frac{\mathbf{s}_t - \sqrt{1-\alpha_t}\epsilon_\theta(\mathbf{s}_t, t|p, q)}{\sqrt{\alpha_t}}\right) + \sqrt{1-\alpha_{t-1}-\sigma_t^2}\cdot\epsilon_\theta(\mathbf{s}_t, t|p, q) + \sigma_t\varepsilon_t \quad (1)$$

where α_t, α_{t-1} and σ_t are pre-defined parameters, ε_t is the Gaussian noise randomly sampled at each timestep. It should be noted that currently, many methods can freely control the number of iterations of the denoising process to accelerate sampling, but the overall process is quite similar, so these methods won't be elaborated here.

Notations and Objectives. We consider a semi-FL setting, where we have K clients with unlabeled datasets $\mathcal{D}_k = \{\mathbf{x}_i^k\}_{i=1}^{N_k}, k = 1, \ldots, K$, where N_k is the number of images on the k-th client, and a server with a labeled dataset $\mathcal{D}_s = \{\mathbf{x}_i^s, y_i\}_{i=1}^{N_s}$, $y_i \in \{1, \ldots, M\}$, where N_s is the number of images on the server and M is the number of categories. The text prompts of these categories are $\mathcal{C}_j, j \in \{1, \ldots, M\}$. The objective of the whole FL framework is:

$$\min_{\mathbf{w}\in\mathbb{R}^d} \frac{1}{K}\sum_{k=1}^{K} \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_k}\left[\ell_k(\mathbf{w}; \mathbf{x})\right] \quad (2)$$

where ℓ_k is the local objective function for the k-th client, $\mathbf{w}$ is the parameters of the global model.

To reduce communication and computation costs and make it suitable for real-world scenarios, such as the devices in autonomous driving, we impose two constraints in this setting: 1) Clients cannot conduct model training and can only conduct model inference. 2) The federated training process can only involve one round of communication.

FedDISC

Our method has four detailed steps: prototype extraction, pseudo labeling, feature processing, and image generation.

Prototype Extraction. Firstly, we utilize a pre-trained CLIP image encoder E_θ to extract the features of all labeled data on the server. We assume that the server contains all possible categories that may appear on the clients, but each category on the server has a relatively single style or belongs to a limited fine-grained subclass.

After obtaining the features of the labeled images on the server, we extract prototypes $\mathbf{p}_j, j \in \{1, \ldots, M\}$ of all categories by calculating the average of all features with the same category:

$$\mathbf{p}_j = \frac{\sum_{(\mathbf{x}_i^s, y_i) \in \mathcal{D}_s} E_\theta(\mathbf{x}_i^s) * \mathbb{I}(y_i = j)}{\sum_{(\mathbf{x}_i^s, y_i) \in \mathcal{D}_s} \mathbb{I}(y_i = j)} \tag{3}$$

where $\mathbb{I}$ is the indicator function. Finally, we send the extracted category prototypes $\mathbf{p}_j, j \in \{1, \ldots, M\}$ and the pre-trained CLIP image encoder E_θ to all the clients.

Pseudo Labeling. For client k, after receiving the encoder E_θ and the prototypes $\mathbf{p}_j$ from the server, the client uses E_θ to extract features of all unlabeled images in $\mathcal{D}_k$ and calculates the similarities between each feature $E_\theta(\mathbf{x}_i^k)$ and all category prototypes $\mathbf{p}_j, j \in \{1, \ldots, M\}$.

$$sim(E_\theta(\mathbf{x}_i^k), \mathbf{p}_j) = \frac{E_\theta(\mathbf{x}_i^k)^\top \mathbf{p}_j}{\left\|E_\theta(\mathbf{x}_i^k)\right\| \left\|\mathbf{p}_j\right\|}, \mathbf{x}_i^k \in \mathcal{D}_k \tag{4}$$

Based on the similarities, each image $\mathbf{x}_i^k$ is assigned with a pseudo-label $\hat{y}_i^k$, where $\hat{y}_i^k = \arg\max_j sim(E_\theta(\mathbf{x}_i^k), \mathbf{p}_j)$. Due to the differences between $\mathcal{D}_s$ and $\mathcal{D}_k$, there is a possibility of making mistakes in pseudo labeling. In traditional semi-FL methods, pseudo-labels are used for self-training. Therefore, various semi-FL methods are required to improve the quality of pseudo-labels. However, in our method, on one hand, in feature processing, clustering can avoid uploading representations of incorrect categories. On the other hand, the introduction of text prompts during the generation process can also prevent the generation of images that do not correspond to the specified categories, which further ensures the correct semantic information of generated images.

