Beyond Labels and Centralisation: Representation Learning Through Data Curation

Özgü Göksu

Submitted in fulfilment of the requirements for the
Degree of Doctor of Philosophy

School of Computing Science
College of Science and Engineering
University of Glasgow

November 2025

Abstract

Deep learning has achieved remarkable progress in vision, language, and multimodal tasks; however, its success remains heavily dependent on centralised, large-scale, and fully labelled datasets such as Imagenet, Open lab V7, etc. In real world, data is frequently limited, unlabelled, privately owned, and distributed across many devices, making traditional supervised learning challenging to scale. These limitations motivate the development of robust, novel representation learning methods capable of addressing under unlabeled, limited, and heterogeneous data constraints.

This thesis addresses these challenges through four main contributions. First, we propose a self-supervised batch curation strategy that scores unlabeled data (batches) using the Fréchet ResNet Distance (FRD) as bad or good, enabling the semantically related and informative batches to improve feature quality under limited and unlabeled data regimes. Second, we introduce FedMPR, a federated parameter-selection framework that adaptively prunes irrelevant weights during local training for each client, improving representation robustness and generalization under highly non-i.i.d. data settings. To further analyse distributional data heterogeneity, the thesis also presents CelebA-Gender, a novel gender classification dataset designed to evaluate complex attribute-based data shifts and compare to the real world cases. Third, we present FedQuad, a framework that incorporates a reformulated quadruplet loss to minimise intra-class distance and maximise inter-class distance while mitigating representational collapse on global representation space. Finally, the thesis investigates partial federated model training combined with self-supervised learning, leveraging a frozen DINOv3 as a backbone and a lightweight projection head (Multilayer) to enable robust and computation-efficient representation learning under extreme client heterogeneity and limited participation.

As a result, our experiments on many benchmarks such as CIFAR10, CIFAR100, Tiny-ImageNet, and CelebA-Gender demonstrate that the proposed algorithms consistently outperform existing baselines in terms of accuracy, representation robustness, and feature consistency across many federated scenarios. To sum up, these contributions advance representation learning by enabling more generalisable, efficient, and data in one place based systems learning without relying on large, labelled, or centrally collected datasets.

Contents

List of Tables

List of Figures

Acknowledgements

This thesis marks the end of a long, difficult and life-changing journey, one that would not have been possible without the incredible people who supported me along the whole way.

I owe my deepest gratitude to my dear supervisor, Dr Nicolas Pugeault, whose guidance has been a constant source of strength, motivation and inspiration. His patience, kindness, and belief in me, even when I doubted myself, made this journey possible. He has not only been a supervisor but also a awesome mentor who taught me how to think critically, embrace uncertainty, and find meaning in the process of research. Without his motivation, encouragement, and support, I would not have been able to complete this PhD. His motivation in me gave me strength during the most challenging times of this research journey.

I would like to thank Dr. Gerardo Aragon-Camarasa and Dr. Tanaya Guha for their invaluable feedback, challenging questions and support throughout these years. Their insights and guidance have enriched both my research and my academic journey. Also, I am deeply grateful to the many academics, colleagues, and friends within the School of Computer Science (SoCS) who have supported me along the way, in particular, my dear friends; Daniela Ivanova, Melonie de Almeida, Le Li, Rory Young, and Ozan Bahadir whose kindness, humour, and brainstorming times brought light and laughter to the final of this PhD journey.

I would like to thank my sincere gratitude to the Republic of Türkiye Ministry of National Education for their support, guidance, and funding throughout my doctoral studies under the YLSY Scholarship Programme.

To my parents, Ercan and Hatice, and to my brothers, Turan and Mehmet, thank you for your unconditional love and endless encouragement and great support. To my lovely friends Elif, Gülşah, Ferda, Hürmüz, Zehra, and Mahla; thank you for your motivation, understanding, and light into the most difficult days. Your friendship has meant more to me than words can ever explain. And thanks to my friends in Scotland; Songül, İdil, Kadir, Bilgi, Ahmet Burak and Elif; you made a foreign place feel like a home.

Finally, my thanks go to everyone who supported and inspired me throughout the journey, each of you has left an unforgettable mark on this work and on my life. A large language model (LLM) tool (ChatGPT) was used exclusively to assist with grammar rules, paraphrasing and punctuation checks. All ideas, contributions, methods, codes and analyses in the thesis are my own work.

Declaration

I hereby declare that this thesis is my original work and has not been submitted previously for any degree or qualification at any other academic institution, or university. The whole source of the information, data, and literature works used in this thesis have been appropriately acknowledged and referenced.

For chapters presented in journal format (Chapters 3, 4, 5, and 6), the sections containing the authors's versions of published or submitted manuscripts are each prefaced with a statement detailing the contribution of all named authors to the work. Also, these submission details are mentioned in each chapter.

Note for the format of this thesis

This thesis includes four chapters presented in journal format; Chapters 3, 4, 5, and 6. Chapters 3, 4, and 5 are based on manuscripts that have been submitted to conferences, while Chapter 6 is currently under review for submission to a conference. These chapters have been integrated in a format consistent with the rest of the thesis, and the numbering of pages, figures, tables, and references follows the overall sequence of the document.

Chapter 1

Introduction

"If intelligence is a cake, the bulk of the cake is unsupervised learning, the icing on the cake is supervised learning and the cherry on the cake is reinforcement learning."
— Yann LeCun's cake analogy at NIPS 2016.

The question of whether machines can deeply think, understand, or learn has been a central topic of interest since Alan Turing's famous question, *"Can machines think?"* Turing's proposal of the Imitation Game shifted the understanding of intelligence from abstract concept to something that can be evaluated through observable behaviour in this work Turing, 2007. This perspective provided the foundation for later the field of Artificial Intelligence (AI) and continues to influence how we study machine intelligence nowadays.

In recent years that followed, AI has expanded into several subfields, mostly machine learning and deep learning, where models are trained to detect patterns, learn from data, and make decisions. Part of the motivation for these learning systems came from neuroscience experiment. Early neurophysiological work (including the well-known experiments by Hubel and Wiesel on the visual cortices of cats mentioned in the paper Hubel and Wiesel, 1959), showed that individual neurons selectively respond to specific visual features, such as edges, lines or orientations. These findings inspired the development of artificial neural networks, where simple computational units (neurons) can learn to combine complex representations increasingly.

Further investigations led to multilayer perceptrons (MLPs) and, eventually, deep neural networks or models, architectures capable of hierarchical feature learning. In these models, early layers capture simple patterns (sometimes they are called as low level features) such as edges or textures, while deeper layers represent increasingly abstract and semantic features (high level features). This paradigm transformed areas like computer vision, natural language processing enabling state-of-the-art performance across a wide range of applications.

In spite of these successes, deep learning struggles with significant limitations. Many cutting-edge deep learning models are fundamentally data-hungry: effective training relies on

large number of samples with high-quality, well-annotated data (labelled) to avoid overfitting and achieve generalisation. However, collecting and labelling data is significantly expensive, time-consuming, and mostly requires domain expertise in a specific area. For example, in medical imaging, annotations or labels typically need to be provided by trained radiologists or clinicians, making dataset creation limited by expert availability. In addition to expertise challenge, medical data contains sensitive patient information and is subject to strict privacy regulations such as HIPAA ("Health Insurance Portability and Accountability Act of 1996 (HIPAA)," n.d.) and GDPR ("General Data Protection Regulation (GDPR), Regulation (EU) 2016/679," n.d.). These legal constraints hinder the sharing of data across many institutions, universities, leading to isolated data silos. Similar issues occur in other privacy or sensitive domains, including finance, defence, and personalised services, where sharing raw data risks exposing confidential or personally information (Kairouz et al., 2021).

