DOES LEARNING FROM DECENTRALIZED NON-IID UNLABELED DATA BENEFIT FROM SELF SUPERVISION?

Lirui Wang, Kaiqing Zhang, Yunzhu Li, Yonglong Tian, Russ Tedrake
MIT CSAIL

ABSTRACT

The success of machine learning relies heavily on massive amounts of data, which are usually generated and stored across a range of diverse and distributed data sources. *Decentralized learning* has thus been advocated and widely deployed to make efficient use of distributed datasets, with an extensive focus on supervised learning (SL) problems. Unfortunately, the majority of real-world data are *unlabeled* and can be highly *heterogeneous* across sources. In this work, we carefully study decentralized learning with unlabeled data through the lens of self-supervised learning (SSL), specifically contrastive visual representation learning. We study the effectiveness of a range of contrastive learning algorithms under a decentralized learning setting, on relatively large-scale datasets including ImageNet-100, MS-COCO, and a new real-world robotic warehouse dataset. Our experiments show that the *decentralized* SSL (Dec-SSL) approach is *robust* to the heterogeneity of decentralized datasets, and learns useful representation for object classification, detection, and segmentation tasks, even when combined with the simple and standard decentralized learning algorithm of Federated Averaging (`FedAvg`). This robustness makes it possible to significantly reduce communication and to reduce the participation ratio of data sources with only minimal drops in performance. Interestingly, using the same amount of data, the representation learned by Dec-SSL can not only perform on par with that learned by centralized SSL which requires communication and excessive data storage costs, but also sometimes outperform representations extracted from decentralized SL which requires extra knowledge about the data labels. Finally, we provide theoretical insights into understanding why data heterogeneity is less of a concern for Dec-SSL objectives, and introduce feature alignment and clustering techniques to develop a new Dec-SSL algorithm that further improves the performance, in the face of highly non-IID data. Our study presents positive evidence to embrace *unlabeled data* in decentralized learning, and we hope to provide new insights into whether and why decentralized SSL is effective and/or even advantageous.[1]

1 INTRODUCTION

The success of machine learning hinges heavily on the access to large-scale and diverse datasets. In practice, most data are generated from different locations, devices, and embodied agents, and stored in a distributed fashion. Examples include a fleet of self-driving cars collecting a massive amount of streaming images under various road and weather conditions during everyday driving, or individuals using mobile devices to take photos of objects and scenery all over the world. Besides being large-scale, these datasets have two salient features: they are *heterogeneous* across data sources, and mostly *unlabeled*. For instance, images of road conditions, which are expensive to label, vary across cars driving on highways vs. rural areas, and under sunny vs. snowy weather conditions (Figure 19).

Methods that can make the best use of these large-scale distributed datasets can significantly advance the performance of current machine learning algorithms and systems. This has thus motivated a surge of research in *decentralized learning/learning from decentralized data*[2] (Konečnỳ et al., 2016; Hsieh et al., 2017; McMahan et al., 2017; Kairouz et al., 2021; Nedic, 2020), where usually a global model is trained on the distributed datasets using communication between the local data sources and

[1] Code is available at `https://github.com/liruiw/Dec-SSL`

[2] Hereafter, we often use *decentralized learning* as a shorthand for *learning from decentralized data*.

a centralized server, or sometimes even only among the local data sources. The goal is typically to reduce or eliminate the exchanges of local raw data to save communication costs and protect data privacy. How to mitigate the effect of *data heterogeneity* remains one of the most important research questions in this area (Zhao et al., 2018; Hsieh et al., 2020; Karimireddy et al., 2020; Ghosh et al., 2020; Li et al., 2021a), as it can heavily downgrade the performance of decentralized learning. Moreover, most existing decentralized learning studies focused on *supervised learning* (SL) problems that require data labels (McMahan et al., 2017; Jeong et al., 2020; Hsieh et al., 2020). Hence, it remains unclear whether and how decentralized learning can benefit from large-scale, heterogeneous, and especially unlabeled datasets typically encountered in the real world.

On the other hand, people have developed effective methods of learning purely from unlabeled data and demonstrated impressive results. Self-supervised learning (SSL), a technique that learns *representations* by generating supervision signals from the data itself, has unleashed the power of unlabeled data and achieved tremendous successes for a wide range of downstream tasks in computer vision (He et al., 2020; Chen et al., 2020; He et al., 2021b), natural language processing (Devlin et al., 2018; Sarzynska-Wawer et al., 2021), and embodied intelligence (Sermanet et al., 2018; Florence et al., 2018). These SSL algorithms, however, are usually trained in a *centralized* fashion by pooling all the unlabeled data together, without accounting for the heterogeneous nature of the decentralized data sources. Very recently, there have been a few contemporaneous/concurrent attempts (He et al., 2021a; Zhuang et al., 2021; 2022; Lu et al., 2022; Makhija et al., 2022) that bridged unsupervised/self-supervised learning and decentralized learning, with focuses on *designing better algorithms* that mitigate the data heterogeneity issue. In contrast, we revisit this new paradigm and ask the question:

Does learning from decentralized non-IID unlabeled data really benefit from SSL?

We focus on *understanding* the use of SSL in decentralized learning when handling unlabeled data. We aim to answer whether and when decentralized SSL (Dec-SSL) is effective (even combined with simple and off-the-shelf decentralized learning algorithms, e.g., `FedAvg` (McMahan et al., 2017)); what are the unique inherent properties of Dec-SSL compared to its SL counterpart; how do the properties play a role in decentralized learning, especially with highly heterogeneous data? We also aim to validate our observations on large-scale and practical datasets. We defer a more detailed comparison with these most related works to §A.

In this paper, we show that unlike in decentralized (supervised) learning, data heterogeneity can be *less concerning* in decentralized SSL, with both empirical and theoretical evidence. This leads to more communication-efficient and robust decentralized learning schemes, which can sometimes even outperform their supervised counterpart that assumes the availability of label information. Among the first studies to bridge decentralized learning and SSL, our study provides positive evidence to embrace unlabeled data in decentralized learning, and provides new insights into this setting. We detail our contributions as follows.

