

# LW-FedSSL: Resource-efficient Layer-wise Federated Self-supervised Learning

Ye Lin Tun<sup>a</sup>, Chu Myaet Thwal<sup>a</sup>, Huy Q. Le<sup>a</sup>, Minh N. H. Nguyen<sup>b</sup>, Choong Seon Hong<sup>a,\*</sup>

<sup>a</sup>Department of Computer Science and Engineering, Kyung Hee University, Yongin-si, Gyeonggi-do 17104, South Korea

<sup>b</sup>Vietnam - Korea University of Information and Communication Technology, Danang, Vietnam

## Abstract

Many studies integrate federated learning (FL) with self-supervised learning (SSL) to take advantage of raw data distributed across edge devices. However, edge devices often struggle with high computational and communication costs imposed by SSL and FL algorithms. With the deployment of more complex and large-scale models, such as Transformers, these challenges are exacerbated. To tackle this, we propose the Layer-Wise Federated Self-Supervised Learning (LW-FedSSL) approach, which allows edge devices to incrementally train a small part of the model at a time. Specifically, in LW-FedSSL, training is decomposed into multiple stages, with each stage responsible for only a specific layer (or a block of layers) of the model. Since only a portion of the model is active for training at any given time, LW-FedSSL significantly reduces computational requirements. Additionally, only the active model portion needs to be exchanged between the FL server and clients, reducing the communication overhead. This enables LW-FedSSL to jointly address both computational and communication challenges in FL. Depending on the SSL algorithm used, it can achieve up to a 3.34 $\times$  reduction in memory usage, 4.20 $\times$  fewer computational operations (GFLOPs), and a 5.07 $\times$  lower communication cost while maintaining performance comparable to its end-to-end training counterpart. Furthermore, we explore a progressive training strategy called Prog-FedSSL, which offers a 1.84 $\times$  reduction in GFLOPs and a 1.67 $\times$  reduction in communication costs while maintaining the same memory requirements as end-to-end training. While the resource efficiency of Prog-FedSSL is lower than that of LW-FedSSL, its performance improvements make it a viable candidate for FL environments with more lenient resource constraints.

**Keywords:** federated learning, self-supervised learning, layer-wise training, resource-efficient.

## 1. Introduction

A significant portion of real-world data, valuable for practical AI applications, resides on edge devices. Such data is often unlabeled, making the adoption of self-supervised learning (SSL) strategies [1, 2, 3] essential. SSL enables models to learn from raw data by generating their own supervisory signals, minimizing the need for labeled datasets. Most SSL methods follow centralized training schemes, often requiring an exhaustive data collection process to build a centrally stored dataset. While some types of data, such as natural images, can be collected with relative ease, privacy-sensitive data—such as medical records—raise significant privacy concerns. Meanwhile, decentralized learning approaches like federated learning (FL) [4, 5] enable privacy-preserving collaborative model training. An FL system operates by collecting trained model parameters from edge devices (a.k.a., clients) instead of their data, thereby preserving data privacy. This mechanism helps safeguard the confidentiality of sensitive information held by different parties. As a result, federated self-supervised learning (FedSSL) has recently emerged as a promising approach for leveraging raw data in distributed environments [6, 7, 8].

Self-supervised learning (SSL), particularly with state-of-the-art models like Transformers [9, 10], imposes substantial computational demands. Unfortunately, clients in an FL environment often operate with limited computational and communication resources. Many clients may lack the necessary resources to participate in a conventional end-to-end FedSSL process. As a result, data from these clients is excluded from the training process, potentially degrading overall model performance. Therefore, it is crucial to explore resource-efficient FedSSL approaches that enable every FL client to participate in collaborative training. A straightforward way to reduce computational demands on low-memory devices is to conduct the training with a smaller batch size. However, using a small batch size in SSL can diminish the quality of learned representations and compromise performance [1, 7]. Moreover, simply reducing the batch size does not lower the FL communication costs, as clients still need to exchange the full model with the server. On the other hand, using a large batch size would exclude low-memory devices from participating in the FL process.

To address these challenges, we introduce LW-FedSSL, a layer-wise training approach where a model is systematically trained one layer at a time. The term “*layer*” in this context can refer to either an individual layer or a block of multiple layers within the model. In LW-FedSSL, the training process is divided into multiple stages, with each stage focusing on a specific portion of the model. At the beginning of each stage,

\*Corresponding author

Email addresses: yelintun@khu.ac.kr (Ye Lin Tun), chumyaet@khu.ac.kr (Chu Myaet Thwal), quanghuy69@khu.ac.kr (Huy Q. Le), nhnminh@vku.udn.vn (Minh N. H. Nguyen), cshong@khu.ac.kr (Choong Seon Hong)

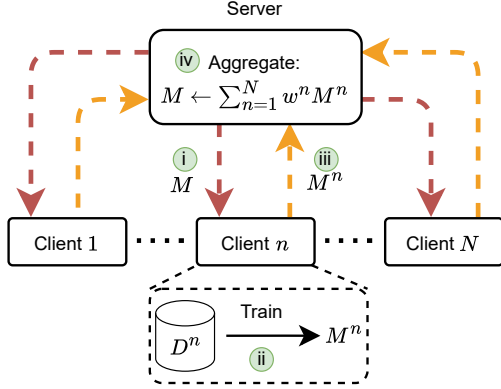


Figure 1: Four key steps in a single FL communication round. (i) The server distributes a base global model  $M$  to all participating clients. (ii) Each client  $n \in [1, N]$  trains the model on its local dataset  $D^n$  to produce a local model  $M^n$ . (iii) The local models are then transmitted back to the server. (iv) Finally, the server aggregates the received local models using weighted averaging to update the global model:  $M \leftarrow \sum_{n=1}^N w^n M^n$  [5].

a new layer is sequentially added to the model, while all previously trained layers from prior stages (if any exists) are frozen, preventing further updates. Only the newly added layer remains active for training, significantly reducing computational requirements for FL clients. Additionally, since only the active layer within a stage needs to be exchanged between the server and FL clients, communication bottlenecks are substantially reduced. As training progresses with each stage, the model depth increases gradually, enabling an incremental and efficient learning process. We also explore a progressive training strategy, Prog-FedSSL, which follows a similar approach to layer-wise training. Like LW-FedSSL, Prog-FedSSL divides the training process into multiple stages, with each stage sequentially adding a new layer to the model. However, instead of freezing previously trained layers, Prog-FedSSL keeps all existing layers active during each stage, allowing the entire sub-model to continue updating. As a result, Prog-FedSSL may require more computational and communication resources than LW-FedSSL, but it has the potential to improve model performance.

Our contributions can be summarized as follows:

- We introduce LW-FedSSL, a novel layer-wise federated self-supervised learning framework designed to significantly enhance the resource efficiency of FL clients. LW-FedSSL can compete with conventional end-to-end training counterparts while simultaneously reducing both computational and communication costs.
- We explore Prog-FedSSL, a progressive training strategy for federated self-supervised learning, that holds the potential to enhance performance. Given its lower resource efficiency compared to LW-FedSSL, Prog-FedSSL is better suited for FL environments with more flexible resource constraints.
- We provide an in-depth exploration of LW-FedSSL and Prog-FedSSL within the FL paradigm, considering various empirical aspects such as computational efficiency, communication overhead, and model performance. We present

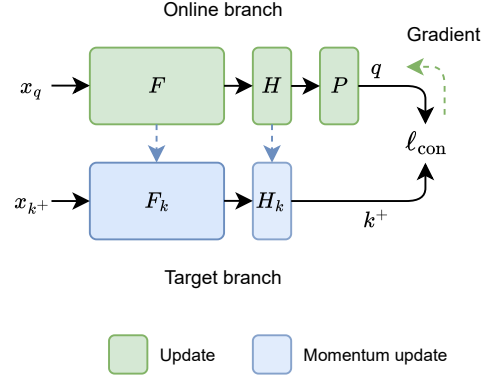


Figure 2: Self-supervised learning with MoCoV3 [11]. We omit the negative samples for clarity.

a thorough evaluation of our proposed approaches across different FL settings and benchmarks.

## 2. Background and Related Work

In this section, we provide a brief overview of key concepts in federated learning and self-supervised learning, along with a discussion of related work.

### 2.1. Federated Learning

A typical FL process involves a central coordinating server and a set of client devices, each storing private data [5]. The goal of FL is to train a global model  $M$ , by learning from the data residing on clients while ensuring their privacy. The FL objective can be expressed as:

$$M^* = \arg \min_M \sum_{n=1}^N w^n \mathcal{L}(M, D^n), \quad (1)$$

where  $N$  is the number of clients, and  $\mathcal{L}$  represents the local loss function. Here,  $w^n = |D^n|/|D|$  is the weight assigned to the  $n$ -th client, with  $|D^n|$  denoting the number of samples in the local dataset  $D^n$ , and  $|D|$  denoting the total number of samples across all clients, i.e.,  $D = \bigcup_{n=1}^N D^n$ . As shown in Fig. 1, a single FL communication round consists of four main steps, which are repeated for a total of  $R$  rounds.

### 2.2. Self-supervised Learning

Self-supervised learning (SSL) offers a way to utilize unlabeled data for model training, which is often more readily available than labeled data. SSL generates its own supervisory signals to learn valuable representations from the unlabeled data. In this work, we primarily use MoCoV3 [11] as the SSL approach.<sup>1</sup> As shown in Fig. 2, MoCoV3 features a Siamese network structure, consisting of an actively trained online branch, and a target branch. The online branch includes an encoder  $F$ , a projection head  $H$ , and a prediction head  $P$ . Meanwhile, the

<sup>1</sup>Unlike in [11], we train the patch projection layer.