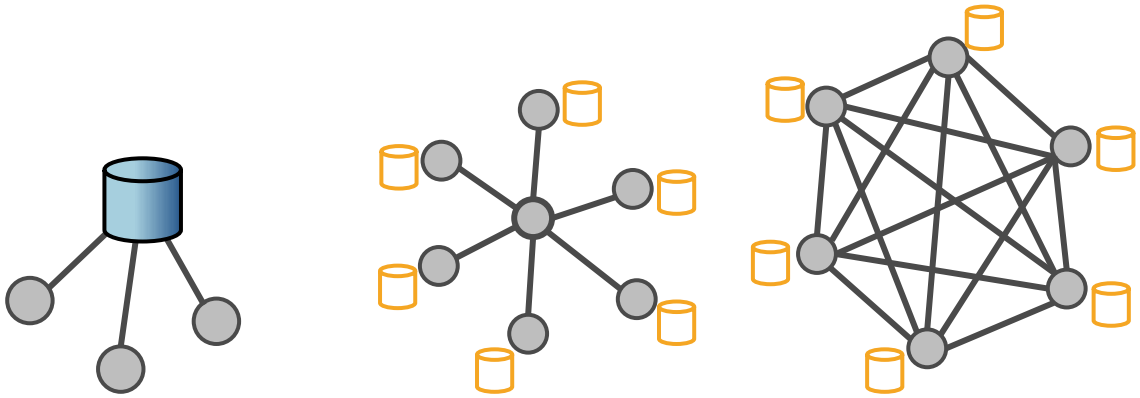
Beyond the data accessibility concerns, centralising massive datasets in one place introduces computational resources and infrastructural burdens. Centralizing data into a single location (cloud, server, etc.) increases memory usage, storage problem, and bandwidth requirements, while also introducing the risks related to data leakage and unauthorised access. These challenges make large-scale centralised training increasingly impractical and inefficient, notably in domains involving sensitive or high-volume data (Bonawitz et al., 2019).

In addition, many real-world systems suffer from limited, heterogeneous, or entirely unlabelled data, which poses a significant barrier to the sustained development of modern deep learning. These challenges are particularly evident in fields such as medicine, autonomous systems, and remote sensing, where collecting large, labelled datasets is either expensive or not feasible. As a result, a major focus of contemporary machine learning research has been on developing methods that can learn robust representations despite data scarcity, heterogeneity, or the absence of supervision.

This research gap raises an important question: *Can machines learn robust representations where data is limited, not homogeneous, and unlabelled?* Expending this gap with highly important today's cases: *Can they collaborate and share knowledge without centralising data?*. These questions are the core motivation of this work, guiding the exploration of decentralised, distributed learning without sharing raw data frameworks capable of extracting robust representations under many data conditions.

An encouraging direction for addressing the challenges of data heterogenity, limitation and distribution lies in unsupervised and decentralised learning paradigms. Unlike supervised learning, which depends heavily on large annotated datasets, unsupervised methods aim to capture feautures and patterns directly from unlabelled data. Approaches such as Vision Transformers (ViTs), self-supervised learning objectives, and contrastive learning have shown that strong features can be learned without explicit labels or annotations, reducing leaning on human annotation. These methods are especially advantageous in domains where labelled data are scarce but

unlabelled datasets have a large number of samples and sufficient diversity and class balance.



(a) Centralized Learning (b) Centralized FL (c) Decentralized FL

Figure 1.1: Three paradigms of learning: (i) *centralised learning*, where all raw data are aggregated on a single server; (ii) *centralised federated learning*, where clients have local data and a central server aggregates model updates to obtain a global model, and (iii) *decentralised federated learning*, where clients exchange model updates peer-to-peer without a central server.

Figure 1.1 illustrates three fundamental paradigms of representation learning based on data: centralised learning, centralised federated learning, and decentralised federated learning. In centralised learning, the typical paradigm in many machine learning algorithms, the whole dataset is aggregated and stored within a single repository. The training of each model proceeds by accessing the centralized data.

In centralised federated learning, a central server has a global model and test data for generalization perfromance. Model training is proceeds by each collobrated local model. and each local model has its own raw data. Rather than sharing raw data, clients transfer model updates such as gradients or parameters, weights, which the server aggregates to update the global model.

On the other hand, in decentralised federated learning, the dependency on a central server is no longer required. Clients engage in direct peer-to-peer communication, exchanging model updates with one another and collaboratively optimising. This paradigm enables distributed learning systems without a central server or model, improving robustness and reducing the impacts of model aggregation risks.

Federated Learning (FL) has emerged as a promising paradigm for distributed machine learning systems, enabling collaborative model training without centralising raw data. In FL, data remain on local devices or institutional spaces (clients), while model weights or gradients are communicated with a central server. This allows multiple clients to contribute to a shared global model without sharing or transferring their private data. By doing so, FL bridges the gap between decentralised training and no data sharing with many clients, making large-scale collaboration practical across hospitals, research centers, institutions, or mobile devices.

However, FL offers significant advantages (data is stored on many local servers, efficient collaboration), it introduces new challenges grounded in data heterogeneity. In real-world settings, client datasets are rarely identically distributed (i.i.d.), they reflect diverse user behaviours, sensor modalities, and absent samples. Such non-i.i.d. can cause local model drift, slow or unstable convergence, and even representational collapse on global representation space, where the learned embeddings fail to capture robust or discriminative features.

This thesis is structured around addressing the limitations of representation learning with limited or unlabelled data, as well as the challenges of learning under data heterogeneity in federated settings with limited data availability. Firstly, we investigate the limitations that arise when learning from limited or unlabelled data in centralised unsupervised environments. Following the presented method, we propose a novel bad-batch elimination strategy that uses similarity measures based on the Fréchet Inception Distance (FID) score. This method analyses augmented batches and removes low-quality transformed views or misleading batches, such as batches distorted by weak data augmentations, thereby improving training stability and boosting generalisation in scenarios where labelled data are limited.

Then, we focus on collaborative learning under federated settings and introduce three robust and novel methods. The metric learning based method, FedQuad (Federated Stochastic Quadruplet Learning for Representational Collapse), incorporates a reformulated quadruplet loss into federated learning. FedQuad relies on learning minimazing the distances between same class samples while maximizing inter-class variance. Therefore, the assumption prevents representational collapse and unstable feature learning across heterogeneous clients. By combining assumptions from metric learning, FedQuad advances the development of FL systems that are data heterogeneous and robust to non-i.i.d. distributions.

The other presented method, FedMPR (Hybrid-Regularised Magnitude Pruning for Robust Federated Learning under Covariate Shift) tackles with data heterogeneity problem. While FedQuad focuses on the distances of learned representations, FedMPR addresses the negative effects of model aggregation where either each weight is assumed equal or explicitly considered. This reduces the influence of irrelevant or noisy features and mitigates the aggregating updates from highly heterogeneous data.

In this part, CelebA-Gender, a gender-classification dataset designed specifically for evaluating complex and diverse federated learning scenarios in the thesis. Unlike traditional approaches that rely on a single attribute or curated labels, this dataset presents genders with multiple CelebA data attributes, making it inherently richer, more complex and representative of real-world non-i.i.d. conditions. As a result, CelebA-Gender provides a more challenging and realistic benchmark for studying federated learning under heterogeneous data distributions.

After addressing data heterogeneity in supervised federated settings, we extend our investigation to unsupervised decentralised federated learning. In this part of the thesis, we introduce a novel method that leverages federated learning with a large number of clients (e.g., 2,000) and

uses a pretrained backbone while training only a lightweight projection head in the federated setting. This framework operates effectively on both limited labelled and limited unlabelled data. Our results demonstrate that a well-pretrained backbone like DiNOv3, combined with high client diversity and considerable data heterogeneity, successfully enhances unsupervised representation learning in large-scale federated environments.

In combination, these contributions advance the broader goal of developing representation learning methods that maintain strong performance despite data heterogeneity, limited supervision, and decentralised data constraints. By integrating insights from self-supervised learning, unsupervised representations, and federated optimisation, this thesis seeks to push the frontier of collaborative training without accessing raw data, robust representations, and efficient distributed machine learning.

1.1 Research Challenges

The process of data collection and annotation in machine learning mostly demands significant human expertise within specific areas. Unfortunately, this process is both time-consuming and expensive, as it requires skilled professionals to accurately label large, diverse datasets. To address these limitations, self-supervised learning (SSL) methods have gained significant attention in recent years since SSL address the representation learning where labeled data is limited. These approaches enable models to learn representations directly from raw data (mostly unlabeled), allowing them to extract and transform features into downstream tasks such as image classification, object detection, thereby considerably reducing dependence on manual annotation or labels.

However, data availability and manual annotation are not the only challenges limiting the scalability of machine learning. The traditional practice of keeping whole data into a single centralised repository introduces serious accessibility, security, and ethical concerns. In sensitive domains such as healthcare, finance, biometrics and other personalised services (e.g. social media), sharing raw data among many clients can poses risks to privacy regulations and may violate ethical standards.