Feature Processing. After obtaining the unlabeled client features and their pseudo labels $\{E_\theta(\mathbf{x}_i^k), \hat{y}_i^k\}_{i=1}^{N_k}$, taking category j as an example, we cluster the client features belonging to category j and select L cluster centroids $\{\mathbf{z}_{j,l}^k\}_{l=1}^L$ to upload. The objective of clustering is as follows:

$$\arg\min_{\mathbf{z}_{j,l}^k} \sum_{l=1}^{L} \sum_{\mathbf{x}_i^k \in \mathcal{D}_k} \left\|E_\theta(\mathbf{x}_i^k) - \mathbf{z}_{j,l}^k\right\|^2 * \mathbb{I}\left(\hat{y}_i^k = j\right) \tag{5}$$

Compared with randomly selecting L features for uploading, the personalized distribution information and semantic information contained in the cluster centroids are clearer. Since only a small number of features are selected and each feature will generate multiple images on the server, the quality of the uploaded features is crucial.

Meanwhile, we obtain the domain-specific representations $\{\mathbf{g}_j^k\}_{j=1}^M$ for each category on client k by averaging all the features belonging to category j. We weaken the individuality of each image and highlight the commonality of each category on the clients. During the conditional generation on the server, we can generate images complying with different client distributions by combining the cluster centroids with different domain-specific representations.

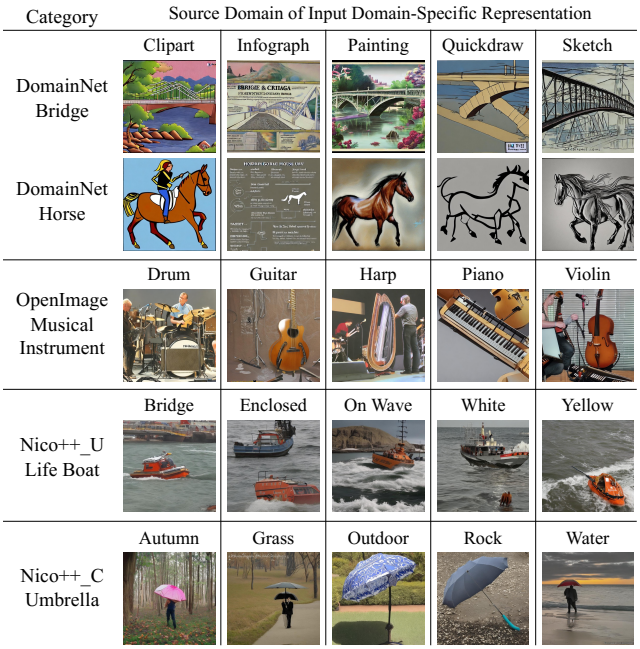

Figure 2: Generated images comply with different distributions on different datasets.

After computing the cluster centroids and the domain-specific representations, for privacy protection, we add noise to all these features. The noise-adding process is as follows:

$$\bar{\mathbf{z}}_{j,l}^k = \sqrt{\alpha_n}\mathbf{z}_{j,l}^k + \sqrt{1-\alpha_n}\varepsilon_1, \bar{\mathbf{g}}_j^k = \sqrt{\alpha_n}\mathbf{g}_j^k + \sqrt{1-\alpha_n}\varepsilon_2$$

where $\varepsilon_1, \varepsilon_2 \sim \mathcal{N}(0, \mathcal{I})$, n is a hyperparameter controlling the intensity of the noise, and $n \in \{0, \ldots, T\}$. We follow the noise-adding process in Stable Diffusion (Rombach et al. 2022) and perform a noise-adding process with a specific timestep to these image features. After this step, cluster centroids $\{\bar{\mathbf{z}}_{j,l}^k\}_{l=1}^L, j = \{1, \ldots, M\}$ and domain-specific representations $\{\bar{\mathbf{g}}_j^k\}_{j=1}^M$ are uploaded to the server.

Image Generation. After receiving $\bar{\mathbf{z}}_{j,l}^k$ and $\bar{\mathbf{g}}_j^k$ uploaded from the clients, for each cluster centroid $\bar{\mathbf{z}}_{j,l}^k$, the server randomly combines $\bar{\mathbf{z}}_{j,l}^k$ with the domain-specific representations which have the same pseudo-label j, the selected domain-specific representations $\mathcal{G}_{j,l}^k = \{\bar{\mathbf{g}}_j^{k_0}, \ldots, \bar{\mathbf{g}}_j^{k_R}\}$ is a random subset of domain-specific representations $\{\bar{\mathbf{g}}_j^k\}_{j=1}^M$.