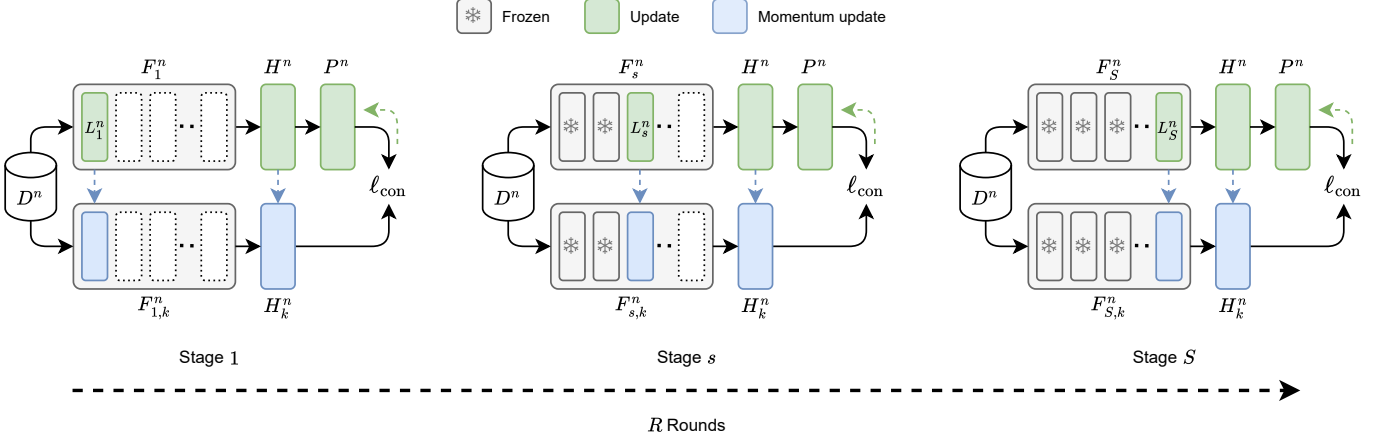


Figure 3: LW-FedSSL: Local training process across different stages for the  $n$ -th client. At the beginning of each stage  $s \in [1, S]$ , a new layer  $L_s$  is sequentially added to the previous encoder  $F_{(s-1)}$ , increasing its depth. During stage  $s$ , only the corresponding layer  $L_s$  is actively updated, while all prior layers (i.e.,  $L_1$  to  $L_{(s-1)}$ ) are kept frozen.

target branch is a moving-averaged copy of the online branch, consisting of a momentum encoder  $F_k$  and a momentum projection head  $H_k$ .

Given an input sample  $x$ , it undergoes augmentation, creating two views,  $x_q$  and  $x_{k^+}$ . These views are then fed into the online and target branches, respectively, to obtain representations  $q$  and its corresponding positive pair,  $k^+$ . Likewise, negative pairs,  $k^-$ , are derived from other samples within the same batch through the target branch. The online branch is trained using the InfoNCE loss [12], defined as:

$$\ell_{\text{con}}(q, k, \tau) = -\log \frac{\exp(q \cdot k^+ / \tau)}{\exp(q \cdot k^+ / \tau) + \sum_{j=1}^K \exp(q \cdot k_j^- / \tau)}, \quad (2)$$

where  $k = (k^+, \{k_j^-\}_{j=1}^K)$  and  $\tau$  is the temperature parameter. Using a momentum parameter  $\mu$ , the target branch is updated as:  $F_k \leftarrow \mu F_k + (1 - \mu)F$  and  $H_k \leftarrow \mu H_k + (1 - \mu)H$ .

### 2.3. Federated Self-supervised Learning

While most studies on FL focus on supervised learning, this requires device owners to label their data—a labor-intensive and often impractical task. As a result, client devices often lack labeled data due to limitations in data collection, labeling resource constraints, and insufficient annotation expertise. Federated self-supervised learning (FedSSL) extends the applicability of FL to new domains where labeled data is scarce or difficult to acquire [13, 14, 15]. Many FedSSL studies build upon state-of-the-art SSL frameworks, such as SimCLR [1], MoCo [2], and BYOL [16]. These studies adapt SSL techniques to address specific FL tasks [17, 18] or tackle FL challenges at hand [19, 20, 21, 6]. In the context of FedSSL, the local dataset  $D^n$  within a client  $n$  is unlabeled, and thus, the local training process relies on a form of SSL loss to train the local model  $M^n$ . When using MoCoV3 as the SSL framework, the local model consists of an encoder  $F^n$  and its accompanying MLP heads,  $H^n$  and  $P^n$ .

### 2.4. Layer-wise Training

Layer-wise training was initially introduced in the context of deep belief networks and restricted Boltzmann machines [22, 23]. The original motivation behind its development was to address challenges associated with training deep neural networks, such as vanishing gradients. While its usage has become less common over the years, we believe it holds great potential to be rediscovered as a highly resource-efficient strategy within FL systems. Incrementally training a model one layer at a time offers a significant resource-saving advantage over end-to-end training. Despite its potential benefits, layer-wise training remains largely unexplored in FL, with very few related studies [24, 25].

The study in [26] investigates layer-wise SSL in the speech domain. Although it was intended for an on-device training scenario, the authors mainly examined it in a centralized setting. A short study [25] on federated layer-wise learning (FLL) introduces a depth dropout technique that randomly drops frozen layers during training to reduce resource overhead. However, FLL is only examined with a large number of communication rounds in each training stage (ranging from 4k to 12k), potentially placing significant strain on clients. Additionally, the exploration of FLL is limited in scope, missing comprehensive evaluations across diverse FL settings and benchmark datasets. In contrast to these studies, our work aims to offer a more thorough and in-depth exploration of layer-wise training through LW-FedSSL. One study [24] investigates the progressive training approach for FL known as ProgFed. Although similar to layer-wise training in gradually increasing the model depth, ProgFed [24] differs by training all existing layers at each stage. Since ProgFed primarily focuses on supervised learning tasks that require labeled data, the nature of progressive training within the FedSSL paradigm remains unexplored. To fill this gap, we introduce a progressive training strategy for FedSSL, referred to as Prog-FedSSL.

---

**Algorithm 1** LW-FedSSL (*Server-side*)

---

**Input:** encoder  $F_0$ , projection head  $H$ , prediction head  $P$ , number of stages  $S$ , number of rounds per stage  $R_s$  where  $s \in [1, S]$

**Output:** encoder  $F_S$

```
1: Server executes:
2: Distribute  $F_0$  to clients
3: for stage  $s = 1, 2, \dots, S$  do
4:   Initialize new layer:  $L_s$ 
5:   for round  $r = 1, 2, \dots, R_s$  do
6:     for client  $n = 1, 2, \dots, N$  in parallel do
7:        $L_s^n, H^n, P^n \leftarrow \text{Train}(n, L_s, H, P)$ 
8:        $w^n = \frac{|D^n|}{|D|}$ , where  $D = \bigcup_{n=1}^N D^n$ 
9:     end for
10:     $L_s \leftarrow \sum_{n=1}^N w^n L_s^n$ 
11:     $H \leftarrow \sum_{n=1}^N w^n H^n$ 
12:     $P \leftarrow \sum_{n=1}^N w^n P^n$ 
13:  end for
14:  Distribute  $L_s$  to clients
15:   $F_s \leftarrow \text{Attach } L_s \text{ to } F_{(s-1)}$ 
16: end for
17: return  $F_S$ 
```

---

### 3. Proposed Method

#### 3.1. LW-FedSSL

Our proposed approach can be integrated with various SSL techniques, and we collectively refer to it as LW-FedSSL. Fig. 3 illustrates the local training process for the  $n$ -th FL client across different stages of LW-FedSSL. For an encoder  $F$  with a total of  $S$  layers, the training process can be generally divided into  $S$  stages. Each stage  $s \in [1, S]$  runs for a fixed number of FL communication rounds,  $R_s$ . The training process starts with an empty encoder  $F_0$  (i.e., with no layers), and each stage  $s$  sequentially adds a new layer  $L_s$  to  $F_{(s-1)}$ , forming  $F_s$  and increasing the encoder's depth. During stage  $s$ , only the newly added layer  $L_s$  is updated, while all prior layers  $[L_1, \dots, L_{(s-1)}]$  within  $F_s$  are kept frozen, serving only for inference. Moreover, in each communication round, clients transmit only the active layer  $L_s$  (along with the MLP heads, depending on the SSL technique used) to the server for aggregation. Likewise, the server only needs to broadcast the aggregated layer  $L_s$  back to clients. The detailed procedure of LW-FedSSL using MoCoV3 as the SSL technique is shown in Algorithms 1 and 2. Algorithm 1 describes the server-side execution, which manages the incremental training stages  $s \in [1, S]$  and the aggregation process. Each client executes Algorithm 2, performing local SSL training to update the active layer  $L_s$ .