Contributions. (i) We show that decentralized SSL, specifically contrastive visual representation learning, is a viable learning paradigm to handle relatively large-scale unlabeled datasets, even when combined with the simple `FedAvg` algorithm. Moreover, we also provide both experimental evidence and theoretical insights that decentralized SSL can be inherently *robust* to the data heterogeneity across different data sources. This allows more local updates, and can significantly improve the *communication efficiency* in decentralized learning. (ii) We provide further empirical and theoretical evidences that even when *labels* are available and decentralized supervised learning (and associated representation learning) is allowed, Dec-SSL still stands out in face of highly non-IID data. (iii) To further improve the performance of Dec-SSL, we design a new Dec-SSL algorithm, `FeatARC`, by using an iterative feature alignment and clustering procedure. Finally, we validate our hypothesis and algorithm in practical and large-scale data and task domains, including a new real-world robotic warehouse dataset.

2 Preliminaries and Overview

Consider a decentralized learning setting with K different data sources, which might correspond to different devices, machines, embodied agents, or datasets/users that can generate and store data locally. The goal is to collaboratively solve a learning problem, by exploiting the decentralized data from all data sources. More specifically, consider each data source $k \in [K]$ has local dataset $D_k = \{x_{k,i}\}_{i=1}^{|D_k|}$, and $x_{k,i} \in \mathcal{X} \subseteq \mathbb{R}^d$ are identically and independently distributed (IID) samples

from probability distribution $\mathcal{D}_k$, i.e., $x_{k,i} \sim \mathcal{D}_k$. Note that the distributions $\mathcal{D}_k$ is in general different across data sources k, yielding an overall *heterogeneous* (i.e., non-IID) data distribution for the data from all the sources. Let $D = \bigcup_{k\in[K]} D_k$ denote the set of all data samples. Moreover, we are interested in situations where no label is provided alongside the data x. To effectively utilize the large-scale *unlabeled* data, we resort to self-supervised learning approaches.

Specifically, SSL approaches extract representations from these unlabeled data, by finding an embedding function $f_w : \mathcal{X} \to \mathbb{R}^m$, where w is the parameter of the embedding function. $z = f_w(x)$ is the representation vector that can be useful for downstream tasks, e.g., classification or segmentation. We summarize several popular SSL approaches here that will be used later in the paper.

Self-supervised representation learning. Now consider a given data source $k \in [K]$. There are two popular methods in the SSL community. In contrastive learning (Chen et al., 2020; He et al., 2020) specifically, a sample x is used to provide supervision signals along with two generated *positive* samples x^+ and x (overloaded for notational simplicity) and (possibly multiple) *negative* samples x^- sampled from the training batch. The goal of SSL is to find an embedding f_w that makes x and x^+ close, while keeping x and x^-s apart, if negative samples are used.

One commonly used loss for SSL is the InfoNCE loss (Oord et al., 2018), which has been used in popular SSL approaches as SimCLR (Chen et al., 2020) and MoCo (He et al., 2020):

$$\mathcal{L}_k(w) := \frac{1}{|D_k|}\sum_{i=1}^{|D_k|} -\log\left(\frac{\exp(-\mathbb{D}(f_w(x_{k,i}), f_w(x^+_{k,i}))/\tau)}{\exp(-\mathbb{D}(f_w(x_{k,i}), f_w(x^+_{k,i}))/\tau) + \sum_j \exp(-\mathbb{D}(f_w(x_{k,i}), f_w(x^-_{k,j}))/\tau)}\right) \tag{2.1}$$

where $\tau > 0$ is a temperature hyperparameter, j is the index for negative samples, $\mathbb{D}(\cdot,\cdot)$ is a distance function such as the cosine distance, i.e., $\mathbb{D}(z_1, z_2) = -\frac{z_1 \cdot z_2}{||z_1||||z_2||}$. Some other effective SSL approaches, such as BYOL (Grill et al., 2020) and SimSiam (Chen & He, 2021), remove the terms related to negative samples in (2.1). These methods also add an additional function g, the *feature predictor*, which only applies to x to create an asymmetry and to avoid the collapsed solutions. This usually leads to the following objective: $\mathcal{L}_k(w) := \frac{1}{|D_k|}\sum_{i=1}^{|D_k|} \mathbb{D}\big(g(f_w(x_{k,i})), f_w(x^+_{k,i})\big)$. In our experiments, we make use of both losses and the SSL approaches associated with them.

Decentralized SSL. To exploit the heterogeneous data distributed at different locations/devices, decentralized SSL optimizes the following global objective:

$$\min_w \sum_{k\in[K]} \frac{|D_k|}{|D|}\mathcal{L}_k(w), \tag{2.2}$$

which can be solved using many existing decentralized learning algorithms. For instance, `FedAvg` (McMahan et al., 2017) is one of the most representative, easy-to-implement, and communication-efficient decentralized learning algorithms which optimizes this objective without data-sharing among data sources. At each iteration t, the server first samples a set of data sources $\mathcal{M}_t$ with size $|\mathcal{M}_t| = \rho K$ and run δ local update steps on each of the local dataset. Then, each data source $k \in \mathcal{M}_t$ sends back the updated local model weight $w_k^{t,\delta}$ to the central server, and the server averages them to be the global model $w^{t+1} = \frac{1}{|\mathcal{M}_t|}\sum_{k\in\mathcal{M}_t} w_k^{t,\delta}$ for the next round $t+1$. The server then broadcasts the global model to each data source to reset $w_k^{t+1,0}$ as w^{t+1}. The number of local updates (δ) determines the communication efficiency (larger δ means less communication); in the experiments, we use E to denote the number of epochs of local updates (as a surrogate for δ). Both E and the participation rate ρ are important factors that determine the efficiency of decentralized learning. The learned representation $f_w(x)$ can then be used in downstream supervised learning tasks. There are many real-world applications of decentralized SSL, including self-driving cars, warehouse robots, and mobile devices. A further discussion can be found in Appendix §D.

2.1 Overview of Our Study

Terminology & setup. We separate our experiment pipeline into **representation learning** (pre-training phase) and **downstream evaluation** (evaluation phase). Our main focus is on the aforementioned **Dec-SSL** approach. We use `FedAvg` (McMahan et al., 2017) with SimCLR (Chen et al., 2020) as the default method. Moreover, we will also compare with settings where the *label information* is available, i.e., the classical decentralized (supervised) learning, which should be more favorable for learning. See Figure 1 for a summary of different settings. The first setting is **Dec-**

SL: we simply run `FedAvg` on the decentralized labeled data, for end-to-end classification.Dec-SL does not learn *representations* explicitly, and serves as a natural baseline when labels are available. The second setting is *representation learning* from Dec-SL, where we train supervised learning with `FedAvg`, and then use the feature extractor network as the backbone for downstream tasks. This way, we can also learn the representation from decentralized labeled data, and make the comparison with Dec-SSL more fair, since both are learning features for various downstream tasks. We term this setting as **Dec-SLRep**.