As for the generating process, following (Liu et al. 2022), since we aim to use the cluster centroids $\bar{\mathbf{z}}_{j,l}^k$, domain-specific representations $\bar{\mathbf{g}}_j^{k_i}$, and the text prompts $\mathcal{C}_j$ as conditions for generation, the conditional probability distribution of the sample $\mathbf{s}$ in diffusion process can be written in the following

	OpenImage						DomainNet					
	client0	client1	client2	client3	client4	average	clipart	infograph	painting	quickdraw	sketch	average
Ceiling	*54.05*	*58.42*	*62.59*	*63.21*	*64.79*	*60.61*	*81.54*	*52.49*	*73.54*	*30.11*	*72.34*	*62.01*
Fine-tune	36.67	46.81	45.43	47.17	42.1	43.64	67.57	45.47	65.28	10.42	62.14	50.17
Zero-shot	56.03	40.61	40.28	44.06	**61.45**	48.47	65.86	40.5	62.25	13.36	57.92	47.98
Prompt	48.61	54.03	59.07	58.42	53.49	54.72	66.42	37.45	59.62	10.73	63.92	47.63
FedAvg	41.11	44.06	46.57	47.45	37.63	43.36	49.95	30.67	51.07	1.74	38.46	34.38
SemiFL	48.15	52.78	61.05	55.23	46.16	52.67	69.55	**47.16**	64.54	7.02	63.32	50.32
RSCFed	28.97	38.04	40.82	33.98	36.35	35.63	71.5	45.73	61.96	11.53	65.03	51.15
FedDISC	**56.11**	**62.49**	**62.53**	**59.16**	56.77	**59.42**	**72.54**	43.47	**67.42**	**17.71**	**67.25**	**53.68**
	NICO++_C						NICO++_U					
	client0	client1	client2	client3	client4	average	client0	client1	client2	client3	client4	average
Ceiling	*89.19*	*91.9*	*89.51*	*90.47*	*85.1*	*89.23*	*96.35*	*96.42*	*96.88*	*97.01*	97.26	*96.78*
Fine-tune	86.5	89.39	83.61	87.21	76.95	84.73	84.75	79.08	81.48	86.58	83.52	83.08
Zero-shot	78.66	85.26	80.01	80.7	72.14	79.35	89.2	89.24	87.19	85.5	88.6	87.94
Prompt	86.94	87.41	**89.73**	82.69	73.51	84.05	90.61	87.14	**89.96**	87.48	88.16	88.67
FedAvg	86.98	90.82	82.68	87.57	74.48	84.51	83.26	73.3	77.93	80.8	79.28	78.91
SemiFL	87.55	89.27	81.93	87.16	77.01	84.58	78.21	74.7	79.87	80.69	77.02	78.09
RSCFed	52.08	60.15	52.6	55.35	43.89	52.81	71.88	64.14	70.82	69.71	69.67	69.24
FedDISC	**87.97**	**92.09**	86.44	**90.52**	**84.17**	**88.24**	**91.73**	**90.82**	89.63	**92.83**	**90.15**	**91.03**

Table 1: The performances of different methods on OpenImage, DomainNet, and NICO++, where the italicized texts represent the inaccessible supervised ceiling performance used solely as a reference, and bold texts represent the best performance excluding the supervised ceiling performance.

form:

$$p(\mathbf{s}|\bar{\mathbf{z}}^k_{j,l}, \bar{\mathbf{g}}^{k_i}_j, \mathcal{C}_j) \propto p(\mathbf{s}|\mathcal{C}_j) p(\bar{\mathbf{z}}^k_{j,l}|\mathbf{s}, \mathcal{C}_j) p(\bar{\mathbf{g}}^{k_i}_j|\mathbf{s}, \mathcal{C}_j)$$

Since $\mathbf{s}$ is initially sampled from a Gaussian distribution, independent of the used cluster centroids and domain-specific representations, the above formula can be rewritten as:

$$p(\mathbf{s}|\bar{\mathbf{z}}^k_{j,l}, \bar{\mathbf{g}}^{k_i}_j, \mathcal{C}_j) \propto p(\mathbf{s}|\mathcal{C}_j) \frac{p(\mathbf{s}|\bar{\mathbf{z}}^k_{j,l}, \mathcal{C}_j)}{p(\mathbf{s}|\mathcal{C}_j)} \frac{p(\mathbf{s}|\bar{\mathbf{g}}^{k_i}_j, \mathcal{C}_j)}{p(\mathbf{s}|\mathcal{C}_j)}$$