Dividing the training process into stages that focus on a single layer (or a block of layers) at a time significantly reduces memory and computational demands, making it more suitable for resource-constrained client devices in FL environments. Additionally, it lowers both upload and download communication costs for clients, as only the active layers need to be exchanged between the server and clients. In FL, many clients may lack the resources required for conventional end-to-end

---

**Algorithm 2** LW-FedSSL (*Client-side*)

---

**Input:** local dataset  $D^n$ , encoder  $F_{(s-1)}^n$ , number of local epochs  $E$ , momentum  $\mu$ , temperature  $\tau$

**Output:** layer  $L_s^n$ , projection head  $H^n$ , prediction head  $P^n$

```
1: Client executes: Train( $n, L_s, H, P$ ):
2: Initialize:  $L_s^n \leftarrow L_s, H^n \leftarrow H, P^n \leftarrow P$ 
3:  $F_s^n \leftarrow \text{Attach trainable } L_s^n \text{ to frozen } F_{(s-1)}^n$ 
4: Target branch:  $F_{s,k}^n \leftarrow F_s^n, H_k^n \leftarrow H^n$ 
5: for epoch  $e = 1, 2, \dots, E$  do
6:   for each batch  $x \in D^n$  do
7:      $x_1 \leftarrow \text{Augment}(x)$ 
8:      $x_2 \leftarrow \text{Augment}(x)$ 
9:      $q_1 \leftarrow P^n(H^n(F_s^n(x_1)))$ 
10:     $q_2 \leftarrow P^n(H^n(F_s^n(x_2)))$ 
11:     $k_1 \leftarrow H_k^n(F_{s,k}^n(x_1))$ 
12:     $k_2 \leftarrow H_k^n(F_{s,k}^n(x_2))$ 
13:     $\mathcal{L} \leftarrow \ell_{\text{con}}(q_1, k_2, \tau) + \ell_{\text{con}}(q_2, k_1, \tau)$ 
14:     $F_s^n, H^n, P^n \leftarrow \text{Update with } \nabla \mathcal{L}$ 
15:     $F_{s,k}^n, H_k^n \leftarrow \text{Momentum update with } \mu, F_s^n, H^n$ 
16:   end for
17: end for
18:  $L_s^n \leftarrow \text{Get active layer from } F_s^n$ 
19: return  $L_s^n, H^n, P^n$ 
```

---

training, preventing them from contributing to model training despite having valuable local data. Our proposed approach can encourage a larger number of client devices with limited computational capacity to participate, leveraging their local data for model training.

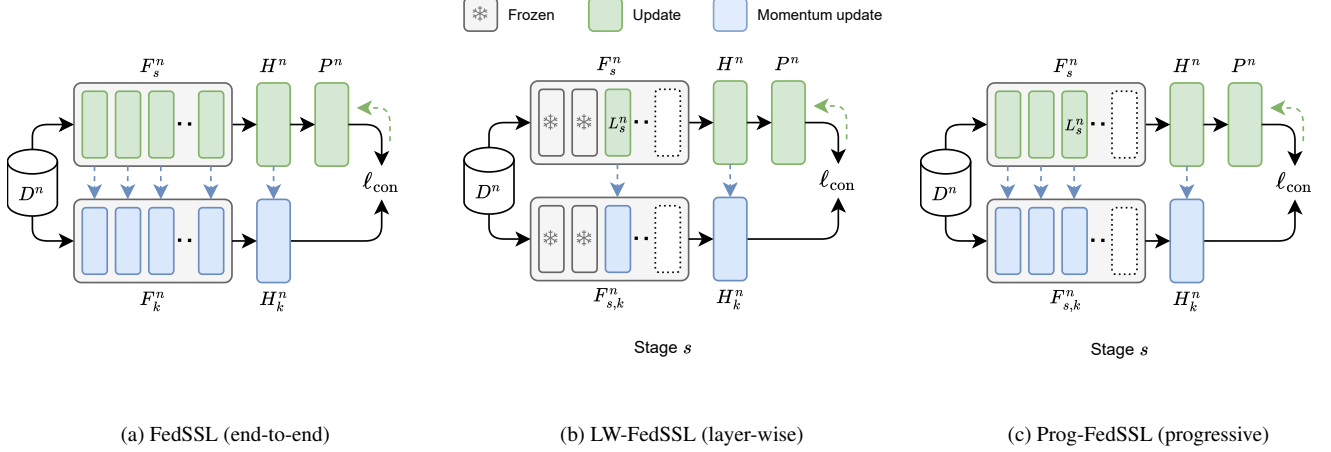
In LW-FedSSL, the number of training stages can be mainly determined by the chosen model architecture and the desired level of resource savings. It can be easily adjusted to incorporate more than a single layer within each stage, though doing so increases computational and communication demands per stage. As a result, there is an inherent trade-off between the number of layers added at each stage and the resources required for training. This flexibility allows for optimizing the training configuration based on system constraints.

#### 3.2. Prog-FedSSL

Inspired by ProgFed [24], we also explore progressive training for federated self-supervised learning, denoted as Prog-FedSSL. The local training process of Prog-FedSSL at stage  $s$  is shown in Fig. 4c, while a more comprehensive illustration across different stages is provided in Fig. A.9 of Appendix A. Similar to layer-wise training, progressive training also starts with an empty encoder  $F_0$  and sequentially adds a new layer  $L_s$  to  $F_{(s-1)}$  at the beginning of each stage  $s$ . However, in contrast to layer-wise training, which updates only a single layer  $L_s$  at a time, progressive training updates all existing layers (i.e.,  $[L_1, \dots, L_s]$ ) during each stage  $s$ . In other words, no layers within  $F_s$  are frozen at any stage. This distinction results in higher computational and communication costs for Prog-FedSSL com-

Table 1: Comparison of different layer management characteristics between FedSSL, LW-FedSSL (ours), and Prog-FedSSL (ours).

	Initialization				At Stage $s \in [1, S]$						
	Training	Number of Stages	Encoder	Layers	Encoder	Layers	New Layer	Frozen Layers	Trainable Layers	Server to Client (Download)	Client to Server (Upload)
FedSSL	end-to-end	—	$F$	$[L_1, \dots, L_S]$	$F$	$[L_1, \dots, L_S]$	—	—	$[L_1, \dots, L_S]$	$[L_1, \dots, L_S]$	$[L_1, \dots, L_S]$
LW-FedSSL	layer-wise	$S$	$F_0$	—	$F_s$	$[L_1, \dots, L_s]$	$L_s$	$[L_1, \dots, L_{(s-1)}]$	$L_s$	$L_s$	$L_s$
Prog-FedSSL	progressive	$S$	$F_0$	—	$F_s$	$[L_1, \dots, L_s]$	$L_s$	—	$[L_1, \dots, L_s]$	$[L_1, \dots, L_s]$	$[L_1, \dots, L_s]$


 Figure 4: Comparison of FedSSL, LW-FedSSL (ours), and Prog-FedSSL (ours) at stage  $s$  using MoCoV3 as the SSL backbone.

pared to LW-FedSSL. Nonetheless, when compared with end-to-end training—where clients must train a full encoder with  $S$  layers (i.e.,  $[L_1, \dots, L_S]$ ) at each round—Prog-FedSSL allows clients to train only  $s$  layers ( $s \leq S$ ). Moreover, clients only need to exchange these  $s$  layers with the server. In essence, Prog-FedSSL reduces both computational and communication costs by avoiding the training and exchanging of  $(S - s)$  layers at each stage  $s$  compared to end-to-end training. The detailed procedure of Prog-FedSSL using MoCoV3 as the SSL technique is presented in Algorithms 3 and 4 of Appendix A, with Algorithm 3 describing the server-side execution and Algorithm 4 describing the client-side execution.

Nevertheless, it is important to recognize that the resource requirements of progressive training gradually increase as the stages progress, eventually matching those of end-to-end training at the final stage (i.e., when  $s = S$ ). This growing demand could prevent resource-constrained devices from participating in training, leading to similar limitations faced by the end-to-end approach and resulting in the loss of valuable data from these devices. Meanwhile, the resource requirements of layer-wise training can be significantly lower at any stage, as it trains only a single layer  $L_s$  at a time. Fig. 4 illustrates a comparison of FedSSL (end-to-end), LW-FedSSL (layer-wise), and Prog-FedSSL (progressive) at stage  $s$ . Additionally, Table 1 provides a detailed comparison of their different layer management characteristics.

## 4. Experiment

### 4.1. Experimental setup

Unless otherwise stated, we use the following settings as the default in our experiments.

**Model.** We use a ViT-Tiny [27] backbone with 12 transformer blocks as the encoder  $F$ . The projection head  $H$  and the prediction head  $P$  are implemented as a 3-layer MLP and a 2-layer MLP, respectively. For both  $H$  and  $P$ , the hidden layer dimensions are set to 512, and the output dimension is set to 256.

**Data.** We use images from the COCO [28] dataset to construct the client datasets. The COCO dataset contains diverse natural scene images, which inherently introduce data heterogeneity. These images are uniformly distributed across 10 FL clients. To evaluate the performance, we use downstream datasets including CIFAR-10/100 [29], Tiny ImageNet [30], and Caltech-101 [31]. For all datasets, we use an image size of  $32 \times 32$  with a patch size of 4.

**Training.** The number of training stages is set to 12 (i.e.,  $S = 12$ ). The total number of FL communication rounds  $R$  is set to 180, which is evenly distributed across the stages  $s \in [1, S]$ . This results in 15 rounds per stage (i.e.,  $R_s = R/S = 15$ ). Each FL client performs local training for 3 epochs using the AdamW optimizer, with a base learning rate of  $1.5e-4$  and a weight decay of  $1e-5$ . The batch size is set to 512, and the learning rate is linearly scaled as  $base\_learning\_rate \times batch\_size/256$  [32, 33].

Table 2: Comparison of resource requirements and performance. For resource requirements, we compare the maximum memory usage (GB), total GFLOPs, and communication costs (download + upload) for a client. Except for the centralized settings, best values within each group are marked in **bold**.