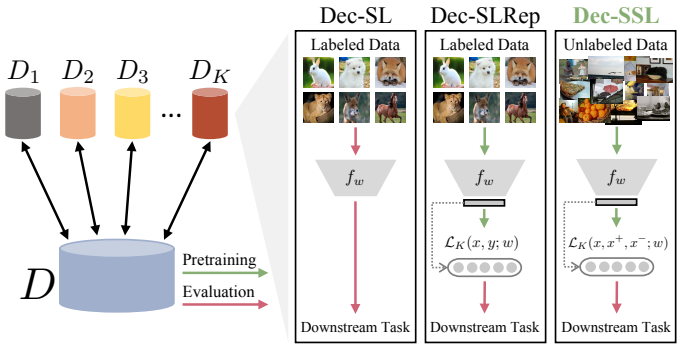

Figure 1: Comparisons among Dec-SL, Dec-SLRep, and Dec-SSL.

The evaluation phase tests the representations from Dec-SSL or Dec-SLRep. We consider two protocols in the evaluation phase: **linear probing** for image classification (Zhang et al., 2016) and **finetuning** for object detection/segmentation (Doersch et al., 2015). For classification, we train a linear classifier on top of the frozen pretrained network and evaluate the top-1 classification accuracy. For object detection/segmentation, we finetune the network by using the pretrained weights as initialization and training in an end-to-end fashion, and then we evaluate the mean Average Precision (mAP) metric. Downstream tasks are performed on centralized train and test dataset. Please refer to Appendix §C.1 for implementation details and Table 3 for experiment setups.

Questions of interest. Through extensive experiments on large-scale datasets, and theoretical analysis in simplified settings, we seek to answer the following questions: (i) How well can decentralized SSL, even instantiated with the simple `FedAvg` algorithm, rival the performance of its centralized counterpart, and handle the non-IIDness of decentralized unlabeled data? (ii) Is there any unique and inherent property of Dec-SSL, compared to its supervised learning counterpart; how and why may the property benefit decentralized learning, even when the label information is available? (iii) Is there a way to further improve the performance of Dec-SSL in face of highly non-IID data? Our hypothesis is that SSL, whose objective is not particularly dependent on the x to y mappings, learns a relatively *uniform* representation across decentralized and heterogeneous unlabeled datasets, thus leading to more efficient and robust decentralized learning schemes. We aim to validate this hypothesis and answer these questions in the following sections.

3 Dec-SSL is Efficient and Robust to Data Heterogeneity

We first seek to address question (i) in §2.1 – how well decentralized SSL performs, in face of non-IID and decentralized unlabeled data. To this end, we first introduce the notion of *data heterogeneity* in decentralized learning, which is usually categorized as *input heterogeneity, label distribution heterogeneity, and the heterogeneity in the relationships between the features and labels*, respectively (Hsieh et al., 2020). We create *label heterogeneity* by distributing each data source with different proportion of classes; we construct the heterogeneity via either sampling from a Dirichlet process with hyperparameter α or via skewness partitioning (Hsieh et al., 2020) with hyperparameter β. We also create *input heterogeneity* by leveraging the feature space of a pretrained network on the data. See §C.2 for more details on how we create data heterogeneity across data sources.

3.1 Experimental observations

CIFAR classification under different types of non-IIDness. In this experiment, we construct input and label non-IIDness using 5 data sources in the CIFAR-10 (Krizhevsky et al., 2009) dataset based on the Dirichlet Process. The sources of non-IIDness are the feature clusters and labels, respectively. We control parameter α to create datasets from very IID (each data source has roughly a uniform distribution over 10 classes / 5 feature clusters) to very non-IID (each data source has data from 2 classes / 1 feature clusters). Recall that E denotes the number of epochs for local updates and ρ denotes the participation ratio of data sources at each round. We use $E = 50$ epochs of local updates in this experiment, which is equivalent to around $\delta = 1000$ iterations, i.e., each local data source updates 50 epochs independently before averaging. The results are shown in Figure 2. Surprisingly,

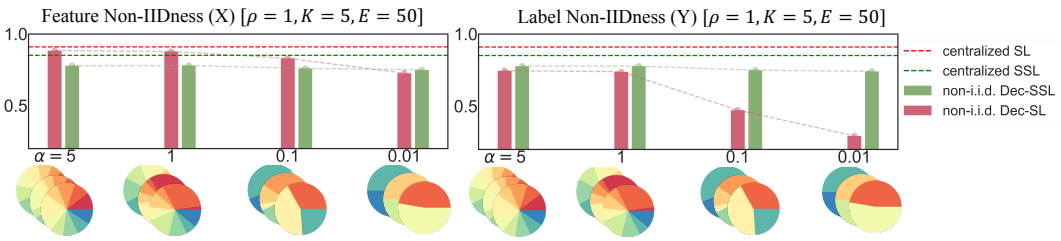

Figure 2: **SSL objective is robust to different types of X and Y heterogeneity on the CIFAR-10 dataset.** In the pie chart below, each pie denotes one data source, and color denotes the sample number of one source of non-IIDness (left to right, more non-IID). We observe that Dec-SSL is surprisingly robust to the non-IIDness in both input (X) and label (Y) and also behaves closer to its centralized counterpart. Y-axis denotes accuracy.

the performance of downstream classification, with representations trained using decentralized SSL, is very *insensitive* to the non-IIDness across the datasets and only bears a slight performance drop. This robustness over data non-IIDness is encouraging, and stands in sharp contrast with most existing decentralized supervised learning algorithms, which are known to suffer from the data heterogeneity in general (Hsieh et al., 2020). As a baseline, we consider the classical decentralized SL approach of `FedAvg`, trained over the same non-IID data, but with label information. Indeed, the performance of decentralized SL can drop significantly as the non-IIDness increases. Finally, we note that the simple use of `FedAvg` in SSL can achieve performance comparable to the centralized SSL, showing that Dec-SSL is an effective decentralized learning scheme to handle unlabeled data.