Therefore, specifically, we use the feature of category prompt $\mathcal{C}_j$ with a cluster centroid $\bar{\mathbf{z}}^k_{c,l}$, a domain-specific representation $\bar{\mathbf{g}}^{k_i}_c$, and without any image feature to respectively obtain three predicted noises. We accumulate these three predicted noises in the following formula to obtain the final predicted noise:

$$\hat{\epsilon}_\theta(\mathbf{s}_t, t|\bar{\mathbf{z}}^k_{j,l}, \bar{\mathbf{g}}^{k_i}_c, \mathcal{C}_j) = \epsilon_\theta(\mathbf{s}_t, t|\mathcal{C}_j) + w_f(\epsilon_\theta(\mathbf{s}_t, t|\bar{\mathbf{z}}^k_{j,l}, \mathcal{C}_j) - \epsilon_\theta(\mathbf{s}_t, t|\mathcal{C}_j)) + w_g(\epsilon_\theta(\mathbf{s}_t, t|\bar{\mathbf{g}}^{k_i}_j, \mathcal{C}_j) - \epsilon_\theta(\mathbf{s}_t, t|\mathcal{C}_j))$$

where w_f and w_g are the weights of the predicted noises. Overall, the generated images are obtained through the denoising process:

$$\mathbf{s}_{t-1} = \sqrt{\alpha_{t-1}}\Big(\frac{\mathbf{s}_t - \sqrt{1-\alpha_t}\hat{\epsilon}_\theta(\mathbf{s}_t, t|\bar{\mathbf{z}}^k_{j,l}, \bar{\mathbf{g}}^{k_i}_j, \mathcal{C}_j)}{\sqrt{\alpha_t}}\Big) + \sqrt{1-\alpha_{t-1}-\sigma_t^2} \cdot \hat{\epsilon}_\theta(\mathbf{s}_t, t|\bar{\mathbf{z}}^k_{j,l}, \bar{\mathbf{g}}^{k_i}_j, \mathcal{C}_j) + \sigma_t \varepsilon_t \quad (6)$$

After obtaining the generated images, since both the cluster centroids and domain-specific representations used for image generation have their corresponding pseudo-labels, the generated images are pseudo-labeled. So we can directly fine-tune a classification model h with the generated dataset for downstream classification tasks. The classification model $h = F_\theta \circ E_\theta$ is a composite of the pre-trained CLIP image encoder E_θ and a linear classifier F_θ.

Experiments

Experimental Setup

Datasets. We adopt three datasets to evaluate the performance of FedDISC: DomainNet (Peng et al. 2019), OpenImage (Kuznetsova et al. 2020), and NICO++ (Zhang et al. 2022b). NICO++ can be divided into the common contexts (NICO++_C) and the unique contexts (NICO++_U). We divide each dataset into six clients based on the inherent domain division of the dataset itself. Due to space constraints, we provide a comprehensive description of our dataset in the supplementary materials.

Compared Methods. We mainly compare our method with 7 methods: 1) **Fine-tune**: Directly fine-tuning the model with the uploaded cluster centroids and corresponding pseudo-labels. 2) **Zero-shot**: Using the zero-shot classification capability of pre-trained CLIP to classify client images without any additional training. 3) **FedAvg**: Using clients' data and corresponding pseudo-labels for conducting FedAvg. In addition, we evaluate the SOTA semi-FL methods: 4) **SemiFL** (Diao, Ding, and Tarokh 2021) and 5) **RSCFed** (Liang et al. 2022). Note that as there is currently no one-shot semi-FL method, the chosen semi-supervised methods require multiple rounds of communication for comparison. 6) **Prompts**: Image generation is directly performed without utilizing any image features as guidance and solely relying on the text prompts. 7) **Ceiling**. As mentioned in the introduction, the inaccessible performance ceiling involves directly uploading and labeling all client images to the server for training the aggregated model, also known as centralized training. And it needs to be mentioned that the **FedAvg**, **RSCFed**, and **SemiFL** all need multiple iterations.

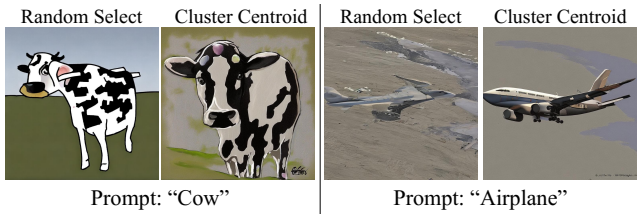

Figure 3: Comparison between generating using clustering centroids and the randomly selected client representations. With the provision of clustering centroids, the introduction of more representative semantic information leads to a significant improvement in the stability of the generated outputs.