Therefore, storing data on a central server or transferring data from the server to other clients is neither practical nor legally compliant in many real-world scenarios. These constraints have motivated the development of federated learning (FL) paradigms, where models are collaboratively trained across distributed data sources without leaking or transferring the underlying data. Thus, federated learning provides a sufficient framework for tackling scalability, data accessibility and ethical challenges in several sensitive domains. However, federated learning assumes that data is preserved and private. However, during the whole communication, gradients and weights may cause data leakage problems and privacy risks (S. Park et al., 2025, L. Jiang et al., 2025). In the thesis, we are focusing on representation learning developments, not data privacy

concerns.

Despite the considerable development enabled by representation learning approaches both in distributed settings and in scenarios with limited or no human supervision, several key challenges remain, especially those related to data availability, accessibility, and the practical constraints imposed by heterogeneous data distributions. These are the main challenges addressed throughout this research;

- **Limited Data:** When data are extremely scarce, deep neural architectures commonly fail to capture adequate features, leading to underfitting and limited generalisation capability. In contrast, models that learn robust representations tend to generalise more effectively across several tasks and seen or unseen domains (Ilyas et al., 2019). In order to achieve this robustness, many existing approaches mostly rely on supervised learning with large-scale, well-annotated datasets such as ImageNet (containing approximately 14 million images, (J. Deng et al., 2009)) or LAION (comprising around 5.85 billion image–text pairs, (Schuhmann et al., 2022)). However, accessing, collecting and curating such extensive datasets is not always feasible due to practical limitations, including high annotation costs, the need for domain expertise, and time-consuming.

- **Data accessibility:** Centralised data collection raises serious accessibility risks, as it frequently involves handling sensitive or personal, individual-specific information. In many domains, such as medical, finance, and biometric data contains highly confidential records, are protected by strict regulations, including GDPR in Europe and HIPAA in the United States. Transferring or storing this data in a centralised repository, server or cloud not only increases the risk of data leakage and unauthorised access but can also lead to critical legal and ethical consequences, which are mentioned in this paper Rieke et al., 2020. Moreover, even when explicit identifiers like social security numbers and phone numbers are removed, re-identification attacks can leverage correlations in high-dimensional data (e.g., facial images, genomic sequences, or location traces) to reconstruct individual identities and disclose personal information to further accessibility risks.

  Therefore, ensuring data confidentiality, adherence to regulations and compliance with ethical standards have become a central challenge in collaborative learning environments. The collaborative organisations are increasingly restricted from sharing or accessing raw data across institutional or local network boundaries, creating isolated data silos that limit the scalability of modern machine learning. These limitations underscore the urgent need for data integrity and secure learning paradigms that enable knowledge sharing without requiring direct data exchange.

- **Data Heterogeneity:** In distributed systems, data heterogeneity, commonly referred to as non-i.i.d. (non-independent and identically distributed), is a key challenge that arises when

each client has its own dataset, and these datasets often exhibit significantly heterogeneous distributions.

This heterogeneity appears in several forms, including feature distribution shifts, class imbalance, unequal sample size, and multi-source image datasets; each of these contributes to biased global model updates and degraded global model performance. For instance, in remote sensing, satellites and aerial platforms often rely on several sensors from different manufacturers or with varying spectral, spatial, and radiometric characteristics (B. Han et al., 2024). These differences, such as resolution, spectral bandwidth, illumination conditions, and atmospheric disturbances, cause each data source to capture distinct feature distributions, leading local models to learn sensor-specific rather than globally consistent representations.

In the same way, in autonomous driving, vehicles operating in different regions encounter diverse lighting, weather, and traffic conditions, resulting in conflicting feature representations across many clients (Marathe et al., 2022). Likewise, label distribution skew further compounds the problem, as seen in facial recognition tasks where some clients's data may consist predominantly of one gender or age group. Another example, in handwritten digit recognition, where clients may possess only a limited subset of digits.

Also, unequal samples size poses a major challenge; for example, in mobile keyboard prediction Hard et al., 2018, active users contribute vast amounts of text while others provide only a few samples, assigning excessive weight to certain clients during global model aggregation. As a whole, these varieties of data heterogeneity cause client or model drift, where local models overfit to their specific data distributions rather than converging toward a shared global optimum. This divergence slows convergence, reduces generalisation ability of global model and clients, and may lead to representational collapse in extreme cases.

In summary, these challenges collectively emphasise the necessity for representation learning frameworks that can operate sufficiently under data scarcity, learning without human supervision, and highly distributed diversity across many clients, while maintaining data confidentiality. To address these issues, this thesis proposes four main approaches: (1) a *self-supervised batch curation strategy* for improving representation quality from unlabelled data when we have centralised and limited data. The rest of the presented approaches are decentralised, distributed learning without raw data sharing; (2) a *FedMPR, federated parameter selection mechanism* that enhances convergence stability and model robustness under non-i.i.d. data, and (3) *FedQuad, minimazing intra class distance based metric learning framework*, a federated stochastic quadruplet learning framework that integrates metric learning principles to mitigate representational collapse and strengthen representation separation via optimising inter-class variance. (4) Pre-trained backbone for partial federated model learning can handle data heterogeneity for unlabeled data with

many client settings. In combination, these contributions advance the broader goal of achieving robust, generalizable, efficient and without raw data sharing based distributed representation learning in several data-constrained environments.

1.2 Thesis Statement

Representations are simplified versions of raw image data that capture its visual patterns and features. Machine learning models rely on these rich, high-quality representations (also known as embeddings) to perform image classification; however, scarce and decentralised data commonly result in weak, non-generalisable and ineffective features. This thesis explores self-supervised and federated learning strategies to overcome these limitations, enabling robust representation learning under limited, heterogeneous data and data accessibility conditions.

1.3 Motivation

Modern machine learning has achieved remarkable success across a wide range of domains. Despite its significant progress, the majority of modern machine learning approaches assume that accessing data is easy, data can be stored in one device and computing the data is practical. However, in the real world, data are mostly limited in quantity, unlabelled, and distributed across multiple sources due to limitations in data sharing, data ownership challenges, and feature heterogeneity. These limitations make traditional centralised learning paradigms impractical, as they struggle to generalise sufficiently when faced with class imbalance, distributional data shifts, or limited human supervision.

This research is driven by the high-priority requirement to develop representational learning frameworks that can operate effectively under realistic and challenging data conditions. Specifically, this thesis aims to enable machines to learn robust and discriminative features from limited, unlabelled, and heterogeneous data without relying on centralised data aggregation. Addressing these challenges is extremely critical not only for advancing the scalability and fairness of machine learning but also for promoting accessibility constraints and inclusive model development across many institutions and devices.

In pursuit of this objective, this research introduces a series of methods designed to overcome the core limitations of traditional data-hungry-based representation learning approaches. By focusing on unsupervised representation learning and federated representation learning, the proposed frameworks seek to achieve robust representations, effective generalisation and high adaptability under distributed, data-scarce, without human annotation types of settings. In conclusion, this thesis aims to balance the trade-offs between data efficiency and representation robustness, contributing to the development of more representative, reproducible, and scalable machine learning systems.

1.3.1 Research Questions

While we have investigated the recent developments in this field, several research questions remain unexplored and unanswered.

- **RQ1** How can we improve self-supervised contrastive learning methods under limited data conditions, where augmentations may introduce misleading (false positive/negative) views?
- **RQ2** How can covariate shift in federated learning be mitigated to enable robust and generalizable training when clients possess limited and non-overlapping datasets?
- **RQ3** What is the effect of inter-sample learning distances within each local model on the global model's generalisation capability?
- **RQ4** How can we design learning frameworks that remain effective under high data heterogeneity across clients, particularly when client participation is intermittent or partial during global aggregation?

To address these research questions, the thesis presents a systematic investigation that progresses through multiple learning paradigms. It begins with traditional centralised representation learning under conditions of limited and unlabelled data. Building on this basis, this study progresses to centralised federated learning challenges, exploring how representation learning can be maintained when data are scarce and distributed across clients. Finally, the research concludes in decentralised federated representation learning scenarios, where data are both unlabelled and limited, aiming to develop robust and generalizable embeddings despite substantial data heterogeneity and data accessibility constraints.