Finetuning ImageNet representation for COCO detection. In this experiment, we finetune the representations learned from ImageNet to COCO detection benchmark (Lin et al., 2014) with the Detectron pipeline (Girshick et al., 2018). Specifically, we use ImageNet-100 with ResNet-18 and $1\times$ training schedule for Mask R-CNN (He et al., 2017) with a ResNet18 FPN being the backbone. Compared to the contemporary works (Zhuang et al., 2022; Lu et al., 2022) on federated self-supervised learning, our setup is more relevant to real-world applications, as it works on larger-scale and more practical datasets and tasks.

We run Dec-SSL on ImageNet-100 dataset with 5 data sources, and with $E = 1$ epoch of local updates, which corresponds to around $\delta = 500$ local updates, to learn the global representation using `FedAvg`. On Table 1 left, we observe that the representation from Dec-SSL almost reaches the performance of the representation from centralized SSL and improves upon baselines that train the model from scratch, i.e., the *no pretrain* row. This conveys that SSL can learn useful representations in decentralized settings, avoiding the heavy communication cost of centralized learning.

Decentralized SSL for real-world package segmentation. The issue of data heterogeneity and communication efficiency is significant for real-world applications such as those in Amazon warehouses, whose fleets of working robots can generate millions of images per day (see Figure 21 for an illustration). We provide details about the Amazon dataset in §D.1. We use data from one sample warehouse site at Amazon, and split the data based on the session ID (which is usually a sequence of days). Each decentralized learner is only allowed to access the local data at one session, which is equivalent to the non-IID case where skewness $\beta = 0$. We then deploy decentralized self-supervised learning on a subset of the enormous warehouse data, which has around 80000 images with contour labels output by the Amazon work-cells. We use SimCLR with `FedAvg` and communication efficiency $E = 1$ number of local update epochs, as the pretraining method.

On the right subtable of Table 1, we compare different ways to initialize weights for finetuning, and show that the representations learned from decentralized SSL outperforms training from scratch and even matches centralized SSL on the Amazon dataset. We also experiment with finetuning segmentation task using Mask R-CNN on different fractions of the data, and show that Dec-SSL can further improve the performance of training from scratch, when there is no as much labeled data.

3.2 Theoretical Insights

We now provide some theoretical insights into why the objective of Dec-SSL leads to more robust performance in face of data heterogeneity. In particular, we analyze the property of the solutions to the local and global objectives of Dec-SSL in a simplified setting, and show that the global objective is not affected significantly by the heterogeneity of local datasets. Our setup is inspired by the very recent work (Liu et al., 2021), where the effect of imbalanced data in centralized SSL was studied in a simplified setting. In particular, we generalize the centralized and 3-way classification setting to a decentralized and $2K$-way one, carefully design the generation of data distribution across data sources, and establish analyses for both local and global objectives in decentralized SSL. We also

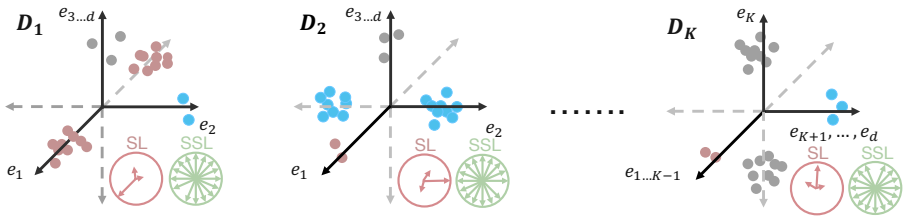

Figure 3: **The learned feature space of SSL is more insensitive to heterogeneity under the linear settings.** In §3.2, we consider a decentralized learning setting where each local dataset has a skewed distribution with most data points (each color is a class) concentrated on one axis. Each basis vector inside the sphere denotes how well it is represented in the learned subspace. For contrastive objectives, the learned feature space (green sphere) of the local model is more uniform and close to the global model. On the other hand, the SL objective (red sphere) tends to overfit to local dataset, and the learned feature spaces become heterogeneous.

ImageNet-100 Pretrain	MS-COCO AP^{bb}	MS-COCO AP^{mk}
no pretrain	20.5	19.4
Central-SLRep	21.2 (+0.7)	20.1 (+0.7)
Central-SSL	23.2 (+2.7)	22.1 (+2.7)
Dec-SLRep	19.8 (−0.7)	19.7 (+0.3)
Dec-SSL	22.1 (+1.6)	20.7 (+1.3)

Amazon Pretrain	Amazon (AP^{mk}) 100%	10%	1%
no pretrain	60.8	59.2	47.0
Central-SSL	61.6 (+0.8)	60.4 (+1.2)	49.5 (+2.5)
Dec-SSL	61.2 (+0.4)	60.1 (+0.9)	48.8 (+1.8)

Table 1: **Left: Object detection and semantic segmentation finetuned on COCO:** The model is pretrained on ImageNet-100 (Tian et al., 2020a) dataset and then finetune on MS-COCO with metrics bounding-box mAP (AP^{bb}) and mask mAP (AP^{mk}). **Right: Finetuning results on the Amazon package segmentation dataset with representations pretrained on the Amazon dataset.** We observe that Dec-SSL reaches similar performance (AP^{mk}) as centralized SSL and also outperforms training from scratch. Note that $100\%, 10\%, 1\%$ denote the portion of the data used for finetuning.

improve some analysis therein, and design new metrics to characterize the performance adapted to the decentralized setting. Due to space limitation, we include an abridged introduction here, and defer more details to Appendix §E.

Setup. Consider a Dec-SSL problem with K data sources. Similar to the SimSiam approach, we first augment x, an anchor sample from the dataset, by sampling $\xi, \xi' \sim \mathcal{N}(0, I)$ IID from the Gaussian distribution. Consider the linear embedding function $f_w(x) = wx$, where $w \in \mathbb{R}^{m \times d}$ and $m \geq 2K$. The SSL objective for data source k is given by

$$\mathcal{L}_k(w) := -\widehat{\mathbb{E}}\big[(w(x_{k,i} + \xi_{k,i}))^\top (w(x_{k,i} + \xi'_{k,i}))\big] + \frac{1}{2}\|w^\top w\|_F^2, \tag{3.1}$$

where $\widehat{\mathbb{E}}$ is taken expectation over the empirical dataset $x_{k,i} \sim D_k$, and the randomness of $\xi_{k,i}$ and $\xi'_{k,i}$. Moreover, recall the global objective is given in (2.2). Note that (3.1) instantiates SimSiam loss with the negative inner-product $\langle a, b \rangle$ as the distance function $\mathbb{D}(a, b)$ and no feature predictor, and with a regularization term for mathematical tractability, as in Liu et al. (2021).