	Resource Requirements			Accuracy (%)				
	Memory (GB)	GFLOPs	Comm. (GB)	CIFAR-10	CIFAR-100	Tiny ImageNet	Caltech-101	Avg
Scratch	—	—	—	63.28	39.82	26.09	22.10	37.82
MoCoV3 [11]								
MoCoV3 [11] (Centralized)	—	—	—	87.22	65.97	43.08	60.61	64.22
FedMoCoV3	8.72 (1.00×)	586 (1.00×)	8.40 (1.00×)	83.95	62.84	<b>41.77</b>	54.60	60.79
LW-FedSSL (ours)	<b>2.61 (0.30×)</b>	<b>139 (0.24×)</b>	<b>1.66 (0.20×)</b>	83.43	61.65	40.00	48.33	58.35
Prog-FedSSL (ours)	8.69 (1.00×)	318 (0.54×)	5.04 (0.60×)	<b>86.53</b>	<b>65.10</b>	40.70	<b>61.60</b>	<b>63.48</b>
BYOL [16]								
BYOL [16] (Centralized)	—	—	—	87.69	66.26	44.84	62.32	65.28
FedBYOL	8.72 (1.00×)	586 (1.00×)	8.40 (1.00×)	83.59	62.71	42.67	53.89	60.72
FedU [6]	8.94 (1.03×)	586 (1.00×)	8.40 (1.00×)	83.50	62.61	42.61	53.88	60.65
FedEMA [13]	8.96 (1.03×)	586 (1.00×)	8.40 (1.00×)	81.35	60.50	41.20	48.26	57.83
LW-FedSSL (ours)	<b>2.61 (0.30×)</b>	<b>139 (0.24×)</b>	<b>1.66 (0.20×)</b>	84.20	61.29	40.56	47.77	58.45
Prog-FedSSL (ours)	8.69 (1.00×)	318 (0.54×)	5.04 (0.60×)	<b>87.51</b>	<b>66.32</b>	<b>43.16</b>	<b>62.19</b>	<b>64.80</b>
SimCLR [1]								
SimCLR [1] (Centralized)	—	—	—	87.49	65.64	43.54	61.12	64.45
FedSimCLR	14.39 (1.00×)	1169 (1.00×)	8.03 (1.00×)	84.38	62.41	42.49	53.45	60.68
FedCA [21]	16.37 (1.14×)	1299 (1.11×)	30.69 (3.82×)	85.30	63.68	<b>43.04</b>	57.75	62.44
LW-FedSSL (ours)	<b>8.59 (0.60×)</b>	<b>278 (0.24×)</b>	<b>1.29 (0.16×)</b>	84.27	61.74	40.70	48.25	58.74
Prog-FedSSL (ours)	14.36 (1.00×)	635 (0.54×)	4.68 (0.58×)	<b>86.38</b>	<b>63.77</b>	41.60	<b>62.08</b>	<b>63.46</b>

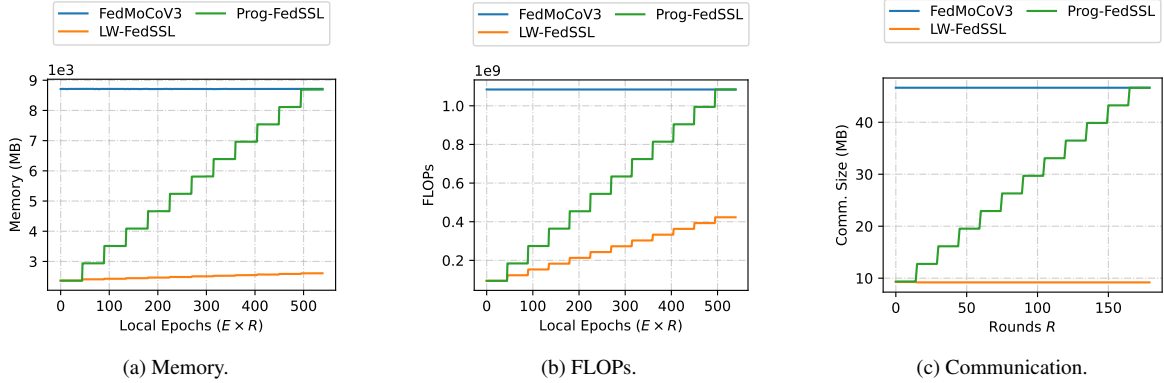


Figure 5: Computational and communication resources required for a client (a) Memory usage. (b) FLOPs consumption. (c) Communication cost.

We set the momentum  $\mu$  to 0.99 and the temperature  $\tau$  to 0.05. Data augmentations for view creation include random resized crop, color jitter, grayscale, horizontal flip, Gaussian blur, and solarization.

**Evaluation.** After training, the encoder  $F$  is retained, while the projection head  $H$  and prediction head  $P$  are discarded. A linear classifier head is added to the encoder  $F$  and fine-tuned on the downstream datasets for 40 epochs, including a warmup period of 10 epochs. We set the batch size to 256 and use the AdamW optimizer with a base learning rate of  $1e-3$  and a weight decay of  $1e-5$ . A cosine decay schedule is applied to the learning rate during training. We use the RandAugment function available in the PyTorch [34] framework for augmentation.

#### 4.2. Performance and Resource Comparison

Table 2 presents a comparison of the resource requirements and performance across different training approaches.<sup>2</sup> For resource requirements, we compare the computational and communication resources needed for a single FL client. Additionally, we demonstrate that our proposed methods, LW-FedSSL and Prog-FedSSL, can be seamlessly integrated with various SSL techniques. As shown in Table 2, we integrate MoCoV3 [11], BYOL [16], and SimCLR [1] with our methods. For clarity, we group similar methods based on the SSL technique used. Fig. 5 plots the resource consumption for a client, illustrating computational resources (i.e., memory and FLOPs) across local epochs and communication cost (download + up-

<sup>2</sup>We report mean per-class accuracy for the Caltech-101 dataset.

Table 3: Resource requirements and performance comparison for different numbers of layers added per stage in LW-FedSSL and Prog-FedSSL.

	Stage Settings			Resource Requirements			Accuracy (%)				
	Layers Added per Stage	Stages $S$	Rounds per Stage $R_s$	Memory (GB)	GFLOPs	Comm. (GB)	CIFAR-10	CIFAR-100	Tiny ImageNet	Caltech-101	Avg
FedMoCoV3	—	—	—	8.72 (1.00×)	586 (1.00×)	8.40 (1.00×)	83.95	62.84	41.77	54.60	60.79
LW-FedSSL (ours)	1	12	15	<b>2.61 (0.30×)</b>	<b>139 (0.24×)</b>	<b>1.66 (0.20×)</b>	83.43	61.65	40.00	48.33	58.35
	2	6	30	3.12 (0.36×)	180 (0.31×)	2.27 (0.27×)	84.85	62.23	40.11	56.19	60.84
	3	4	45	3.67 (0.42×)	220 (0.38×)	2.88 (0.34×)	85.32	63.07	40.28	58.48	61.79
	4	3	60	4.23 (0.48×)	261 (0.45×)	3.49 (0.42×)	85.54	64.13	40.71	58.64	62.26
	6	2	90	5.34 (0.61×)	342 (0.58×)	4.71 (0.56×)	86.58	66.18	42.79	62.72	64.57
Prog-FedSSL (ours)	1	12	15	8.69 (1.00×)	318 (0.54×)	5.04 (0.60×)	86.53	65.10	40.70	61.60	63.48
	2	6	30	8.69 (1.00×)	342 (0.58×)	5.35 (0.64×)	86.57	65.39	42.00	62.84	64.20
	3	4	45	8.69 (1.00×)	367 (0.63×)	5.65 (0.67×)	86.61	65.73	42.13	63.52	64.50
	4	3	60	8.69 (1.00×)	391 (0.67×)	5.96 (0.71×)	87.33	65.80	43.12	62.56	64.70
	6	2	90	8.69 (1.00×)	440 (0.75×)	6.57 (0.78×)	<b>88.08</b>	<b>66.78</b>	<b>43.80</b>	<b>64.21</b>	<b>65.72</b>

Table 4: Comparison of resource requirements and performance with ResNet-18 as the encoder.

	Resource Requirements			Accuracy (%)				
	Memory (GB)	GFLOPs	Comm. (GB)	CIFAR-10	CIFAR-100	Tiny ImageNet	Caltech-101	Avg
FedMoCoV3	1.83 (1.00×)	41 (1.00×)	11.09 (1.00×)	85.20	57.57	33.71	61.58	59.52
LW-FedSSL (ours)	<b>1.59 (0.87×)</b>	<b>17 (0.40×)</b>	<b>3.41 (0.31×)</b>	<b>85.64</b>	<b>57.84</b>	34.40	61.84	59.93
Prog-FedSSL (ours)	1.81 (0.99×)	27 (0.67×)	4.24 (0.38×)	85.57	57.44	<b>34.63</b>	<b>63.59</b>	<b>60.31</b>

Table 5: Communication cost for the worst-case scenario of LW-FedSSL.

	Comm. (GB)
FedMoCoV3	8.40 (1.00×)
LW-FedSSL (ours)	<b>3.35 (0.40×)</b>
Prog-FedSSL (ours)	5.04 (0.60×)

load) across FL rounds. For FLOPs calculation, we consider only a single input sample.

The results in Table 2 show that LW-FedSSL significantly reduces resource requirements. Specifically, when using MoCoV3, LW-FedSSL achieves a 3.34× reduction in memory usage, a 4.20× fewer computational operations (GFLOPs), and a 5.07× reduction in communication costs compared to conventional FedMoCoV3. Additionally, Table 2 demonstrates that LW-FedSSL also maintains comparable performance to its end-to-end training counterparts. While Prog-FedSSL is not as resource-efficient as LW-FedSSL, it still consumes fewer GFLOPs and communication resources compared to FedMoCoV3. Furthermore, Prog-FedSSL achieves better performance in most cases, suggesting that progressive training can also benefit self-supervised learning, aligning with the findings for supervised learning in [24]. Fig. 5 illustrates that the resource consumption for LW-FedSSL remains relatively flat across local epochs and FL rounds due to the focus on training a single layer at each stage. Meanwhile, the resource consumption for Prog-FedSSL gradually increases as the number of trainable layers grows with each stage, eventually matching that of conventional FedMoCoV3.