Figure 4: The inclusion of domain-specific representations and their impact on the generated results. We can effectively alter the style of the generated images by controlling the added domain-specific representations, thereby enhancing the diversity of generated samples.

Main Results

Table 1 shows the performance of our method and various compared methods on four datasets. We highlight several observations:

- In addition to the Ceiling, FedDISC achieves the best average performance with only 1 communication round. This demonstrates the potential of DMs in FL.
- From the results on DomainNet, we can see that FedDISC has good performance on all clients except *infograph*. A possible reason is that the Stable Diffusion still has limited support for text in the images currently.
- On OpenImage, NICO++_C, and NICO++_U, compared with other baselines, FedDISC exhibits a significant performance improvement. That's because Stable Diffusion is exposed to more realistic images during pre-training.
- Despite the powerful ability of CLIP, directly using CLIP to perform classification still cannot achieve the best performance, therefore further fine-tuning is needed.
- Compared with Prompts Only, without the guidance of client image features, generated images would be heavily biased towards the most common distributions, demonstrating the necessity of guidance.
- Compared with Ceiling, FedDISC does not exhibit significant performance lag. It can even surpass in some domains, affirming the ideas posited in the introduction.

	client0	client1	client2	client3	client4
$L = 3$ Fine-tune	36.01	46.01	44.59	45.55	41.87
$L = 3$ FedDISC	**56.33**	**61.93**	**58.62**	**56.71**	**58.74**
$L = 5$ Fine-tune	36.67	46.81	45.43	47.17	42.10
$L = 5$ FedDISC	**56.11**	**62.49**	**62.53**	**59.16**	**56.77**
$L = 10$ Fine-tune	37.55	45.86	44.85	46.01	42.15
$L = 10$ FedDISC	**57.16**	**63.84**	**61.12**	**57.91**	**59.13**

Table 2: The influence of the number of cluster centroids.

	client0	client1	client2	client3	client4
$R = 3$	53.41	62.15	61.31	56.87	55.49
$R = 5$	54.58	**63.47**	61.26	58.19	**57.57**
$R = 10$	**56.11**	62.49	**62.53**	**59.16**	56.77

Table 3: The influence of the number of generated images.

From Figure 2 and the other visualization results in the supplementary materials, it can be seen that on DomainNet, OpenImage, NICO++_C, and NICO++_U our method can generate high-quality images that comply with various client distributions while being semantically correct, collectively underscoring the superior performance of FedDISC.

Ablation Experiments

The Number of Uploaded Cluster Centroids. We perform experiments on OpenImage to discuss the influence of the number of the uploaded cluster centroids L. Since this number is related to the performance of fine-tuning, we also test the performance of fine-tuning under different L for comparison. From Table 2, we can see that in most cases, uploading a small number of cluster centroids is already sufficient to represent the semantics of the subcategories. Increasing the number of cluster centroids enhances the availability of client information during the generation process, thereby further elevating the quality of generated images.

The Number of Generated Images. We discuss the number of images generated by each cluster centroid on OpenImage. From Table 3, we can find that the overall performance of the method gradually improves as R increases. This is reasonable as the increment of R indicates the increment in the variety of generated data. However, generating a small number of images is sufficient since the randomly sampled initial noise can bring some varieties as well.

The Roles of Domain-specific Representations and Cluster Centroids. We discuss the roles of Domain-specific Representations (DR) and Cluster Centroids (CC) on DomainNet. We compare FedDISC with cases where DR are not used during generation, and cases where CC are not uploaded, but an equal number of client features are randomly uploaded. As shown in Table 4, the removal of either DR or CC has a significant influence on the performance.

The visualization results in Figure 3 demonstrate that without the cluster centroids, the semantic information of the generated images becomes ambiguous and may lead to generating images that do not match the given text prompts.

DR	CC	client0	client1	client2	client3	client4
		66.42	37.45	59.62	10.73	63.92
✓		67.79	40.02	63.59	13.77	60.57
	✓	65.83	38.27	64.56	14.30	60.37
✓	✓	**72.54**	**43.47**	**67.42**	**17.71**	**67.25**

Table 4: The influence of different conditions.

	Upload Params (M)	Client Compute (Gflops)
FedAvg	30*632.08	30*1004.19
SemiFL	500*632.08	500*1004.19
RSCFed	100*632.08	100*1004.19
Ceiling	925.88	-
FedDISC	**4.23**	**334.73**

Table 5: Comparison about communication and computation.