1.3.2 Scope

This research mainly focuses on developing robust representation learning frameworks that are capable of operating effectively under limited, unlabelled, and diverse data conditions within both centralised and decentralised environments. The study systematically investigates representation learning algorithms from three different perspectives:

- **Centralised Learning:** Investigating robust representation learning methods using limited and unlabelled data to establish an essential understanding of data efficiency and feature robustness.
- **Centralised Federated Learning:** Extending representation learning to multi-client environments in which data are partitioned and subject to communication limitations, analysing how distributed optimisation affects representation quality and model generalisation.

- **Centralised Federated with Unsupervised Learning:** Advancing to distributed scenarios where data are not only limited and heterogeneous but also unlabelled, this thesis places particular emphasis on data accessibility, generalizable, and collaborative representation learning under both cross-device and cross-silo settings.

This thesis focuses on the development of image-based representation learning methods, with particular focus on federated learning and self-supervised learning paradigms. The scope is deliberately limited to settings in which data are distributed, heterogeneous, and unlabelled, reflecting real-world limitations experienced in cross-device and cross-silo environments. Within this domain, the thesis explores metric-based objectives, contrastive learning frameworks, and federated strategies that address challenges of data scarcity, heterogeneity, accessibility, and representational robustness.

Several areas lie outside the scope of this research. Primarily, reinforcement learning, high-dimensional representation learning (3D, 4D), video-based learning tasks, sequence modelling, and natural language processing tasks are not considered, as the methodological focus centres on visual representations. Likewise, the new design and training of generative models are excluded. In addition, this work does not investigate communication-efficient protocols, secure aggregation mechanisms, homomorphic encryption, personalisation, fully secure distributed systems or differential accessibility beyond their conceptual relevance to federated settings.

These boundaries allow this research to maintain a straightforward focus on representation-based approaches while providing methodological depth. The objective of this research is to develop a coherent and realistically practical framework for learning robust, data accessibility constraints, and generalizable representations in distributed environments by focusing on the availability of limited image samples and label-efficient training.

1.4 Contributions

This thesis makes the following key contributions to the fields:

1. **Self-Supervised Batch Curation:** introduces a novel self-supervised batch curation strategy that evaluates and selects informative batches from unlabelled datasets using the Fréchet ResNet Distance (FRD) scores. This approach enhances the quality of learned representations by reducing redundancy and highlighting weakness of augmentations with limited data. The approach enables more efficient and effective self-supervised learning in scenarios where labelled data are inadequate or unavailable. (Chapter 3)

2. **Federated Parameter Selection and Novel Dataset for Evaluation:** Proposes a federated parameter selection framework that dynamically adjusts the contribution of each client during global model aggregation to improve convergence stability and robustness of embeddings under non-i.i.d. conditions.

In addition, the thesis introduces CelebA-Gender benchmark, a novel gender classification dataset derived from the widely used CelebA dataset (Z. Liu et al., 2018). CelebA-Gender is notably designed to simulate complex attribute-based data heterogeneity by splitting data across many clients based on gender and specified multi-facial attributes. This dataset provides a beneficial benchmark for evaluating federated learning algorithms under more realistic, heterogeneous, and imbalanced data distributions. (Chapter 4)

3. **Federated Metric Learning:** develops FedQuad, a Federated Stochastic Quadruplet Learning framework that integrates a reformulated quadruplet loss with distance-based representation learning. FedQuad, explicitly balances intra-class variance and inter-class variance while addressing representational collapse across decentralised federated learning. This framework enhances global model generalisation and bot local and global representation alignment in federated scenarios, especially under high heterogeneity. (Chapter 5)

4. **Unsupervised Federated Partial Model Training** investigates federated projection head training under a self-supervised learning paradigm, focusing on scenarios with highly diverse and limited data distributions and no human supervision. The proposed approach leverages a pretrained backbone network while training only a lightweight projection head across many clients. By enabling distributed learning without requiring full model training from scratch or a large amount of labelled data, this method clearly demonstrates that even with a small proportion of participating clients among a large, heterogeneous population, it is possible to learn robust and generalizable feature representations under significant data heterogeneity. (Chapter 6)

Considering these key contributions together, they advance the state-of-the-art in unsupervised and federated learning with deep analysis. The presented methods and dataset address critical challenges related to data scarcity, non-i.i.d. distribution, and client data imbalance, limited client participation and enabling more robust features for representation learning.

1.4.1 List of Papers

These are the list of published papers during my research;

- "The Bad Batches: Enhancing Self-Supervised Learning in Image Classification Through Representative Batch Curation", Ozgu Goksu, Nicolas Pugeault, IEEE WCCI-International Joint Conference on Neural Networks (IJCNN), 2024.

- "Hybrid-Regularised Magnitude Pruning for Robust Federated Learning under Covariate Shift", Ozgu Goksu, Nicolas Pugeault, The International Symposium on Edge Intelligence, Trustworthy and Decentralised Artificial Intelligence (iEDGE), 2025.

- "FedQuad: Federated Stochastic Quadruplet Learning to Mitigate Data Heterogeneity", Ozgu Goksu, Nicolas Pugeault, The 3rd IEEE International Conference on Federated Learning Technologies and Applications (FLTA25), 2025.

1.5 Thesis Overview

This thesis is organised into seven chapters, each building upon the previous to develop a comprehensive understanding of visual representation learning under many scenarios.

Chapter 1 – Introduction: presents the problem definition, motivation, thesis statement and research challenges associated with representation learning from limited, heterogeneous, no human supervision and without raw data sharing. The chapter presents the main research questions, outlines the research scope, explains the thesis statement and summarises the key contributions of this thesis.

Chapter 2 – Background: section examines the key foundations of deep learning, with emphasis on unsupervised and self-supervised representation learning, and reviews federated learning as a paradigm associated with data accessibility, collaborative training cases for visual representation learning. Despite this progress, existing methods show notable limitations under *data scarcity* (few labels, limited samples per client, low-data regımes) and *data heterogeneity* (non-i.i.d. distributions, class imbalance, covariate shift, label skewness). In particular, (i) most unsupervised contrastive objectives rely on sufficient pairs or misleading views of positive/negative images, which usually cause weak embeddings; (ii) Global weight averaging commonly weakens the client-specific structure, resulting in unstable optimisation or representational overlaps. These challenges mainly motivate the approaches proposed in this thesis, which aim to learn robust, efficient representations under limited supervision with heterogeneous, distributed data.

Chapter 3 – Self-Supervised Batch Curation: introduces a novel self-supervised batch curation strategy, designed to enhance representation learning from unlabeled and limited datasets. The chapter explains the benefits of the Fréchet Resnet Distance (FRD) for evaluating batch-level image consistency by computing batch scores from embeddings during training, and demonstrates improvements in representation robustness.

Chapter 4 – Federated Parameter Selection and CelebA-Gender Dataset: proposes a novel federated parameter selection method to stabilise global model convergence and improve global model performance under non-i.i.d. data. This chapter also presents the CelebA-Gender benchmark, a novel gender classification dataset designed to simulate complex attribute-based heterogeneity for evaluating federated learning algorithms.

Chapter 5 – FedQuad: Federated Stochastic Quadruplet Learning: This study develops and evaluates FedQuad, a federated learning framework that integrates a reformulated quadruplet loss to mitigate representational collapse and minimise intra-class variance within each client.

Extensive experiments demonstrate that FedQuad effectively enhances representation consistency and generalisation under limited, diverse, and heterogeneous data conditions.

Chapter 6 – FedDINOv3: proposes a framework where a DINOv3-based backbone, combined with a federated projection head under a self-supervised training, to address data heterogeneity. The study investigates practical model training in large-scale federated settings (cross-silo) with highly diverse data distributions and limited client participation per round. As a result, our findings in the section show robust representation learning can be achieved under partial model training locally with high data heterogeneity.

Chapter 7 – Conclusion and Future Work: This chapter mainly summarises the key findings, outcomes, and generally discusses the importance of the proposed algorithms, and outlines potential research directions for future data privacy, unsupervised, centralised representation and decentralised federated representation learning.

To sum up, these chapters present an organised progression from centralised, data-limited representation learning towards decentralised federated systems. The work bridges theoretical insights and ideas with practical implementations, contributing to the development of more robust visual representation learning methods and more generalisable machine learning frameworks capable of collaborative learning under many data curation cases, such as data limitation and heterogeneity, without sharing raw data.