Data heterogeneity. The K data sources collaboratively solve (2.2) to learn a representation for a $2K$-way classification task. The K local datasets are generated in a way that for each fixed $k \in [K]$, the labels are skewed in that data from classes $2k - 1$ and $2k$ constitute the majority of the data, while other classes are rare, or even unseen. More details on the specifications of data heterogeneity can be found in §E.1. We visualize the heterogeneity of the data distributions in Figure 3.

To compare the representations learned across data sources and that learned from jointly solving (2.2), we introduce the following definition on the representability of the representation space.

Definition 3.1 (Representability vector)**.** Let $\mathcal{S} \subseteq \mathbb{R}^d$ be the subspace spanned by the rows of the learned feature matrix $w \in \mathbb{R}^{m \times d}$, where the embedding function $f_w(x) = wx$. The *representability* of $\mathcal{S}$ is defined as a vector $\boldsymbol{r} = [r_1, \cdots, r_d]^\top \in \mathbb{R}^d$, such that $r_i = \|\Pi_{\mathcal{S}}(e_i)\|_2^2$ for $i \in [d]$, where $\Pi_{\mathcal{S}}(e_i) \in \mathbb{R}^d$ is the projection of standard basis e_i onto $\mathcal{S}$, and thus $r_i = \sum_{j=1}^{s} \langle e_i, v_j \rangle^2$ where $s = \dim(\mathcal{S})$ and $\{v_1, \cdots, v_s\}$ is a set of orthonormal bases for $\mathcal{S}$.

The intuition of this definition is that a good feature space should have the property that many standard unit bases among $e_1, \cdots, e_d$, which can be used to represent any vectors in $\mathbb{R}^d$, can be represented well by the feature space, i.e., have large projections onto it. Note that as a vector, $\boldsymbol{r}$ provides a quantitative way to compare the representability of two feature spaces across different directions (i.e., different unit basis). In the following theorem, we compare the representability learned by local objectives and the global one, for Dec-SSL.

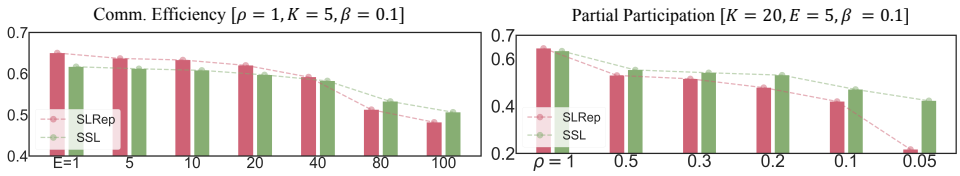

Figure 4: **Dec-SSL performance on ImageNet-100 dataset**. Compared to supervised learning, we observe that under non-IID settings, decentralized SSL can perform better under communication constraints (left) and partial participation constraints (right).

Theorem 3.2 (Representability of local v.s. global objectives for Dec-SSL)**.** For decentralized SSL in the setting described above, with high probability, the representability vector learned from any local objective of source k, denoted by $\boldsymbol{r}^k = [r_1^k, \cdots, r_d^k]^\top$, satisfies that $1 - O(d^{-4/5}) \leq r_i^k \leq 1$ for all $i \in [K] \backslash \{k\}$. Moreover, the representability vector learned from the global objective, denoted by $\bar{\boldsymbol{r}} = [\bar{r}_1, \cdots, \bar{r}_d]^\top$, satisfies that $1 - O(d^{-4/5}) \leq \bar{r}_i \leq 1$ for all $i \in [K]$.

Theorem 3.2 states that the feature spaces learned from local SSL objectives are relatively *uniform*, in the sense that for the K basis directions $e_1, \cdots, e_K$ that generate the data, any two data sources have similar representability in *all of them but two* directions, especially when the dimension d of the data is large. Furthermore, when solving the global objective (2.2), the learned representation is also uniform, and its representability differs *at most one* direction from that of each local data source. Note that the results hold with highly heterogeneous data across data sources. In other words, Dec-SSL is not affected significantly by the non-IIDness of the data, justifying the empirical observations in §3.1. Illustration of the results can also be found in Figure 3.

Intuition & implication. The main intuition behind Theorem 3.2 is that, the objective of SSL is not *biased* by the heterogeneous distribution of labels at each local dataset, and tends to learn uniform representations. Related arguments have also been made in the recent works on the theoretical understanding contrastive learning/SSL (Wang & Isola, 2020; Liu et al., 2021). In the decentralized setting, this insensitivity to data heterogeneity becomes even more relevant, as it potentially allows each local data source to perform much more local updates, without drifting the iterates significantly. This enables more communication-efficient decentralized learning schemes, in contrast to most existing ones that are vulnerable to data non-IIDness. We validate these points next.

4 Dec-SSL Can be Favorable Even When Labels are Available

We here seek to address question (ii) in §2.1 – how does the unique property of Dec-SSL, such as the robustness to data heterogeneity, benefit decentralized learning? While lack of labels seems a limitation, we show that this might not be the case in decentralized learning with heterogeneous data. First, it is known that decentralized SL in general performs poorly when the data is highly heterogeneous (Zhao et al., 2018; Hsieh et al., 2020). Further, even in the decentralized representation learning setting when labels are available, Dec-SSL still stands out in face of highly non-IID data.

To make a fair comparison, we mainly compare Dec-SSL with Dec-SLRep (recall the definition in §2.1), which are both decentralized *representation learning* approaches. We defer the comparison with Dec-SL to Appendix §B. We conduct experiments on both ImageNet and CIFAR-10 datasets, and evaluate the performance of the learned representations in terms of the variations of two commonly used metrics in decentralized learning – the number of local updates epochs E, and the participation ratio of data sources ρ. We observe consistently that Dec-SSL indeed outperforms Dec-SLRep in learning representations in terms of communication efficiency and participation ratio, especially with highly non-IID data. We remark that such observations are also consistent with those on object detection and semantic segmentation given in Table 1.

4.1 Experimental observations

In this experiment, we train and evaluate the feature backbone on ImageNet-100 in a decentralized setting. We create non-IIDness across the local datasets based on label skewness and use $\beta = 0.1$ (each data source has only 10% of its data coming from the uniform class distributions).