#### 4.3. Worst-case Communication Scenario for LW-FedSSL

By default, we assume that clients only need to download the layer  $L_s$  at each FL round for LW-FedSSL. This assumption

holds only if all clients participate in every FL round without dropping out or no new clients join. Otherwise, new clients or those that previously dropped out would need to download the latest model  $F_s$  (i.e.,  $[L_1, \dots, L_s]$ ) instead of just  $L_s$ . Therefore, we compare the worst-case communication cost for LW-FedSSL in Table 5, where a client downloads  $F_s$  at every FL round. Note that client still needs to upload only  $L_s$  after local training. The results show that even in the worst-case scenario, LW-FedSSL significantly reduces communication costs compared to end-to-end FedMoCoV3.

#### 4.4. Number of Layers Added per Stage

Both LW-FedSSL and Prog-FedSSL allow each stage to be adjusted to incorporate multiple new layers instead of just a single layer. However, increasing the number of layers per stage also raises resource requirements, including memory, computational cost, and communication overhead. The number of layers added per stage presents a trade-off between training efficiency and resource constraints, which can be adjusted based on the resource capacity of the targeted FL clients. Table 3 compares different settings for LW-FedSSL and Prog-FedSSL, where varying numbers of new layers are added at each stage. In all cases, the total number of training rounds  $R$  is kept constant at 180. As more layers are added per stage, the number of stages decreases, and resource requirements increase, but this is accompanied by an improvement in model performance.

#### 4.5. ResNet-18

To demonstrate that our proposed approaches can also work well with different model architectures, we use ResNet-18 [37] as the encoder  $F$  in Table 4. We set the number of stages to  $S = 4$ , with each stage sequentially adding a new ResNet block to the encoder. The total number of FL rounds is set to  $R = 120$ ,



Table 6: Resource requirements and performance comparison when integrating Depth Dropout (DD), Proximal Alignment (PA), and Representation Alignment (RA) into LW-FedSSL and Prog-FedSSL.

	Resource Requirements			Accuracy (%)				
	Memory (GB)	GFLOPs	Comm. (GB)	CIFAR-10	CIFAR-100	Tiny ImageNet	Caltech-101	Avg
LW-FedSSL	2.61 (1.00×)	139 (1.00×)	<b>1.66 (1.00×)</b>	83.43	61.65	40.00	48.33	58.35
LW-FedSSL + DD [25]	<b>2.49 (0.95×)</b>	<b>111 (0.80×)</b>	<b>1.66 (1.00×)</b>	84.10	60.78	40.18	<b>48.90</b>	<b>58.49</b>
LW-FedSSL + PA [35]	2.63 (1.01×)	139 (1.00×)	<b>1.66 (1.00×)</b>	<b>84.46</b>	<b>61.74</b>	<b>40.72</b>	44.31	57.81
LW-FedSSL + RA [36]	2.86 (1.10×)	318 (2.29×)	<b>1.66 (1.00×)</b>	84.31	61.69	40.18	45.29	57.87
Prog-FedSSL	8.69 (1.00×)	318 (1.00×)	5.04 (1.00×)	86.53	65.10	40.70	61.60	63.48
Prog-FedSSL + DD [25]	<b>5.24 (0.60×)</b>	<b>233 (0.73×)</b>	<b>3.97 (0.79×)</b>	86.69	64.41	41.59	<b>63.46</b>	64.04
Prog-FedSSL + PA [35]	8.72 (1.00×)	318 (1.00×)	5.04 (1.00×)	84.71	62.09	38.65	56.35	60.45
Prog-FedSSL + RA [36]	9.00 (1.04×)	496 (1.56×)	5.04 (1.00×)	<b>87.05</b>	<b>66.05</b>	<b>42.37</b>	62.59	<b>64.52</b>

Table 7: Performance comparison under partial client participation, where a subset of clients is selected at each FL round. The setting considers 25 or 50 out of 100 clients participating in each round.

	Accuracy (%)				
	CIFAR-10	CIFAR-100	Tiny ImageNet	Caltech-101	Avg
Number of Participating Clients: 25/100					
FedMoCoV3	71.85	49.19	32.98	33.95	46.99
LW-FedSSL (ours)	84.33	61.96	39.46	43.80	57.39
Prog-FedSSL (ours)	<b>85.18</b>	<b>62.80</b>	<b>40.05</b>	<b>55.88</b>	<b>60.98</b>
Number of Participating Clients: 50/100					
FedMoCoV3	72.73	49.62	33.58	34.67	47.65
LW-FedSSL (ours)	84.23	62.17	41.17	44.28	57.96
Prog-FedSSL (ours)	<b>84.46</b>	<b>62.87</b>	<b>40.19</b>	<b>56.96</b>	<b>61.12</b>

with each stage  $s$  allocated  $R_s = 30$  rounds. Whenever a new ResNet block is added and the output dimension of the encoder changes, we add a temporary linear layer to reshape the output to 512. This reshaping layer is discarded at the end of each stage. The results in Table 4 demonstrate that our approaches also perform well with ResNet-18.

#### 4.6. Integration of Additional Mechanisms

In both LW-FedSSL and Prog-FedSSL, each stage can be viewed as an independent FL process. At the beginning of each stage, some layers retain weights from the previous stage, but otherwise, the FL process within each stage remains independent. This modularity allows easy integration of existing FL mechanisms into LW-FedSSL and Prog-FedSSL. To demonstrate this flexibility, we incorporate depth dropout (DD) [25], a technique that selectively drops frozen layers during training; alignment using the proximal term (PA) [35], which constrains local updates to prevent drastic weight divergence across clients; and a similar alignment mechanism using representation (RA) [36]. Table 6 presents the evaluation results, including resource requirements and downstream performance across multiple datasets.

#### 4.7. Partial Client Participation

In FL environments, it is common for only a subset of clients to participate in the training at each communication round. Clients may drop out of the training due to various constraints, such as unstable network connections or power limitations. In

Table 8: Performance comparison under partial client participation, where each client is assigned a dropout probability of 25% or 50%. The total number of clients is set to 100.

	Accuracy (%)				
	CIFAR-10	CIFAR-100	Tiny ImageNet	Caltech-101	Avg
Client Dropout Probability: 25 %					
FedMoCoV3	72.38	49.49	33.48	34.38	47.43
LW-FedSSL (ours)	84.33	62.12	40.32	43.19	57.49
Prog-FedSSL (ours)	<b>85.11</b>	<b>62.73</b>	<b>40.34</b>	<b>54.69</b>	<b>60.72</b>
Client Dropout Probability: 50 %					
FedMoCoV3	73.08	49.67	33.19	35.14	47.77
LW-FedSSL (ours)	83.93	61.66	40.06	44.77	57.61
Prog-FedSSL (ours)	<b>84.78</b>	<b>63.08</b>	<b>40.67</b>	<b>54.02</b>	<b>60.64</b>

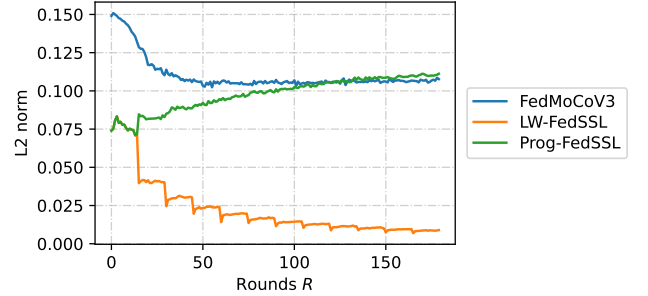


Figure 6: Distance between local model weights. At the end of each FL round, we randomly select a client and measure the L2 norm between its local model weights and those of all other clients, plotting the average.

Table 7, we set the total number of clients to 100 and randomly select 25 or 50 clients to participate in each FL round. Furthermore, in Table 8, we use another setting where we assign a dropout probability to each client, using dropout rates of 25% or 50%.

Interestingly, we observe that LW-FedSSL consistently outperforms the end-to-end FedMoCoV3 under both settings. To investigate this, in Fig. 6, we measure the L2 norm distance between local model weights at the end of each FL round. We use the setting where 25 clients are selected per round. To measure the distance, we randomly select a client and compute the average L2 norm between its local model weights and those of all other clients. Fig. 6 indicates that gradually adding new trainable layers, stage by stage, helps limit the divergence between



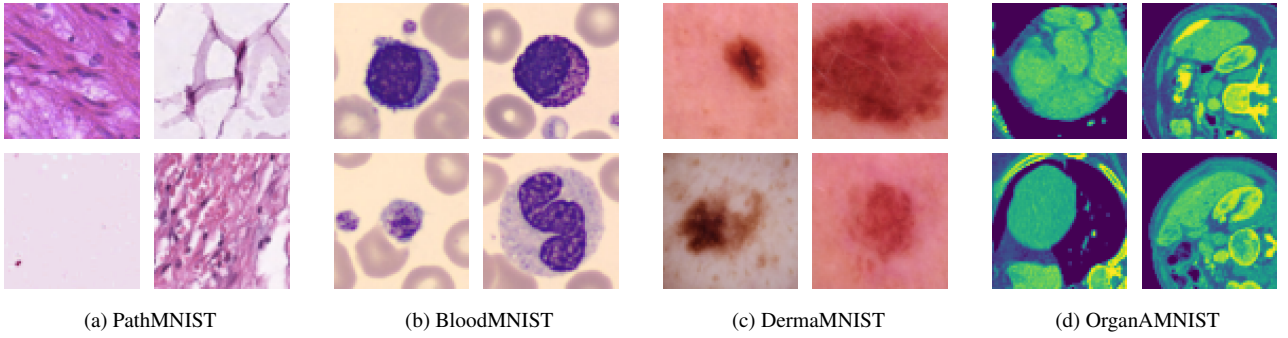


Figure 7: Sample images for (a) PathMNIST, (b) BloodMNIST, (c) DermaMNIST, and (d) OrganAMNIST from the MedMNIST collection. Each dataset represents different biomedical imaging modalities used for classification tasks.