Chapter 2

Background

This chapter presents the theoretical foundations and related work that form the basis of the proposed research. The chapter begins by introducing the key concepts of deep representation learning, which provide an understanding of how neural networks capture and are capable to structure complex patterns in visual data. Following this, this chapter explores centralised representation learning and unsupervised learning, especially self-supervised learning algorithms.

The theoretical explanation and literature review follow the shifts to federated learning, a distributed paradigm that enables collaborative model training across multiple clients while preserving data in each local client. This section examines the key challenges of data heterogeneity, non-i.i.d. distributions, which are significantly critical in federated systems settings where client data vary widely in quality and distribution.

In addition, metric learning algorithms are reconsidered as a method of improving representation robustness through structured embedding spaces. By integrating these insights from the research fields, this chapter identifies the existing limitations of current approaches and motivates the development of the proposed methods.

2.1 Representation Learning

Representation learning lies at the core of modern deep learning and refers to the process of *directly extracting robust representations (also called as embeddings) or features from raw data*, which is mentioned in these works; Goodfellow et al., 2016, Bengio et al., 2013. Representation learning aims to transform input data $x \in \mathcal{X}$ into a vector of features $f_\theta(x) \in \mathcal{Z}$ that capture the underlying descriptive aspects of variation in the data. These representations serve as the basis for downstream tasks such as image classification, object detection, or image generation. Unlike traditional machine learning frameworks, which mostly depend on handcrafted features, deep representation learning allows deep neural networks to learn directly hierarchical embeddings from raw data through multiple layers with nonlinear transformations.

Theoretically, a deep neural network model defines a parametric mapping as a function $f(\cdot)$

$$f_\theta : \mathcal{X} \to \mathcal{Z}, \tag{2.1}$$

where $\mathcal{X}$ denotes the input space, $\mathcal{Z}$ the latent feature space, and θ the learnable model parameters. Given an input $x \in \mathcal{X}$, the network model produces a representation $z = f_\theta(x)$ that generally captures the most informative and invariant characteristics of x with respect to the target task. Each layer in a deep network model performs a nonlinear transformation ϕ_l, L is the total number of layers in the network model, and each f_l is the function performed by the l^{th} layer, leading to a hierarchy of embeddings:

$$f_\theta(x) = \phi_L(\phi_{L-1}(\dots\phi_1(x))), \tag{2.2}$$

where lower layers extract local or structural patterns such as edges, corners and textures (low-level features), while deeper layers encode higher-level semantic concepts such as object parts or categories (Goodfellow et al., 2016). This hierarchical composition allows network models to learn representations that are invariant to nuisance transformations (e.g., scale, illumination, or viewpoint) and robust across varying data distributions.

In supervised learning framework settings, the network model parameters θ are optimised over labelled data:

$$\min_\theta \ \mathbb{E}_{(x,y)\sim\mathcal{D}} \ \mathcal{L}(f_\theta(x), y), \tag{2.3}$$

where $\mathcal{L}$ is a task-specific loss (e.g., cross-entropy) and $\mathcal{D}$ denotes the data distribution, y represents label for input x. The parameters are typically updated via stochastic gradient descent:

$$\theta \leftarrow \theta - \eta\nabla_\theta\mathcal{L}(f_\theta(x), y), \tag{2.4}$$

where η is the learning rate controlling the update magnitude.

A robust representation should benefit several desirable properties: (i) *invariance* to irrelevant transformations, (ii) *disentanglement* of independent factors of variation, and (iii) *sufficiency* for tasks such as generalisation. From a theoretic perspective, this can be formalised through the *Information Bottleneck* (IB) principle, which is shown in the paper; Kawaguchi et al., 2023, seeks representations Z that maintain optimal information about the target variable Y while eliminating irrelevant details from the input X:

$$\max I(Z;Y) \quad \text{subject to} \quad I(Z;X) \le \kappa, \tag{2.5}$$

where $I(\cdot;\cdot)$ represents mutual information and κ controls the trade-off between compression (e.g. embedding vector) and relevance (label-embedding relation). This formulation highlights the balance between preserving predictive information of data and eliminating unnecessary details.

This paper Goodfellow et al., 2016 indicates the importance of *shallow* and *deep* represen-

tations: while shallow network models capture mostly limited low-level patterns or features. On the other hand, deeper network architectures are capable to learn semanticly reletaed, high-level features that generalise across many tasks and domains. The capacity to learn hierarchical, transferable, and robust representations is what makes deep learning models particularly powerful for perception, reasoning, and decision-making tasks.

In self-supervised and unsupervised contexts, where labels are unavailable or missing, representation learning relies on auxiliary objectives, such as reconstruction-based methods (e.g., autoencoders, GANs) or similarity-based contrastive losses (e.g., NT-Xent loss) to capture features directly from unlabelled data. These approaches are especially valuable in domains where labelled data are not annotated or expensive to obtain.

In summary, this chapter present that representation learning literature review, background includes both the theoretical and practical foundations of deep learning. The chapter enables deep network models to extract hierarchical, abstract, and transferable features that support model generalisation, robustness, and scalability. Understanding these principles is crucial for developing efficient representation learning frameworks capable of handling limited, unlabelled, and heterogeneous data conditions that motivate the federated and self-supervised representation learning strategies proposed in this research.

2.2 Self-Supervised Representation Learning

Self-supervised learning (SSL) has emerged as a powerful paradigm for representation learning without requiring manually annotated labels. In contrast to traditional supervised learning approaches, where the model is trained on labelled image pairs (x, y) drawn from a distribution $\mathcal{D}$. Self-supervised representation learning leverages fundamental patterns within the data itself to define supervisory signals. The main goal is to learn feature representations that capture semantic and invariant properties of the input data distribution $p(x)$, thus enabling the transfer of learned features to downstream tasks. The key difference of SSL is that these learned features can be directly transferred or fine-tuned with specific datasets for supervised or unsupervised tasks.

Self-Supervised Learning Framework

Self-supervised learning (SSL) typically lies in two main stages: (1) a *pretext task training stage*, where the network model learns fundamental representations from unlabelled data, and (2) a *downstream task*, where the learned representations are adapted for specific supervised or unsupervised tasks with transfer learning.

Step 1: Pretext Task Training

Figure 2.1 shows the pretext task stage, where the model is trained on an auxiliary self-supervised objective, such as contrastive, predictive, or reconstruction-based learning to extract

invariant representations from unlabelled data. The network model learns to capture semantic relationships by solving assigned tasks without labels.

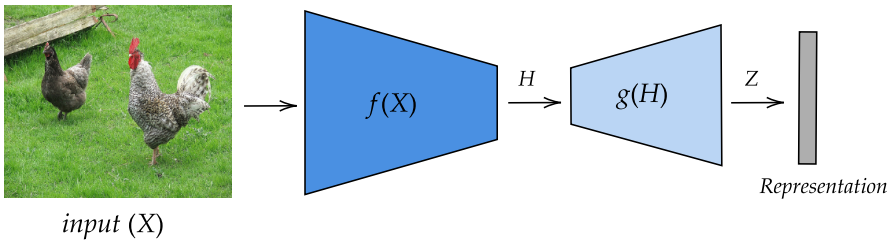


Figure 2.1: Stage 1: The network model learns to extract invariant embeddings from unlabelled data using a self-supervised objective function (e.g., contrastive or reconstruction loss).

Step 2: Downstream Task Training Figure 2.1 demonstrates the second stage, where the pretrained encoder from Step 1 is transferred and fine-tuned with a specific dataset on a downstream task, such as image classification or segmentation. The previously learned representations provide a strong initialisation, low-level embeddings, allowing the network model to achieve fast convergence, better generalisation with limited labelled data.

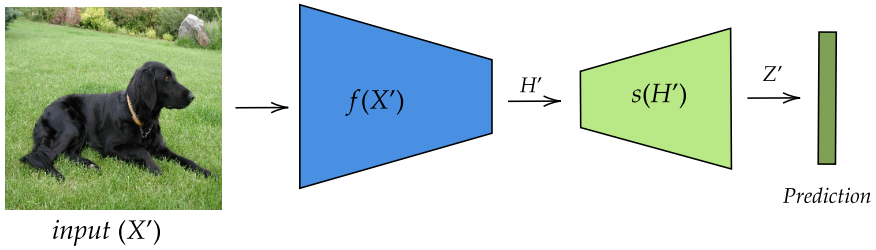


Figure 2.2: Stage 2: The pretrained encoder is fine-tuned for a downstream task using its own data, leveraging the representations learned during the pretext task.