Communication efficiency under high non-IIDness. In Figure 4, we show that under the non-IID scenario, averaging weights with an infrequent communication schedule causes less trouble to Dec-SSL than to Dec-SLRep. In `FedAvg`, the idea of averaging weights after multiple epochs might sound sub-optimal, but we notice that decentralized SSL is very robust with respect to this parameter. Intuitively, the robustness of Dec-SSL allows each local model to drift longer, leading to a lower communication frequency for decentralized learning.

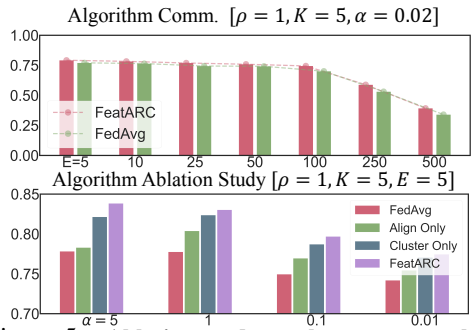

Figure 5: **Ablation study on the `FeatARC` algorithm.** We observe that under non-IIDness and communication constraints, `FeatARC` outperforms the baseline variants of the algorithm and `FedAvg`.

Method / Setting	IID	non-IID
FURL (Zhang et al., 2020a)	71.25	68.01
EMA (Zhuang et al., 2022)	86.26	83.34
Per-SSFL (He et al., 2021a)	N/A	83.10
FEDU (Zhuang et al., 2021)	83.96	80.52
`FeatARC` (Ours)	**86.74**	**84.63**

CIFAR-100 Pretrain	CIFAR-10 100%	CIFAR-10 10%	CIFAR-10 1%
no pretrain	0.31	0.27	0.25
Dec-SLRep IID	0.65	0.60	0.47
Dec-SSL IID	**0.71**	**0.67**	**0.57**
Dec-SLRep Non-IID	0.43	0.35	0.32
Dec-SSL Non-IID	**0.70**	**0.66**	**0.57**

Table 2: **Top). Algorithm performance comparison. Bottom). CIFAR-10 Linear probing on the representation of CIFAR-100.** Our algorithm surpasses previous works on federated SSL both in the IID and non-IID settings.

Participation ratio under high non-IIDness. In this experiment, we split ImageNet-100 into 20 data sources and use local update $E = 5$ epochs. We measure the performance of decentralized learning algorithms with respect to the participation ratio of data sources at each round. For instance, when $\rho = 1$, at each round, all data sources update their local weights and upload to the server, while $\rho = 0.05$ means that each round a single random data source is selected for update. On the right of Figure 4, we show that with non-IID data, the convergence of Dec-SSL is more stable to less participants compared to Dec-SLRep. This allows more efficient decentralized learning, especially when deployed with extremely large number of data sources and unstable communication channels.

4.2 Theoretical Insights

To shed light on the above observations, we provide analysis for the feature spaces learned by the local objective of Dec-SLRep, under the same setup as in §3.2. For Dec-SLRep and each data source k, we consider learning a two-layer linear network $g_{u_k,v_k}(x) := v_k u_k x$ as classifier, where $u_k \in \mathbb{R}^{m\times d}$ and $v_k \in \mathbb{R}^{c\times m}$, and use $u_k x$ as the learned representation for downstream tasks. The network is learned by minimizing $\|(u_k)^\top u_k\|_F^2 + \|(v_k)^\top v_k\|_F^2$ subject to the margin constraint that $[g_{u_k,v_k}(x)]_y \geq [g_{u_k,v_k}(x)]_{y'} + 1$ for all data (x, y) in the local dataset k with all $y' \neq y$. . We now have the following proposition on the representations learned by Dec-SLRep across data sources.

Proposition 4.1 (Representations learned by Dec-SLRep across heterogeneous data sources)**.** With high probability, the features $u_k = [u_{k,1}, \cdots, u_{k,m}]^\top \in \mathbb{R}^{m\times d}$ learned from the local dataset D_k satisfies that $\sum_{i=1}^m \langle u_{k,i}, e_j\rangle^2 \leq O(d^{-\frac{1}{10}})$, for $j \in [K]\backslash\{k\}$; while $\sum_{i=1}^m \langle u_{k,i}, e_k\rangle^2 \geq 1 - O(d^{-\frac{1}{20}})$. In other words, the correlation between the learned features in w_k and e_j is small for all $j \in [K] \setminus \{k\}$, while the correlation between the features and e_k is large.

The proposition suggests that the feature spaces learned by Dec-SLRep differ significantly across local data sources, given the highly heterogeneous data. More specifically, we show that most of the unit bases in $\{e_1, \cdots, e_K\}$ have small correlations with the features learned at each local data source, while these feature spaces themselves vary significantly across data sources. The unit bases that are not learned might be significant for various other downstream tasks, making the learned representations less favorable. This heterogeneity among local solutions is not in favor of *local updates*, as too many local updates would drift the iterates towards its local solution, and the iterates would become too far away from each other, hurting the convergence of decentralized learning. Hence, compared with the Dec-SSL case and Theorem 3.2, Dec-SLRep can be less robust to data heterogeneity and less communication-efficient. We note that the advantage of Dec-SSL does not come from *using more data*, since we use exactly the same data for training Dec-SLRep and Dec-SSL. The intuition is also illustrated in Figure 3. Finally, we remark that the *uniformity* of features, which is believed to be the key to better transfer performance in SSL (Wang & Isola, 2020; Caron et al., 2020), is not always preferred given *specific* learning tasks (Burgess et al., 2018).

5 Our Algorithm – `FeatARC` (Feature Alignment and Clustering)

Although Dec-SSL tends to learn relatively uniform features that are robust across datasets, the uniformity itself might not imply the alignment of features across datasets: the representation network from different local data sources can still map the same data point to different regions in the feature space. This misalignment becomes more significant when the data is highly non-IID and can have an adverse effect on the model aggregation process in decentralized learning (Zhang et al., 2020a).

To mitigate this issue and address question (iii) in §2.1, we propose to use the same feature distance loss as an auxiliary local objective to align the local models with the global model. The alignment between two features is defined as the negative cosine distance metric $\mathbb{D}(z_1, z_2) = -\frac{z_1 \cdot z_2}{||z_1|| ||z_2||}$.