Table 9: Performance evaluation across medical datasets, including PathMNIST, BloodMNIST, DermaMNIST, and OrganAMNIST, from the MedMNIST collection.

	Precision	Recall	F1	Accuracy
PathMNIST				
FedMoCoV3	86.53	86.13	85.04	85.92
LW-FedSSL (ours)	88.13	87.92	87.22	87.24
Prog-FedSSL (ours)	<b>90.59</b>	<b>90.25</b>	<b>89.80</b>	<b>90.03</b>
BloodMNIST				
FedMoCoV3	91.18	90.84	90.68	90.82
LW-FedSSL (ours)	92.90	93.15	92.76	93.20
Prog-FedSSL (ours)	<b>94.12</b>	<b>93.73</b>	<b>93.77</b>	<b>93.62</b>
DermaMNIST				
FedMoCoV3	33.56	27.24	27.85	26.97
LW-FedSSL (ours)	<b>40.03</b>	<b>30.29</b>	<b>30.78</b>	<b>30.31</b>
Prog-FedSSL (ours)	27.39	24.63	24.24	25.42
OrganAMNIST				
FedMoCoV3	77.28	74.40	74.31	74.19
LW-FedSSL (ours)	77.05	74.27	74.10	74.32
Prog-FedSSL (ours)	<b>78.44</b>	<b>76.61</b>	<b>76.03</b>	<b>76.59</b>

Table 10: Resource requirements for the BloodMNIST dataset.

	Resource Requirements (BloodMNIST)		
	Memory (GB)	GFLOPs	Comm. (GB)
FedMoCoV3	11.48 (1.00×)	2705 (1.00×)	8.45 (1.00×)
LW-FedSSL (ours)	<b>3.10 (0.27×)</b>	<b>640 (0.24×)</b>	<b>1.66 (0.20×)</b>
Prog-FedSSL (ours)	11.46 (1.00×)	1468 (0.54×)	5.09 (0.60×)

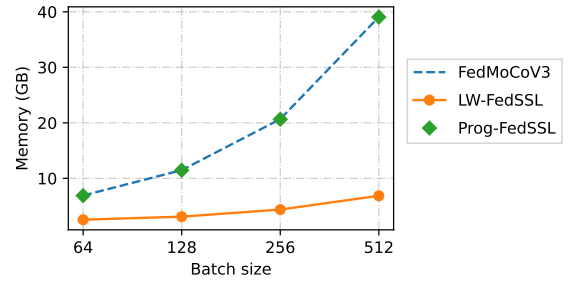


Figure 8: Peak memory usage across different batch sizes on the BloodMNIST dataset.

local models. We believe this can have a positive impact on the aggregation process, especially with a large number of clients, leading to improved performance.

#### 4.8. Medical Data

We evaluate our proposed approaches on MedMNIST [38], a collection of diverse biomedical image datasets. Specifically, we use four subsets within MedMNIST: PathMNIST, BloodMNIST, DermaMNIST, and OrganAMNIST. Sample images from each dataset are shown in Fig. 7. Each dataset contains three splits: train, validation, and test. We use the train split to construct the local datasets for FL clients, the validation split as the downstream dataset for fine-tuning, and the test split for final evaluation. To introduce data heterogeneity, we partition the data among clients using the Dirichlet distribution [39] with a concentration parameter of  $\beta = 5$ . We use an image size of  $64 \times 64$  and a batch size of 128 for all experiments. In Table 10, we present resource requirements only for BloodMNIST, as the experimental settings—such as batch size, input dimensions,

and model architecture—are consistent across all datasets, resulting in similar resource requirements. In Table 9, we evaluate the performance of our proposed methods on the medical datasets. For all datasets, we report mean per-class accuracy.

#### 4.9. Batch Size

We evaluate the impact of batch size on memory consumption using the BloodMNIST dataset. We consider batch sizes of 64, 128, 256, and 512 while all other settings remain the same as described in Section 4.8. Fig. 8 shows the peak memory usage across different batch sizes, while Table 12 presents the evaluation results. We only report memory usage in Fig. 8, as the FLOPs and communication costs across different batch sizes remain similar to those presented in Table 10.<sup>3</sup> Given that Prog-FedSSL does not involve frozen layers, its peak memory requirement—which occurs at the final stage  $S$ —is equiva-

<sup>3</sup>This is because we consider only a single input sample for FLOPs calculations, and communication cost is independent of the training batch size.

Table 11: Evaluation on the BloodMNIST dataset across different input dimensions and patch size settings.

	Memory (GB)	GFLOPs	Comm. (GB)	Precision	Recall	F1	Accuracy
Image Size: $64 \times 64$ , Patch Size: $4 \times 4$ , Number of Patches: 256							
FedMoCoV3	11.48 (1.00 $\times$ )	2705 (1.00 $\times$ )	8.45 (1.00 $\times$ )	91.18	90.84	90.68	90.82
LW-FedSSL (ours)	<b>3.10 (0.27<math>\times</math>)</b>	<b>640 (0.24<math>\times</math>)</b>	<b>1.66 (0.20<math>\times</math>)</b>	92.90	93.15	92.76	93.20
Prog-FedSSL (ours)	11.46 (1.00 $\times$ )	1468 (0.54 $\times$ )	5.09 (0.60 $\times$ )	<b>94.12</b>	<b>93.73</b>	<b>93.77</b>	<b>93.62</b>
Image Size: $128 \times 128$ , Patch Size: $8 \times 8$ , Number of Patches: 256							
FedMoCoV3	11.54 (1.00 $\times$ )	2717 (1.00 $\times$ )	8.48 (1.00 $\times$ )	89.72	89.77	89.53	89.61
LW-FedSSL (ours)	<b>3.15 (0.27<math>\times</math>)</b>	<b>645 (0.24<math>\times</math>)</b>	<b>1.66 (0.20<math>\times</math>)</b>	92.16	92.23	92.03	92.03
Prog-FedSSL (ours)	11.52 (1.00 $\times$ )	1479 (0.54 $\times$ )	5.13 (0.60 $\times$ )	<b>94.41</b>	<b>93.65</b>	<b>93.87</b>	<b>93.61</b>
Image Size: $224 \times 224$ , Patch Size: $16 \times 16$ , Number of Patches: 196							
FedMoCoV3	8.88 (1.00 $\times$ )	2028 (1.00 $\times$ )	8.62 (1.00 $\times$ )	90.20	90.21	89.94	90.05
LW-FedSSL (ours)	<b>2.84 (3.12<math>\times</math>)</b>	<b>487 (0.24<math>\times</math>)</b>	<b>1.67 (0.19<math>\times</math>)</b>	93.16	93.01	92.88	92.80
Prog-FedSSL (ours)	8.87 (1.00 $\times$ )	2028 (0.55 $\times$ )	5.26 (0.61 $\times$ )	<b>95.23</b>	<b>95.26</b>	<b>95.13</b>	<b>95.14</b>

Table 12: Evaluation on the BloodMNIST dataset across different batch sizes.

	Precision	Recall	F1	Accuracy
Batch Size: 64				
FedMoCoV3	92.28	91.56	91.71	91.29
LW-FedSSL (ours)	92.66	92.85	92.61	92.61
Prog-FedSSL (ours)	<b>93.80</b>	<b>94.25</b>	<b>93.87</b>	<b>94.23</b>
Batch Size: 128				
FedMoCoV3	91.18	90.84	90.68	90.82
LW-FedSSL (ours)	92.90	93.15	92.76	93.20
Prog-FedSSL (ours)	<b>94.12</b>	<b>93.73</b>	<b>93.77</b>	<b>93.62</b>
Batch Size: 256				
FedMoCoV3	91.48	91.78	91.35	91.65
LW-FedSSL (ours)	92.33	92.74	92.31	92.58
Prog-FedSSL (ours)	<b>94.67</b>	<b>94.72</b>	<b>94.54</b>	<b>94.68</b>
Batch Size: 512				
FedMoCoV3	90.10	90.48	90.02	90.13
LW-FedSSL (ours)	91.95	92.25	91.86	92.01
Prog-FedSSL (ours)	<b>93.80</b>	<b>93.91</b>	<b>93.65</b>	<b>93.89</b>

lent to that of end-to-end FedMoCoV3. Peak memory requirements for both FedMoCoV3 and Prog-FedSSL sharply rise as the batch size grows, while those of LW-FedSSL remain relatively flat. Many SSL approaches that rely on contrastive loss require large batch sizes to leverage a greater number of negative samples [1]. The low-memory footprint of LW-FedSSL can accommodate large-batch training, making it more scalable in resource-constrained environments.

#### 4.10. Impact of Input Dimensions

In Table 11, we evaluate our proposed approaches using different input image sizes and corresponding patch sizes on the BloodMNIST dataset. The choice of input size and patch size can significantly impact computational resource requirements. Larger input dimensions demand increased memory and FLOPs for processing. Meanwhile, larger patch sizes reduce the number of patches but may affect feature granularity, potentially influencing model performance. We compare three different input configurations: (1) a  $64 \times 64$  image with  $4 \times 4$  patches, (2) a  $128 \times 128$  image with  $8 \times 8$  patches, and (3) a

Table 13: Evaluation on the BloodMNIST dataset across different levels of data heterogeneity.