The theoretical foundation of self-supervised representation learning lies in the direction that low-level, fundamental representations can be learned by solving *pretext tasks*, auxiliary objectives extract these representations from raw image data. These pretext tasks are designed to understand the underlying basic patterns of the input data, implicitly leveraging these representations into the new network model (the downstream task's model) to provide more robust features for the main task. Most common examples for pretest tasks include predicting missing parts of an image, like in this work Pathak et al., 2016, partitioning the representation space appropriately based on similarity of augmented views of input images like SimCLR, or reconstructing inputs in the paper Vincent et al., 2010.

Theoretically, $\{x_i\}_{i=1}^{N}$ represents an unlabelled dataset, the main objective of SSL is to learn an encoder $f_\theta : \mathcal{X} \to \mathcal{Z}$ that maps inputs to a latent representation space $\mathcal{Z}$. The learned

representation $z_i = f_\theta(x_i)$ should maintain the most accurate, invariant, and transferable features of the data distribution $p(x)$.

The prior work (T. Chen, Kornblith, Norouzi, and Hinton, 2020) shows that SSL can be interpreted as maximising the mutual information between related views of the same data samples while minimising the distance of irrelevant or dissimilar samples. Let x represent a data point and x' an augmented view of the data point (e.g., obtained via random data transformation $t \sim \mathcal{T}$). The encoder f_θ is trained to maximise:

$$I(Z; Z') = I(f_\theta(x); f_\theta(x')), \tag{2.6}$$

subject to invariance constraints. This formulation aligns with the Information Bottleneck (IB) principle Kawaguchi et al., 2023, where the representation Z preserves information relevant to recognising x' (the positive sample) while discarding irrelevant variations (negative pairs).

In the real world, exacting mutual information from samples is difficult to compute, thus SSL methods approximate the mutual information by using contrastive or redundancy-reduction losses like Sohn, 2016. The typical formula of a contrastive objective in SSL is:

$$\mathcal{L}_{\text{contrastive}} = -\mathbb{E}\left[\log \frac{\exp(\text{sim}(z_i, z_i^+)/\tau)}{\sum_{j=1}^{N} \exp(\text{sim}(z_i, z_j^-)/\tau)}\right], \tag{2.7}$$

where $\text{sim}(\cdot,\cdot)$ is a similarity measurement (e.g., cosine similarity), τ is a temperature scaling factor, z_i^+ is a positive (related or similar) view, and z_j^- are negative (unrelated or dissimilar) views. This function is standardized through contrastive objectives such as the *InfoNCE* loss van den Oord et al., 2018, which can be interpreted as a lower bound on the mutual information $I(Z; Z^+)$ between a representation Z and its positive (correlated) pair Z^+. By maximising this bound, the encoder $f_\theta(\cdot)$ is encouraged to preserve shared similar information between positive pairs while throwing away noise and inconsistent correlations of negative pairs.

Normalised Temperature-scaled Cross-Entropy (NT-Xent) Loss

A widely popular formula of the InfoNCE objective is the NT-Xent loss, which is introduced in the SimCLR T. Chen, Kornblith, Norouzi, and Hinton, 2020 paper. Given an image sample x, two random augmentations $\mathcal{T}_1$ and $\mathcal{T}_2$ are applied to produce two correlated views:

$$x_i = \mathcal{T}_1(x), \qquad x_j = \mathcal{T}_2(x), \tag{2.8}$$

Then, each augmented view is passed through an encoder network model $f(\cdot)$ to obtain the corresponding embeddings:

$$z_i = f(x_i), \qquad z_j = f(x_j). \tag{2.9}$$

The objective is to maximise the agreement between the embeddings of positive pairs (similar views) (z_i, z_j) transformed from the same image sample, while minimising mutual information

between embeddings of other augmented views (treated as negatives) in the batch.

Given a batch of N samples, each transformed into two views, the calculation of loss for one positive pair (i, j) is defined as:

$$\mathcal{L}_{i,j} = -\log \frac{\exp(\text{sim}(z_i, z_j)/\tau)}{\sum_{k=1,k\neq i}^{2N} \exp(\text{sim}(z_i, z_k)/\tau)}, \tag{2.10}$$

where $\text{sim}(z_i, z_j)$ represents the cosine similarity between two embeddings to measure directional similarity,

$$\text{sim}(z_i, z_j) = \frac{z_i^\top z_j}{\|z_i\| \, \|z_j\|}. \tag{2.11}$$

and $\tau > 0$ is a temperature parameter that controls the sharpness of the similarity distribution. The overall objective is calculated as the mean over positive pairs in the batch, considering the symmetry between each pair of augmented views:

$$\mathcal{L}_{\text{NT-Xent}} = \frac{1}{2N} \sum_{i=1}^{N} \left(\mathcal{L}_{i,j(i)} + \mathcal{L}_{j(i),i}\right). \tag{2.12}$$

The numerator in Eq. (2.10) increases the similarity between the positive pair (z_i, z_j), while the denominator penalises that if there is any similarity with other negative samples within the batch. Therefore, this objective function performs *instance discrimination* (Z. Wu et al., 2018), where each instance is treated as a unique class. Thus, the embeddings encode features that are invariant to data augmentations while being discriminative across individual samples, effectively learning semantically correct and transferable representations.

2.2.1 Invariance, Disentanglement, and Compression

The main theoretical assumption in self-supervised representation learning is that each observation $x \in \mathcal{X}$ is generated by a set of latent data transformations $\mathbf{v} = (v_1, v_2, \ldots, v_k)$, which describe the underlying patterns of the data manifold. The goal of self-supervised learning (SSL) is to learn a representation $z = f_\theta(x)$ that captures the invariant aspects of these augmented views while discarding non-informative variations that are irrelevant to downstream tasks, which explain with these papers (Bengio et al., 2013, Cohen and Welling, 2019).

Within this context, invariance refers to the properties that the representation remains unchanged under transformations. Therefore, these representations preserve semantic identity, whereas sensitivity (or equivariance, transformation-consistent) ensures that the representation responds efficiently to changes in several attributes of the input. By learning the representations, SSL aims to approximate the underlying abundant factors of the data without requiring labels. Thus, this perspective is grounded in the information-theoretic view of representation learning, where the optimal representation z is expected to maximise the mutual information with the

relevant latent factors (latent space embeddings) while minimising its dependence on irrelevant noise or transformations (dissimilar embeddings). This representation alignment is mentioned these (Kawaguchi et al., 2023, Bachman et al., 2019). This is strongly linked to the principles of *invariance* and *equivariance*:

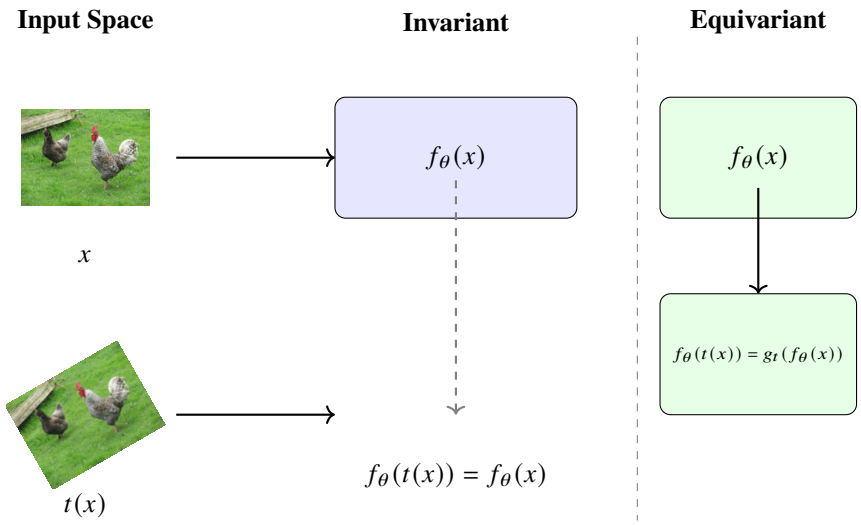


Figure 2.3: Illustration of invariance and equivariance. Invariance indicates that the learned representation preserves semantic relation and transformation-invariance, i.e., $f_\theta(t(x)) = f_\theta(x)$. Equivariance means that transformations in input space contribute to predictable transformations in latent feature space, i.e., $f_\theta(t(x)) = g_t(f_\theta(x))$.