To further improve the Dec-SSL algorithm, we propose to learn *multiple models* using clustering-based approach. In particular, instead of learning a single global model as in (2.2), we learn C models and separate the K data sources into C clusters. The update of C models and the assignment of data sources to C clusters are conducted alternatively. When $C = K$, the algorithm reduces to learning K local models; when $C = 1$, it reduces to learning a single global one. The clustering approach intuitively learns multiple models to interpolate the performance between learning a *single global* model and K *local* models, thus achieving a good bias-variance tradeoff when testing on each local dataset (Mansour et al., 2020; Ghosh et al., 2020). However, unlike the supervised learning case, we do not use the loss of the decentralized learning (i.e., (2.1)) as the metric for clustering. This is because for contrastive learning, it has been observed that the SSL loss might not be indicative enough for the performance of the representation on downstream tasks (Robinson et al., 2021). Hence, we here again use the feature alignment distance $\mathbb{D}(\cdot, \cdot)$ as the metric for clustering.

We adopt the alignment regularization and clustering techniques, and developed a new Dec-SSL algorithm `FeatARC`, summarized in Algorithm 1 and Algorithm 2 in Appendix. We show the performance of `FeatARC` in Figure 5, in comparison with different baselines including `FedAvg`, under different levels of data heterogeneity and communication frequency. It is shown that `FeatARC` outperforms the baselines consistently, including the variants that only uses alignment ("Align Only") or clustering ("Cluster Only"). Moreover, on the top of Table 2, we show that `FeatARC` also outperforms other recent decentralized self-supervised learning algorithms on CIFAR-10 dataset.

6 Extensions

In this section, we discuss a few extended experiments of our framework. Please see Appendix §B for a thorough set of experiments and ablation studies with visualizations.

6.1 Fully decentralized case and different network topology

We conduct experiments on the *fully decentralized* learning in Appendix §B.5, where the local data sources are only allowed to communicate with their neighbors over a peer-to-peer network, without a centralized server. In short, most observations we had regarding Dec-SSL in the setting with a centralized server still hold, even under several different network topologies. This aligns with our theoretical insight provided in Section 3, which came from the benign properties of the *solution* to the Dec-SSL *objective*, instead of the properties of *specific algorithms* (averaging the iterates via a star or other network topologies) that achieves the solution.

6.2 Extremely heterogeneous case for decentralized learning

In Figure 13, we show that even in the extremely heterogeneous case where each local source only owns *one* class, the Dec-SSL framework is still robust to the non-IIDness of the data. This also holds true when we scale to more clients, as shown in Figure 15. The Dec-SSL objective would not be *biased* by the highly heterogeneous class labels at each local dataset, while the Dec-SL objective could be biased by it. This is also consistent with our theoretical insights in Section §3.2 and the key reason for the success of Dec-SSL is that, despite only having one single class, the information of features obtained from local datasets may still be useful for the jointly classifying of all the classes.

6.3 Comparison of FeatARC with other algorithms

We also compare our algorithm with the Dec-SSL algorithms that are combined with other federated learning algorithms, including Li et al. (2020a) (FedProx) and Li et al. (2020b) (FedBN). In Figure 16 (Left), we show that our proposed `FeatARC` can outperform these two baselines.

7 Conclusion

We propose the framework of decentralized SSL that learns representations from non-IID unlabeled data and conduct an empirical study on the robustness of Dec-SSL to different types of heterogeneity, communication constraints, and participation rates of data sources. We also provide findings and theoretical analyses of Dec-SSL compared to its supervised learning counterpart, as well as developing a new algorithm to further address the high heterogeneity in decentralized datasets.

Acknowledgement. This work is supported in part by Amazon.com Services LLC, PO2D-06310236 and Defense Science & Technology Agency, DST00OECI20300823. L.W. was supported by the MIT EECS Xianhong Wu Graduate Fellowship. K.Z. also acknowledges support from Simons-Berkeley Research Fellowship. We thank MIT Supercloud for providing compute resources. The authors would like to thank many helpful discussions from Phillip Isola at MIT and Andrew Marchese at Amazon.

REFERENCES

Sanjeev Arora, Hrishikesh Khandeparkar, Mikhail Khodak, Orestis Plevrakis, and Nikunj Saunshi. A theoretical analysis of contrastive unsupervised representation learning. *arXiv preprint arXiv:1902.09229*, 2019.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Christopher P Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in backslash beta-vae. *arXiv preprint arXiv:1804.03599*, 2018.

Mathilde Caron, Piotr Bojanowski, Julien Mairal, and Armand Joulin. Unsupervised pre-training of image features on non-curated data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2959–2968, 2019.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems*, 33:9912–9924, 2020.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15750–15758, 2021.

Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 215–223. JMLR Workshop and Conference Proceedings, 2011.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE international conference on computer vision*, pp. 1422–1430, 2015.

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

Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.

Peter R Florence, Lucas Manuelli, and Russ Tedrake. Dense object nets: Learning dense visual object descriptors by and for robotic manipulation. *arXiv preprint arXiv:1806.08756*, 2018.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bygh9j09KX.

Avishek Ghosh, Jichan Chung, Dong Yin, and Kannan Ramchandran. An efficient framework for clustered federated learning. *Advances in Neural Information Processing Systems*, 33:19586–19597, 2020.

Ross Girshick, Ilija Radosavovic, Georgia Gkioxari, Piotr Dollár, and Kaiming He. Detectron. https://github.com/facebookresearch/detectron, 2018.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in Neural Information Processing Systems*, 33:21271–21284, 2020.

Jeff Z HaoChen, Colin Wei, Adrien Gaidon, and Tengyu Ma. Provable guarantees for self-supervised deep learning with spectral contrastive loss. *Advances in Neural Information Processing Systems*, 34, 2021.

Chaoyang He, Zhengyu Yang, Erum Mushtaq, Sunwoo Lee, Mahdi Soltanolkotabi, and Salman Avestimehr. Ssfl: Tackling label deficiency in federated learning via personalized self-supervision. *arXiv preprint arXiv:2110.02470*, 2021a.

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

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, 2017.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. *arXiv preprint arXiv:2111.06377*, 2021b.

Dan Hendrycks, Mantas Mazeika, Saurav Kadavath, and Dawn Song. Using self-supervised learning can improve model robustness and uncertainty. *Advances in Neural Information Processing Systems*, 32, 2019.

Kevin Hsieh, Aaron Harlap, Nandita Vijaykumar, Dimitris Konomis, Gregory R Ganger, Phillip B Gibbons, and Onur Mutlu. Gaia: {Geo-Distributed} machine learning approaching {LAN} speeds. In *USENIX Symposium on Networked Systems Design and Implementation (NSDI 17)*, pp. 629–647, 2017.