	Precision	Recall	F1	Accuracy
$\beta = 50$				
FedMoCoV3	91.05	91.09	90.85	91.09
LW-FedSSL (ours)	92.60	92.92	92.56	92.77
Prog-FedSSL (ours)	<b>94.33</b>	<b>94.46</b>	<b>94.27</b>	<b>94.49</b>
$\beta = 5$				
FedMoCoV3	91.18	90.84	90.68	90.82
LW-FedSSL (ours)	92.90	93.15	92.76	93.20
Prog-FedSSL (ours)	<b>94.12</b>	<b>93.73</b>	<b>93.77</b>	<b>93.62</b>
$\beta = 0.5$				
FedMoCoV3	92.43	92.26	92.10	92.14
LW-FedSSL (ours)	92.55	93.15	92.68	92.93
Prog-FedSSL (ours)	<b>93.34</b>	<b>93.49</b>	<b>93.24</b>	<b>93.59</b>

$224 \times 224$  image with  $16 \times 16$  patches. All other experimental settings are kept as described in Section 4.8. The results indicate that LW-FedSSL consistently maintains its resource efficiency across different input dimensions while achieving strong performance. Meanwhile, Prog-FedSSL achieves the best overall performance across all input sizes.

#### 4.11. Data Heterogeneity

Data heterogeneity is one of the main challenges in an FL environment. The concentration parameter  $\beta$  in the Dirichlet distribution can be used to determine the strength of data heterogeneity, with a lower  $\beta$  value indicating a higher degree of heterogeneity. In Table 13, we conduct experiments using different levels of data heterogeneity by setting  $\beta$  values to 50, 5, and 0.5.<sup>4</sup> All other experimental settings are kept the same as in Section 4.8. Table 13 shows that all approaches are relatively robust to different  $\beta$  values. This observation aligns with prior studies [8] that highlight the robustness of SSL-based approaches to data heterogeneity in FL environments. This also indicates that LW-FedSSL and Prog-FedSSL exhibit robustness to data heterogeneity inherent in end-to-end SSL training.

<sup>4</sup>The resulting client data distributions are visualized in Fig. A.10.

## 5. Conclusion

Edge devices in distributed environments often struggle to meet the computational and communication demands of federated self-supervised learning. In response to these challenges, we propose a layer-wise training approach named LW-FedSSL, which allows FL clients to perform self-supervised learning with significantly improved resource efficiency. By dividing the training process into multiple stages and focusing on a subset of layers at each stage, LW-FedSSL effectively reduces resource consumption. When applied to MoCoV3, LW-FedSSL requires  $3.34\times$  less memory, consumes  $4.20\times$  fewer GFLOPs, and reduces total transmission costs (download + upload) by  $5.07\times$ , while maintaining comparable performance to its end-to-end counterpart, FedMoCoV3. Furthermore, we explore a progressive training approach named Prog-FedSSL. Although Prog-FedSSL is less resource-efficient than LW-FedSSL, it remains more efficient than end-to-end training and achieves superior performance in most cases. Through extensive experiments across various FL settings, datasets, and ablations, we demonstrate the effectiveness of LW-FedSSL and Prog-FedSSL. The results highlight the potential of layer-wise and progressive training strategies for enhancing the scalability and practicality of self-supervised learning in resource-constrained FL environments.

## Appendix A. Additional Details

### Appendix A.1. Details on FLOPs Calculation

Here, we discuss the details on FLOPs calculation used in our experiments. In layer-wise training, calculating FLOPs for inactive (frozen) layers involves simply considering FLOPs associated with the forward pass. However, for active layers, the calculation must account for both the forward and backward passes. While numerous works have addressed FLOPs calculation for the forward pass (inference), limited studies are available on the practice of FLOPs calculation for the backward pass. Notably, existing related works [24, 25] also do not provide specific details on how FLOPs are computed. Some studies [40, 41, 42, 43] suggest that the number of operations in a backward pass of a neural network is often twice that of a forward pass. Following these studies, we adopt a backward-forward ratio of 2:1 to calculate the total FLOPs for active layers. We use the `FlopCountAnalysis` tool in the `fvcore` [44] library for calculating the FLOPs of the forward pass. We only consider a single input sample for FLOPs calculation.

### Appendix A.2. Data Distribution

Fig. A.10 illustrates the data distribution among clients for different  $\beta$  values used in Section 4.11. Lower  $\beta$  values correspond to increasingly heterogeneous data distributions among clients.

---

### Algorithm 3 Prog-FedSSL (Server-side)

---

**Input:** encoder  $F_0$ , projection head  $H$ , prediction head  $P$ , number of stages  $S$ , number of rounds per stage  $R_s$  where  $s \in [1, S]$

**Output:** encoder  $F_S$

```

1: Server executes:
2: Distribute  $F_0$  to clients
3: for stage  $s = 1, 2, \dots, S$  do
4:   Initialize new layer:  $L_s$ 
5:    $F_s \leftarrow$  Attach  $L_s$  to  $F_{(s-1)}$ 
6:   for round  $r = 1, 2, \dots, R_s$  do
7:     for client  $n = 1, 2, \dots, N$  in parallel do
8:        $F_s^n, H^n, P^n \leftarrow$  Train( $n, F_s, H, P$ )
9:        $w^n = \frac{|D^n|}{|D|}$ , where  $D = \bigcup_{n=1}^N D^n$ 
10:    end for
11:     $F_s \leftarrow \sum_{n=1}^N w^n F_s^n$ 
12:     $H \leftarrow \sum_{n=1}^N w^n H^n$ 
13:     $P \leftarrow \sum_{n=1}^N w^n P^n$ 
14:  end for
15: end for
16: return  $F_S$ 

```

---



---

### Algorithm 4 Prog-FedSSL (Client-side)

---

**Input:** local dataset  $D^n$ , number of local epochs  $E$ , momentum  $\mu$ , temperature  $\tau$

**Output:** encoder  $F_s^n$ , projection head  $H^n$ , prediction head  $P^n$

```

1: Client executes: Train( $n, F_s, H, P$ ):
2: Initialize:  $F_s^n \leftarrow F_s, H^n \leftarrow H, P^n \leftarrow P$ 
3: Target branch:  $F_{s,k}^n \leftarrow F_s^n, H_k^n \leftarrow H^n$ 
4: for epoch  $e = 1, 2, \dots, E$  do
5:   for each batch  $x \in D^n$  do
6:      $x_1 \leftarrow$  Augment( $x$ )
7:      $x_2 \leftarrow$  Augment( $x$ )
8:      $q_1 \leftarrow P^n(H^n(F_s^n(x_1)))$ 
9:      $q_2 \leftarrow P^n(H^n(F_s^n(x_2)))$ 
10:     $k_1 \leftarrow H_k^n(F_{s,k}^n(x_1))$ 
11:     $k_2 \leftarrow H_k^n(F_{s,k}^n(x_2))$ 
12:     $\mathcal{L} \leftarrow \ell_{\text{con}}(q_1, k_2, \tau) + \ell_{\text{con}}(q_2, k_1, \tau)$ 
13:     $F_s^n, H^n, P^n \leftarrow$  Update with  $\nabla \mathcal{L}$ 
14:     $F_{s,k}^n, H_k^n \leftarrow$  Momentum update with  $\mu, F_s^n, H^n$ 
15:  end for
16: end for
17: return  $F_s^n, H^n, P^n$ 

```

---

## References

- [1] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A simple framework for contrastive learning of visual representations, in: International conference on machine learning, PMLR, 2020, pp. 1597–1607.
- [2] K. He, H. Fan, Y. Wu, S. Xie, R. Girshick, Momentum contrast for unsupervised visual representation learning, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 9729–9738.

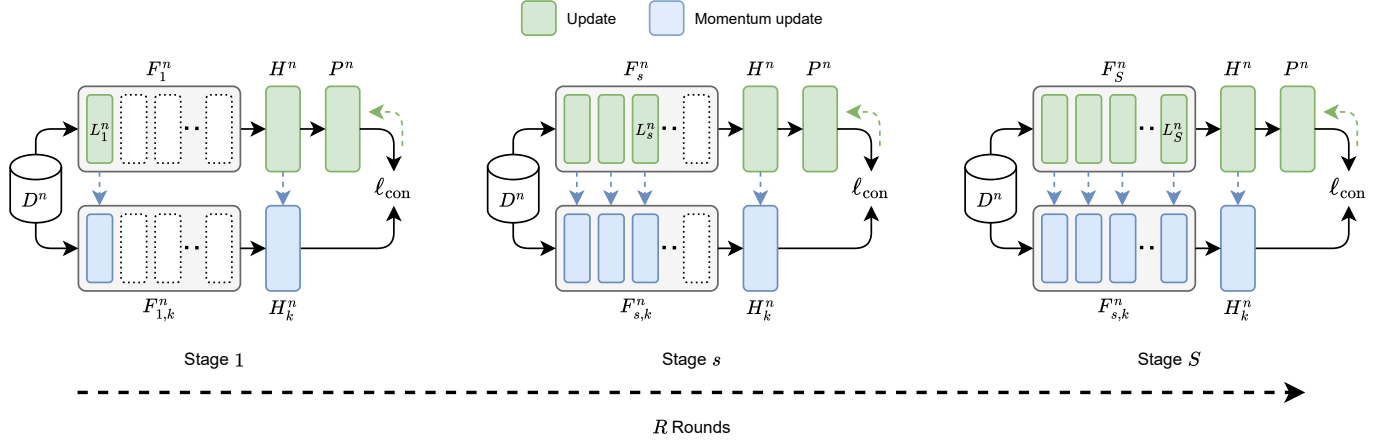


Figure A.9: Prog-FedSSL: Local training process across different stages for the  $n$ -th client. At the beginning of each stage  $s \in [1, S]$ , a new layer  $L_s$  is sequentially added to the encoder  $F_{(s-1)}$ , increasing its depth. During stage  $s$ , Prog-FedSSL updates all existing layers (i.e.,  $[L_1, \dots, L_s]$ ) within  $F_s$ .