Figure 4 shows that removing the domain-specific representations leads to the style of the generated images being monotonous, which reduces the diversity of the generated images. These results are consistent with our goals of using them mentioned in the introduction.

Discussions

The Privacy Issues. As mentioned in (Shao et al. 2023), the transmission of features is one of the existing ways of information sharing in FL. And considering the amount of data uploaded from clients, our method exhibits significantly lower privacy leakage compared to other FL methods. Moreover, to explore the feasibility of recovering private information from SD using the noise-added features, we focus on validating whether the generated images contain any privacy-sensitive information, such as text, faces, etc. For example, in some images with private text, we claim that the generated images with completely different texts do not leak the user's privacy, even though they appear to be similar. In Figure 5, we select four categories from OpenImage that may involve privacy-sensitive information and show some client images and their corresponding generated images. These results demonstrate that FedDISC has a low risk of leaking privacy-sensitive information during generation. Due to the space limitation, we provide further discussions in supplementary materials.

Communication and Computational Complexity. In Table 5, to compare communication and computational complexity with other compared methods, we conduct statistical analyses on DomainNet. Among the compared methods requiring iterations, the uploading and downloading of the image encoder is needed in each round. Ceiling needs to upload all client images. In contrast, FedDISC solely requires one time of downloading image encoder and uploading image features, incurring negligible communication.

Concerning computational complexity, since the server is generally not constrained by device performance due to tasks involving client scheduling and model aggregation, we only assess computation on the clients. All computation of Ceiling proceeds on the server and is devoid of reference

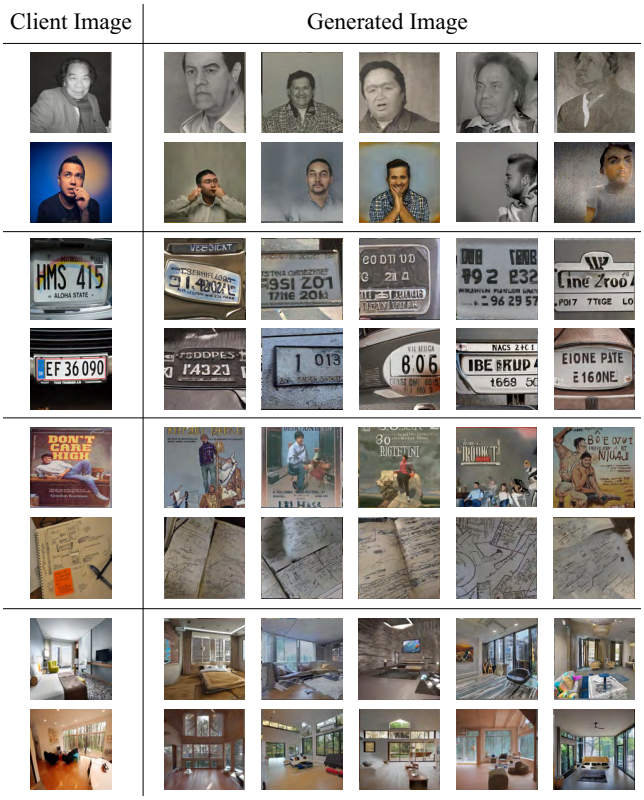

Figure 5: The comparison between the raw client images and their generated images. It can be observed that the generated images do not leak any sensitive privacy present in the original images, such as faces, text, etc. The generated images exhibit only a stylistic resemblance to the original images. Restoring original images starting from high-dimensional features without any training is nearly impossible.

value. Among the compared methods requiring iterations, multiple rounds of backpropagation exhibit more computing. But FedDISC needs a single forward propagation for feature extraction, resulting in approximately one-third of the computation compared to backpropagation. The aforementioned experiments demonstrate that FedDISC holds significant advantages in terms of communication and client computational complexity, underscoring its practicality.

Conclusion

In this paper, we explore the task of one-shot semi-FL and propose FedDISC, a new method that integrates pre-trained DMs into the semi-FL framework for the first time. In a single communication round and without any client training, our method achieves performance comparable to the ceiling performance and even surpasses it in some cases. The introduction of domain representations and clustering centroids further enhances the quality and stability of generation. Extensive quantitative and visualization experiments demonstrate the excellent performance of our method and underscore the potential and prospects of DMs within the FL.

Acknowledgements

This work was supported in part by the National Key R&D Program of China (No.2021ZD0112803), the National Natural Science Foundation of China (No.62176061), STCSM project (No.22511105000), the Shanghai Research and Innovation Functional Program (No.17DZ2260900), and the Program for Professor of Special Appointment (Eastern Scholar) at Shanghai Institutions of Higher Learning.