- **Invariance:** $f_\theta(x) = f_\theta(t(x))$ for transformation function $t \in \mathcal{T}$ that do not affect semantic content of images. In figure 2.3 shows the original image and the augmented view of the image.
- **Equivariance:** for certain tasks, $f_\theta(t(x))$ transforms predictably with t, i.e., $f_\theta(t(x)) = g_t(f_\theta(x))$ for some transformation g_t in latent representation space.

To sum up, these principles encourage the network model to extract stable, generalisable, robust features that correspond to fundamental latent factors of variation, rather than shallow pixel-level visual data details.

According to the Information Bottleneck interpretation mentioned in the work Kawaguchi et al., 2023, effective representations are capable of maximising mutual information (statistical dependency) $I(Z;Y)$ with the target variable (or a task) while minimising $I(X;Z)$, thereby throwing away irrelevant information. From this perspective, Self-supervised methods go further than the bottleneck interpretation work by leveraging unlabelled data. Therefore, limited labelled data is not a bottleneck in this domain.

2.2.2 Connections to Typical Representation Learning Theory

Self-supervised representation learning can be linked to classical principles of statistical representational learning theory. Traditional supervised learning methods mostly address the labelled data-based task, while SSL seeks to minimise an *unsupervised generalisation bound* through pretext tasks that constrain the space of f_θ. Therefore, supervised generalisation limits in labelled data space, and it has poor performance on many unseen labelled data. Unlike many supervised methods, SSL is capable of expanding the bounds of generalisation with unlabelled data.

Thus, this can be interpreted as imposing a pattern recognition prior that biases the network model toward fundamental learning features aligned with typical data geometry (Balestriero et al., 2023). By learning features to solve various pretext tasks, the network model effectively regularises its parameter space, resulting in sufficient sample usage and better generalisation under limited or heterogeneous data conditions.

2.2.3 Representative SSL Methods

In recent years, several SSL frameworks have been proposed to address these limitations:

SimCLR. The prior work in SSL T. Chen, Kornblith, Norouzi, and Hinton, 2020 presents the SimCLR framework which is shown in the figure 2.4 and provides a simple but powerful implementation of contrastive representation learning. SimCLR relies on large batch sizes to provide a diverse set of negatives/positives and employs plenty of random data augmentation (e.g., random crops, colour jitter, Gaussian blur) to generate many pairs. In their assumption, large batch sizes are capable of addressing misleading, wrong or weak augmented views. In the design of the framework, a two-layer simple projection head is added to the backbone network to map features into a contrastive latent space, where the NT-Xent loss (Eq. 2.10) calculates the cost, error between embeddings. SimCLR demonstrates that sufficient, various data augmentation and many batches are beneficial for contrastive learning instead of fully supervised training.

MoCo (Momentum Contrast). In this work K. He et al., 2020 introduces two key innovations over SimCLR: a *momentum-updated encoder "MoCo"* and a *dynamic memory queue* for maintaining a large and consistent set of negative examples in the figure 2.5. While SimCLR totally relies on in-batch negatives, requiring significantly large batch sizes (e.g., 4096, 8192) to achieve high accuracy, MoCo decouples the number of negative samples from the batch size. Thus, dynamic memory maintains a *first-in, first-out (FIFO)* principle, the queue of encoded embeddings from previous mini-batches, which serves as a frequently updated embeddings dictionary of negative examples.

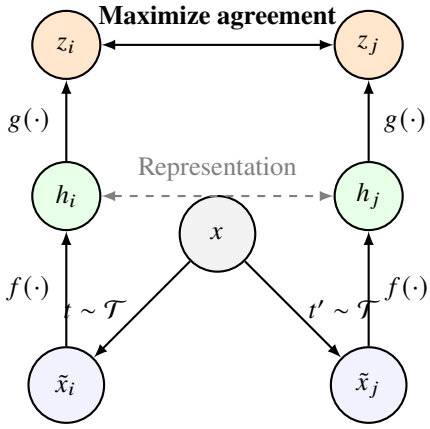


Figure 2.4: A simple framework T. Chen, Kornblith, Norouzi, and Hinton, 2020 for contrastive learning of visual representations. In here, two data augmentation operators are applied to each image sample x to obtain two transformed views $\tilde{x}_i$ and $\tilde{x}_j$ of one image. A shared base encoder $f(\cdot)$ and a small projection head $g(\cdot)$ are trained to maximise agreement between their projected embeddings (z_i, z_j) using a contrastive loss. After training, the projection head $g(\cdot)$ is eliminated, and the encoder $f(\cdot)$ with representation h is used for downstream tasks.

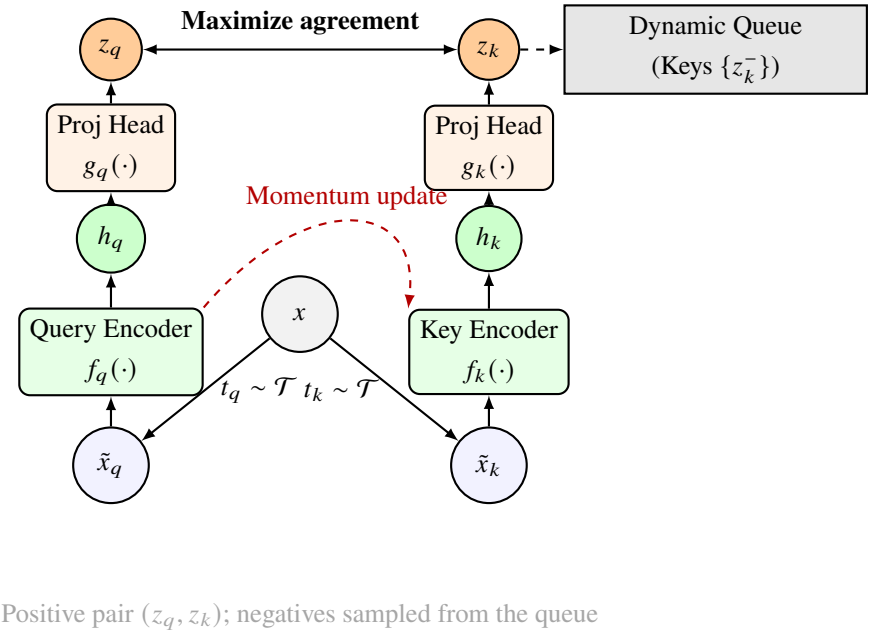


Figure 2.5: Overview of the (MoCo) K. He et al., 2020 framework. Two random augmentations $\mathcal{T}_q$ and $\mathcal{T}_k$ are applied to the same input image x to produce augmented views $(\tilde{x}_q, \tilde{x}_k)$. The query encoder f_q is updated by gradient descent, while the key encoder f_k is updated via a momentum mechanism from f_q. Representations (z_q, z_k) form positive pairs, and a dynamic queue provides a large set of negative keys for contrastive learning. The model is trained to maximise agreement between positive pairs while distinguishing them from negative samples in the queue.

BYOL (Bootstrap Your Own Latent). This approach (Grill et al., 2020) eliminates the need for negatives by training an online network to predict the representation of a momentum-updated target network. The method uses two asymmetric network models (models are not the same):

an *online network* with a predictor and a *target network* updated via exponential moving average (EMA). Despite lacking negatives, BYOL prevents representational collapse through architectural model asymmetry and momentum updates.