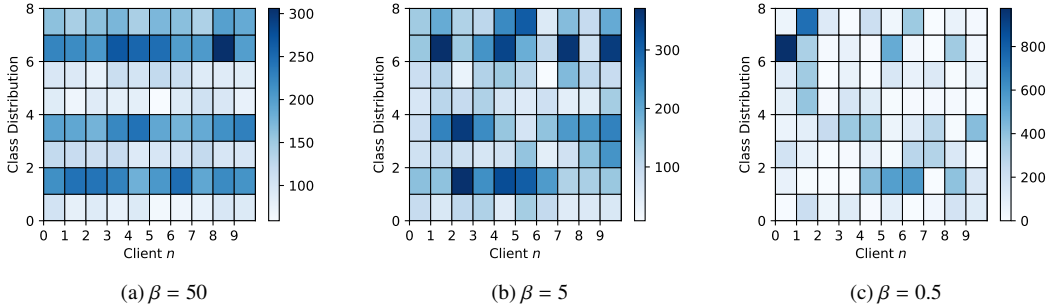


Figure A.10: Client data distribution with different  $\beta$  values for the BloodMNIST dataset. A darker color denotes a higher number of data samples for a specific class in a client.

- [3] X. Chen, K. He, Exploring simple siamese representation learning, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 15750–15758.
- [4] J. Konečný, H. B. McMahan, D. Ramage, P. Richtárik, Federated optimization: Distributed machine learning for on-device intelligence, arXiv preprint arXiv:1610.02527 (2016).
- [5] B. McMahan, E. Moore, D. Ramage, S. Hampson, B. A. y Arcas, Communication-efficient learning of deep networks from decentralized data, in: Artificial intelligence and statistics, PMLR, 2017, pp. 1273–1282.
- [6] W. Zhuang, X. Gan, Y. Wen, S. Zhang, S. Yi, Collaborative unsupervised visual representation learning from decentralized data, in: Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 4912–4921.
- [7] J. Li, L. Lyu, D. Iso, C. Chakrabarti, M. Spranger, Mocosfl: Enabling cross-client collaborative self-supervised learning, in: The Eleventh International Conference on Learning Representations, 2023. URL <https://openreview.net/forum?id=2QGJXyMNoPz>
- [8] L. Wang, K. Zhang, Y. Li, Y. Tian, R. Tedrake, Does learning from decentralized non-IID unlabeled data benefit from self supervision?, in: The Eleventh International Conference on Learning Representations, 2023. URL <https://openreview.net/forum?id=2L9gzS80tA4>
- [9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).
- [10] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, in: International Conference on Learning Representations, 2021. URL <https://openreview.net/forum?id=YicbFdNTTy>
- [11] X. Chen, S. Xie, K. He, An empirical study of training self-supervised vision transformers, in: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), IEEE Computer Society, Los Alamitos, CA, USA, 2021, pp. 9620–9629. doi:10.1109/ICCV48922.2021.00950. URL <https://doi.ieeecomputersociety.org/10.1109/ICCV48922.2021.00950>
- [12] A. Van den Oord, Y. Li, O. Vinyals, Representation learning with contrastive predictive coding, arXiv e-prints (2018) arXiv:1807.
- [13] W. Zhuang, Y. Wen, S. Zhang, Divergence-aware federated self-supervised learning, arXiv preprint arXiv:2204.04385 (2022).
- [14] A. Saeed, F. D. Salim, T. Ozcelebi, J. Lukkien, Federated self-supervised learning of multisensor representations for embedded intelligence, IEEE Internet of Things Journal 8 (2) (2020) 1030–1040.
- [15] S. Ek, R. Rombourg, F. Portet, P. Lalanda, Federated self-supervised learning in heterogeneous settings: Limits of a baseline approach on har, in: 2022 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops), IEEE, 2022, pp. 557–562.
- [16] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar, et al., Bootstrap your own latent: A new approach to self-supervised learning, Advances in neural information processing systems 33 (2020) 21271–21284.
- [17] Y. Wu, D. Zeng, Z. Wang, Y. Shi, J. Hu, Distributed contrastive learning for medical image segmentation, Medical Image Analysis 81 (2022) 102564.
- [18] N. Dong, I. Voiculescu, Federated contrastive learning for decentralized unlabeled medical images, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2021, pp. 378–387.
- [19] D. Makhija, N. Ho, J. Ghosh, Federated self-supervised learning for heterogeneous clients, arXiv preprint arXiv:2205.12493 (2022).

- [20] Y. Wu, Z. Wang, D. Zeng, M. Li, Y. Shi, J. Hu, Decentralized unsupervised learning of visual representations, arXiv preprint arXiv:2111.10763 (2021).
- [21] F. Zhang, K. Kuang, L. Chen, Z. You, T. Shen, J. Xiao, Y. Zhang, C. Wu, F. Wu, Y. Zhuang, et al., Federated unsupervised representation learning, *Frontiers of Information Technology & Electronic Engineering* 24 (8) (2023) 1181–1193.
- [22] G. E. Hinton, S. Osindero, Y.-W. Teh, A fast learning algorithm for deep belief nets, *Neural Comput.* 18 (7) (2006) 1527–1554. doi:10.1162/neco.2006.18.7.1527. URL <https://doi.org/10.1162/neco.2006.18.7.1527>
- [23] Y. Bengio, P. Lamblin, D. Popovici, H. Larochelle, Greedy layer-wise training of deep networks, in: *Proceedings of the 19th International Conference on Neural Information Processing Systems, NIPS'06*, MIT Press, Cambridge, MA, USA, 2006, p. 153–160.
- [24] H.-P. Wang, S. Stich, Y. He, M. Fritz, Progfed: Effective, communication, and computation efficient federated learning by progressive training, in: *International Conference on Machine Learning*, PMLR, 2022, pp. 23034–23054.
- [25] P. Guo, W. Morningstar, R. Vemulapalli, K. Singhal, V. Patel, P. Mansfield, Towards federated learning under resource constraints via layer-wise training and depth dropout, in: *The 4th Workshop on practical ML for Developing Countries: learning under limited/low resource settings @ ICLR 2023*, 2023. URL [https://pml4dc.github.io/iclr2023/pdf/PML4DC\\_ICLR2023\\_5.pdf](https://pml4dc.github.io/iclr2023/pdf/PML4DC_ICLR2023_5.pdf)
- [26] Z. Huo, D. Hwang, K. C. Sim, S. Garg, A. Misra, N. Siddhartha, T. Strohmman, F. Beaufays, Incremental layer-wise self-supervised learning for efficient speech domain adaptation on device, arXiv preprint arXiv:2110.00155 (2021).
- [27] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, in: *International Conference on Learning Representations*, 2021. URL <https://openreview.net/forum?id=YicbFdNTTy>
- [28] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick, Microsoft coco: Common objects in context, in: *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*, Springer, 2014, pp. 740–755.
- [29] A. Krizhevsky, G. Hinton, et al., Learning multiple layers of features from tiny images, Technical report, University of Toronto, 2009 (2009). URL <https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf>
- [30] Y. Le, X. Yang, Tiny imagenet visual recognition challenge, *CS 231N* 7 (7) (2015) 3.
- [31] F.-F. Li, M. Andreto, M. Ranzato, P. Perona, Caltech 101 (Apr 2022). doi:10.22002/D1.20086.
- [32] P. Goyal, P. Dollár, R. Girshick, P. Noordhuis, L. Wesolowski, A. Kyrola, A. Tulloch, Y. Jia, K. He, Accurate, large minibatch sgd: Training imagenet in 1 hour, arXiv preprint arXiv:1706.02677 (2017).
- [33] A. Krizhevsky, One weird trick for parallelizing convolutional neural networks, arXiv preprint arXiv:1404.5997 (2014).
- [34] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, PyTorch: An Imperative Style, High-Performance Deep Learning Library, in: H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, R. Garnett (Eds.), *Advances in Neural Information Processing Systems 32*, Curran Associates, Inc., 2019, pp. 8024–8035.
- [35] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, V. Smith, Federated optimization in heterogeneous networks, *Proceedings of Machine Learning and Systems* 2 (2020) 429–450.
- [36] Q. Li, B. He, D. Song, Model-contrastive federated learning, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 10713–10722.
- [37] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [38] J. Yang, R. Shi, D. Wei, Z. Liu, L. Zhao, B. Ke, H. Pfister, B. Ni, Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification, *Scientific Data* 10 (1) (2023) 41.
- [39] T. S. Ferguson, A bayesian analysis of some nonparametric problems, *The annals of statistics* (1973) 209–230.
- [40] Lennart, Jsevimol, M. Hobbhahn, T. Besiroglu, anson.ho, Estimating training compute of deep learning models (2022). URL <https://www.lesswrong.com/posts/HvqQm6o8KnwxbdmhZ/estimating-training-compute-of-deep-learning-models>
- [41] M. Hobbhahn, Jsevimol, What's the backward-forward flop ratio for neural networks? (2021). URL <https://www.lesswrong.com/posts/fnjKpBoWJXcSDwhZk/what-s-the-backward-forward-flop-ratio-for-neural-networks>
- [42] D. Amodei, D. Hernandez, G. Sastry, J. Clark, G. Brockman, I. Sutskever, Ai and compute, <https://openai.com/research/ai-and-compute> (2018).
- [43] M. Hobbhahn, J. Sevilla, What's the backward-forward flop ratio for neural networks?, accessed: 2023-12-15 (2021). URL <https://epochai.org/blog/backward-forward-FLOP-ratio>
- [44] c. v. t. FAIR, fvc core library (2021). URL <https://github.com/facebookresearch/fvc core/>