References

Caruana, R. 1997. Multitask learning. *Machine learning*, 28(1): 41–75.

Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34: 8780–8794.

Diao, E.; Ding, J.; and Tarokh, V. 2021. SemiFL: Communication efficient semi-supervised federated learning with unlabeled clients. *arXiv preprint arXiv:2106.01432*, 3.

Dinh, C. T.; Tran, N. H.; and Nguyen, T. D. 2020. Personalized Federated Learning with Moreau Envelopes. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Du, Y.; Durkan, C.; Strudel, R.; Tenenbaum, J. B.; Dieleman, S.; Fergus, R.; Sohl-Dickstein, J.; Doucet, A.; and Grathwohl, W. 2023. Reduce, Reuse, Recycle: Compositional Generation with Energy-Based Diffusion Models and MCMC. *arXiv preprint arXiv:2302.11552*.

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

Fantauzzo, L.; Fanì, E.; Caldarola, D.; Tavera, A.; Cermelli, F.; Ciccone, M.; and Caputo, B. 2022. Feddrive: Generalizing federated learning to semantic segmentation in autonomous driving. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 11504–11511. IEEE.

Guo, T.; Guo, S.; Wang, J.; and Xu, W. 2022. PromptFL: Let Federated Participants Cooperatively Learn Prompts Instead of Models–Federated Learning in Age of Foundation Model. *arXiv preprint arXiv:2208.11625*.

Heinbaugh, C. E.; Luz-Ricca, E.; and Shao, H. 2022. Data-Free One-Shot Federated Learning Under Very High Statistical Heterogeneity. In *The Eleventh International Conference on Learning Representations*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33: 6840–6851.

Huang, L.; Chen, D.; Liu, Y.; Shen, Y.; Zhao, D.; and Zhou, J. 2023. Composer: Creative and controllable image synthesis with composable conditions. *arXiv preprint arXiv:2302.09778*.

Huang, Y.; Chu, L.; Zhou, Z.; Wang, L.; Liu, J.; Pei, J.; and Zhang, Y. 2021. Personalized cross-silo federated learning on non-iid data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 7865–7873.

Jeong, W.; Yoon, J.; Yang, E.; and Hwang, S. J. 2021. Federated Semi-Supervised Learning with Inter-Client Consistency & Disjoint Learning. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Jin, Z.; Shen, X.; Li, B.; and Xue, X. 2023. Training-free Diffusion Model Adaptation for Variable-Sized Text-to-Image Synthesis. *arXiv preprint arXiv:2306.08645*.

Kairouz, P.; McMahan, H. B.; Avent, B.; Bellet, A.; Bennis, M.; Bhagoji, A. N.; Bonawitz, K.; Charles, Z.; Cormode, G.; Cummings, R.; et al. 2021. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2): 1–210.

Karimireddy, S. P.; Kale, S.; Mohri, M.; Reddi, S.; Stich, S.; and Suresh, A. T. 2020. SCAFFOLD: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning*, 5132–5143. PMLR.

Kim, G.; Kwon, T.; and Ye, J. C. 2022. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2426–2435.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. *arXiv preprint arXiv:2304.02643*.

Kuznetsova, A.; Rom, H.; Alldrin, N.; Uijlings, J.; Krasin, I.; Pont-Tuset, J.; Kamali, S.; Popov, S.; Malloci, M.; Kolesnikov, A.; et al. 2020. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, 128(7): 1956–1981.

Li, B.; Chi, M.; Fan, J.; and Xue, X. 2007. Support cluster machine. In *Proceedings of the 24th International Conference on Machine Learning*, 505–512.

Li, T.; Sahu, A. K.; Talwalkar, A.; and Smith, V. 2020a. Federated learning: Challenges, methods, and future directions. *IEEE signal processing magazine*, 37(3): 50–60.

Li, T.; Sahu, A. K.; Zaheer, M.; Sanjabi, M.; Talwalkar, A.; and Smith, V. 2020b. Federated Optimization in Heterogeneous Networks. In Dhillon, I. S.; Papailiopoulos, D. S.; and Sze, V., eds., *MLSys*. mlsys.org.

Li, X.; Huang, K.; Yang, W.; Wang, S.; and Zhang, Z. 2019. On the convergence of fedavg on non-iid data. *arXiv preprint arXiv:1907.02189*.

Liang, X.; Lin, Y.; Fu, H.; Zhu, L.; and Li, X. 2022. RSCFed: random sampling consensus federated semi-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10154–10163.