Kevin Hsieh, Amar Phanishayee, Onur Mutlu, and Phillip Gibbons. The non-iid data quagmire of decentralized machine learning. In *International Conference on Machine Learning*, pp. 4387–4398. PMLR, 2020.

Wonyong Jeong, Jaehong Yoon, Eunho Yang, and Sung Ju Hwang. Federated semi-supervised learning with inter-client consistency & disjoint learning. In *International Conference on Learning Representations*, 2020.

Ziwei Ji and Matus Telgarsky. Gradient descent aligns the layers of deep linear networks. In *International Conference on Learning Representations*, 2018.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pp. 4904–4916. PMLR, 2021.

Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210, 2021.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning*, pp. 5132–5143. PMLR, 2020.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Jakub Konečnỳ, H Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*, 2016.

Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-10 and cifar-100 datasets. *URl: https://www. cs. toronto. edu/kriz/cifar. html*, 6(1):1, 2009.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Learning representations for automatic colorization. In *European conference on computer vision*, pp. 577–593. Springer, 2016.

Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.

Jason D Lee, Qi Lei, Nikunj Saunshi, and Jiacheng Zhuo. Predicting what you already know helps: Provable self-supervised learning. *Advances in Neural Information Processing Systems*, 34, 2021.

Qinbin Li, Yiqun Diao, Quan Chen, and Bingsheng He. Federated learning on non-iid data silos: An experimental study. *arXiv preprint arXiv:2102.02079*, 2021a.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine Learning and Systems*, 2:429–450, 2020a.

Xiaoxiao Li, Meirui JIANG, Xiaofei Zhang, Michael Kamp, and Qi Dou. FedBN: Federated learning on non-iid features via local batch normalization. In *International Conference on Learning Representations*, 2020b.

Xiaoxiao Li, Meirui Jiang, Xiaofei Zhang, Michael Kamp, and Qi Dou. Fedbn: Federated learning on non-iid features via local batch normalization. *arXiv preprint arXiv:2102.07623*, 2021b.

Xiangru Lian, Ce Zhang, Huan Zhang, Cho-Jui Hsieh, Wei Zhang, and Ji Liu. Can decentralized algorithms outperform centralized algorithms? A case study for decentralized parallel stochastic gradient descent. *Advances in Neural Information Processing Systems*, 30, 2017.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Hong Liu, Jeff Z. HaoChen, Adrien Gaidon, and Tengyu Ma. Self-supervised learning is more robust to dataset imbalance. In *NeurIPS 2021 Workshop on Distribution Shifts: Connecting Methods and Applications*, 2021. URL https://openreview.net/forum?id=vUz4JPRLpGx.

Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. 2018.

Nan Lu, Zhao Wang, Xiaoxiao Li, Gang Niu, Qi Dou, and Masashi Sugiyama. Federated learning from only unlabeled data with class-conditional-sharing clients. In *International Conference on Learning Representations*, 2022.

Disha Makhija, Nhat Ho, and Joydeep Ghosh. Federated self-supervised learning for heterogeneous clients. *arXiv preprint arXiv:2205.12493*, 2022.

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

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pp. 1273–1282. PMLR, 2017.

Angelia Nedic. Distributed gradient methods for convex machine learning problems in networks: Distributed optimization. *IEEE Signal Processing Magazine*, 37(3):92–101, 2020.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2536–2544, 2016.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.

Joshua Robinson, Li Sun, Ke Yu, Kayhan Batmanghelich, Stefanie Jegelka, and Suvrit Sra. Can contrastive learning avoid shortcut solutions? *Advances in Neural Information Processing Systems*, 34, 2021.

Justyna Sarzynska-Wawer, Aleksander Wawer, Aleksandra Pawlak, Julia Szymanowska, Izabela Stefaniak, Michal Jarkiewicz, and Lukasz Okruszek. Detecting formal thought disorder by deep contextualized word representations. *Psychiatry Research*, 304:114135, 2021.

Pierre Sermanet, Corey Lynch, Yevgen Chebotar, Jasmine Hsu, Eric Jang, Stefan Schaal, Sergey Levine, and Google Brain. Time-contrastive networks: Self-supervised learning from video. In *2018 IEEE international conference on robotics and automation (ICRA)*, pp. 1134–1141. IEEE, 2018.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *European conference on computer vision*, pp. 776–794. Springer, 2020a.

Yonglong Tian, Yue Wang, Dilip Krishnan, Joshua B Tenenbaum, and Phillip Isola. Rethinking few-shot image classification: a good embedding is all you need? In *European Conference on Computer Vision*, pp. 266–282. Springer, 2020b.

Christopher Tosh, Akshay Krishnamurthy, and Daniel Hsu. Contrastive learning, multi-view redundancy, and linear models. In *Algorithmic Learning Theory*, pp. 1179–1206. PMLR, 2021.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pp. 1096–1103, 2008.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pp. 9929–9939. PMLR, 2020.

Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3733–3742, 2018.

Mikhail Yurochkin, Mayank Agarwal, Soumya Ghosh, Kristjan Greenewald, Nghia Hoang, and Yasaman Khazaeni. Bayesian nonparametric federated learning of neural networks. In *International Conference on Machine Learning*, pp. 7252–7261. PMLR, 2019.

Fengda Zhang, Kun Kuang, Zhaoyang You, Tao Shen, Jun Xiao, Yin Zhang, Chao Wu, Yueting Zhuang, and Xiaolin Li. Federated unsupervised representation learning. *arXiv preprint arXiv:2010.08982*, 2020a.

Michael Zhang, Karan Sapra, Sanja Fidler, Serena Yeung, and Jose M Alvarez. Personalized federated learning with first order model optimization. *arXiv preprint arXiv:2012.08565*, 2020b.

Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European conference on computer vision*, pp. 649–666. Springer, 2016.

Yue Zhao, Meng Li, Liangzhen Lai, Naveen Suda, Damon Civin, and Vikas Chandra. Federated learning with non-iid data. *arXiv preprint arXiv:1806.00582*, 2018.

Weiming Zhuang, Xin Gan, Yonggang Wen, Shuai Zhang, and Shuai Yi. Collaborative unsupervised visual representation learning from decentralized data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4912–4921, 2021.

Weiming Zhuang, Yonggang Wen, and Shuai Zhang. Divergence-aware federated self-supervised learning. In *International Conference on Learning Representations*, 2022.