Liu, N.; Li, S.; Du, Y.; Torralba, A.; and Tenenbaum, J. B. 2022. Compositional visual generation with composable diffusion models. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVII*, 423–439. Springer.

Mammen, P. M. 2021. Federated learning: Opportunities and challenges. *arXiv preprint arXiv:2101.05428*.

McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; and y Arcas, B. A. 2017. Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*, 1273–1282. PMLR.

Nguyen, A.; Do, T.; Tran, M.; Nguyen, B. X.; Duong, C.; Phan, T.; Tjiputra, E.; and Tran, Q. D. 2022. Deep federated learning for autonomous driving. In *2022 IEEE Intelligent Vehicles Symposium (IV)*, 1824–1830. IEEE.

Nichol, A.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; McGrew, B.; Sutskever, I.; and Chen, M. 2021. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*.

Peng, X.; Bai, Q.; Xia, X.; Huang, Z.; Saenko, K.; and Wang, B. 2019. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1406–1415.

Preechakul, K.; Chatthee, N.; Wizadwongsa, S.; and Suwajanakorn, S. 2022. Diffusion autoencoders: Toward a meaningful and decodable representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10619–10629.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Reddi, S.; Charles, Z.; Zaheer, M.; Garrett, Z.; Rush, K.; Konečnỳ, J.; Kumar, S.; and McMahan, H. B. 2020. Adaptive federated optimization. *arXiv preprint arXiv:2003.00295*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10684–10695.

Saharia, C.; Chan, W.; Chang, H.; Lee, C.; Ho, J.; Salimans, T.; Fleet, D.; and Norouzi, M. 2022a. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*, 1–10.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, K.; Gontijo Lopes, R.; Karagol Ayan, B.; Salimans, T.; et al. 2022b. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35: 36479–36494.

Shao, J.; Li, Z.; Sun, W.; Zhou, T.; Sun, Y.; Liu, L.; Lin, Z.; and Zhang, J. 2023. A Survey of What to Share in Federated Learning: Perspectives on Model Utility, Privacy Leakage, and Communication Efficiency. *arXiv preprint arXiv:2307.10655*.

Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; and Ganguli, S. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, 2256–2265. PMLR.

Su, S.; Li, B.; and Xue, X. 2023. One-shot Federated Learning without server-side training. *Neural Networks*, 164: 203–215.

Su, S.; Yang, M.; Li, B.; and Xue, X. 2022a. Cross-domain Federated Adaptive Prompt Tuning for CLIP. *arXiv preprint arXiv:2211.07864*.

Su, X.; Song, J.; Meng, C.; and Ermon, S. 2022b. Dual diffusion implicit bridges for image-to-image translation. In *The Eleventh International Conference on Learning Representations*.

Wang, J.; Liu, Q.; Liang, H.; Joshi, G.; and Poor, H. V. 2020. Tackling the objective inconsistency problem in heterogeneous federated optimization. *Advances in Neural Information Processing Systems 33*.

Wang, T.; Zhang, T.; Zhang, B.; Ouyang, H.; Chen, D.; Chen, Q.; and Wen, F. 2022. Pretraining is all you need for image-to-image translation. *arXiv preprint arXiv:2205.12952*.

Yang, M.; Su, S.; Li, B.; and Xue, X. 2023. One-Shot Federated Learning with Classifier-Guided Diffusion Models. *arXiv preprint arXiv:2311.08870*.

Yu, H.; Wang, X.; Li, B.; and Xue, X. 2023. Chinese Text Recognition with A Pre-Trained CLIP-Like Model Through Image-IDS Aligning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11943–11952.

Zhang, J.; Chen, C.; Li, B.; Lyu, L.; Wu, S.; Ding, S.; Shen, C.; and Wu, C. 2022a. Dense: Data-free one-shot federated learning. *Advances in Neural Information Processing Systems*, 35: 21414–21428.

Zhang, L.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*.

Zhang, M.; Sapra, K.; Fidler, S.; Yeung, S.; and Alvarez, J. M. 2021a. Personalized Federated Learning with First Order Model Optimization. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Zhang, X.; Zhou, L.; Xu, R.; Cui, P.; Shen, Z.; and Liu, H. 2022b. Nico++: Towards better benchmarking for domain generalization. *arXiv preprint arXiv:2204.08040*.

Zhang, Z.; Yang, Y.; Yao, Z.; Yan, Y.; Gonzalez, J. E.; Ramchandran, K.; and Mahoney, M. W. 2021b. Improving semi-supervised federated learning by reducing the gradient diversity of models. In *2021 IEEE International Conference on Big Data (Big Data)*, 1214–1225. IEEE.