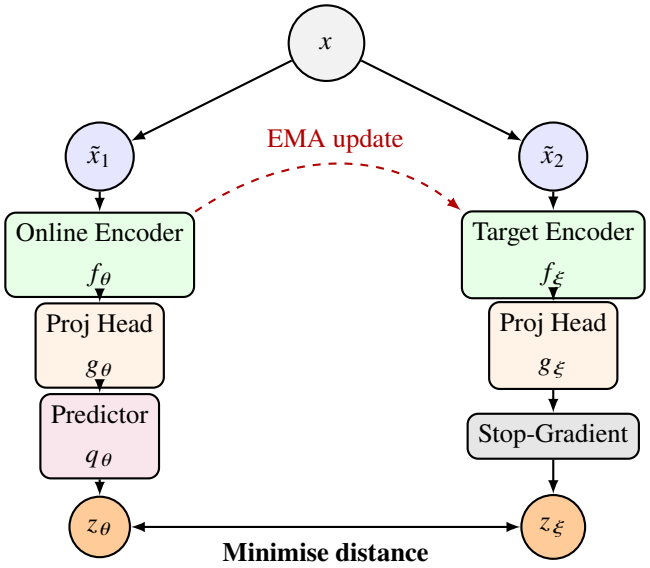


Figure 2.6: Overview of the BYOL framework Grill et al., 2020. Two augmented views $(\tilde{x}_1, \tilde{x}_2)$ of the same image are passed through an *online network* and a *target network*, each composed of an encoder f, a projection head g, and a predictor q (online only). The online network learns to predict the target network's representation, while the target parameters are updated using an exponential moving average (EMA) of the online weights. No negative samples are used; training stability is achieved through the stop-gradient operation on the target branch.

DINO (Self-Distillation with No Labels). DINO (Caron et al., 2021) extends the self-distillation paradigm by training a student network to match the output distribution of a teacher network. Thus, the teacher parameters are updated as an exponential moving average of the student's model parameters. DINO treats different augmented views as positive pairs and aligns their output distributions using a cross-entropy objective over non-normalised features. Therefore, this approach captures semantic patterns and yields powerful visual representations without requiring negative samples. These approaches clearly show that negatives play a critical role in representation learning.

2.2.4 Recent Advances in Self-supervised Representation Learning

While pioneer contrastive learning methods such as SimCLR, MoCo, BYOL and DINO established the fundamental paradigm of maximising agreement between positive views and avoiding negative transformed views of image samples. More recent work has extended, refined and

examined these representation learning concepts in several important ways. In this section, we highlight a few practical directions.

No-Negative and redundancy-reduction approaches. Self-supervised learning (SSL) has increasingly moved away from the traditional focus on straightforward negative pairs and large batch sizes. Alternatively, recent approaches underline redundancy reduction, invariance across augmentations, and architectural asymmetry to prevent representational collapse.

Some methods in this area, such as Barlow Twins(Zbontar et al., 2021) and VICReg (Bardes et al., 2022), introduced two objectives that enforce augmentation-invariance while penalising dimension-wise correlations or feature collapse. These frameworks showed that robust representations can be learned by shaping the statistical structure of embeddings rather than contrasting them against negatives. Like SimSiam (X. Chen and He, 2021), it demonstrates that feature collapse can be avoided purely through architectural model asymmetry and a stop-gradient operation, without momentum encoders or negative pairs.

A series of following works expanded this principle by replacing it with whitening, alignment, or prediction-based constraints. W-MSE (Ermolov et al., 2021) has leveraged whitening image transformations and direct feature matching, respectively, to provide stable training while removing dependence on batch size. Variance-aware predictive models like the influential I-JEPA Assran et al., 2023 shifted SSL toward masked prediction frameworks that learn global semantic patterns by predicting latent representations instead of pixel-level reconstructions.

More recent large-scale methods continue to refine this paradigm. DINOv2 (Oquab et al., 2024) has improved teacher–student distillation, feature regularisation, and large-scale iamge augmentation strategies, becoming a dominant network model for vision tasks without relying on any particular contrastive losses. Hybrid alignment–variance approaches, including UniGrad (Tao et al., 2022), integrate the contrastive and non-contrastive areas by implicitly preserving variance while aligning embeddings through learned soft negatives or gradient-based constraints.

In general, these achievements illustrate a clear trend: modern self-supervised vision models increasingly rely on predictive asymmetry, variance preservation, feature decorrelation, and teacher–student network model training as primary mechanisms. The main aims of those approaches mostly prevent feature collapse, enabling scalable and highly accurate network models, robust representation learning without straightforward negative sampling.

Multimodal contrastive learning and large-scale self-distillation. The state-of-the-art, CLIP (Radford et al., 2021) and ALIGN (C. Jia et al., 2021) significantly enlarged the scope of contrastive representation learning by extending it to multimodal settings, where image–text pairs are aligned in a shared embedding space. Thus, trained on hundreds of millions of noisy web-scale image–caption pairs, these models exhibit highly accurate zero-shot generalisation, sparking a shift toward large-scale, weakly supervised pretraining as an alternative to traditional

supervised pipelines. After implementing multimodal methods including Florence (L. Yuan et al., 2021), CoCa (Yu et al., 2022), and MetaCLIP (Y.-S. Chuang et al., 2025), refined dataset curation, alignment objectives, and transformer model architectures to further enhance cross-modal robustness and transferability.

Following the developments of large-scale self-distillation or transformers, DINOv2 scaled self-supervised distillation to massive curated datasets and incorporated improved data pipelines, regularisation, and embedding stabilisation, producing robust models with exceptional performance across classification, retrieval, and various prediction tasks. Recently, DINOv3 (Siméoni et al., 2025) has been presented as a new network model. The network model relies on data scale, integrating masked prediction, improved positional encoding, and architectural refinements to achieve state-of-the-art transfer, especially in high-resolution settings.

On the whole, these multimodal contrastive and large-scale self-distillation frameworks highlight a major trend: scaling data and models, combined with cross-modal alignment or predictive teacher–student objectives, yields highly generalisable visual representations that increasingly serve as universal backbones for downstream computer vision tasks. The entire works clearly indicate that the quality of learned features or representation directly affects the accuracy of the network models or model generalisation.

2.3 Federated Learning

Federated Learning (FL) enables collaborative deep network model training across multiple clients without directly sharing the raw data of each client. The pioneer paper McMahan et al., 2017 presents FedAvg, distributed representation learning data preserving in many local places. Following FedAvg, another federated model aggregation method Sharma et al., 2022 is proposed to handle the global model aggregation problem. The key difference of federated learning is that each client (e.g., mobile device, institution, or organisation) owns its local dataset and contributes to the global model update through communication of model parameters, weights or gradients. Thus, this distributed paradigm preserves data privacy (as no access to any local data) and security (the global model cannot use data directly) while enabling large-scale knowledge sharing across heterogeneous data sources.

2.3.1 Federated Learning Objective

The primary objective of Federated Learning (FL) is to collaboratively optimise a shared global model across a set of clients without directly exchanging their raw data. In general, FL aims to minimise a global model loss function that aggregates each local model's weights of participating clients:

$$\min_{w} \; F(w) = \sum_{k=1}^{K} p_k \, F_k(w), \tag{2.13}$$

where K represents the total number of clients, $w \in \mathbb{R}^d$ defines d dimesional global model parameters, and $p_k = \frac{n_k}{\sum_{i=1}^{K} n_i}$ is weight of client k determined by its local dataset size n_k. Each client's local objective $F_k(w)$ is defined as:

$$F_k(w) = \frac{1}{n_k} \sum_{i=1}^{n_k} \ell(w; x_i^{(k)}, y_i^{(k)}), \tag{2.14}$$

where $\ell(\cdot)$ denotes the loss function evaluated on local samples $(x_i^{(k)}, y_i^{(k)})$ drawn from the client's data distribution $\mathcal{D}_k$.

The optimisation process is typically performed in a set of communication rounds between a central server (global model) and the distributed clients. At the beginning of each round t, the server broadcasts the global model w_t to a subset of selected clients $\mathcal{S}_t \subseteq \{1, \ldots, K\}$ (broadcasting). Then, each participating client performs several steps (epochs) of local stochastic gradient descent (SGD) on its private data for an updated model $w_t^{(k)}$. Following that, these local updates (mostly local model parameters) are transmitted to the central server, which aggregates them to optimise the global model parameters:

$$w_{t+1} = \sum_{k \in \mathcal{S}_t} p_k \, w_t^{(k)}. \tag{2.15}$$

When we check the literature pioneer works, this aggregation rule, popularised by FedAvg (McMahan et al., 2017), performs as the cornerstone of most federated optimisation algorithms.

Despite this model aggregation's simplicity, the FedAvg optimisation struggles several critical challenges. The most well-known challenge is *data heterogeneity*, the non-identical and independently distributed (non-i.i.d.) nature of each client datasets $\{\mathcal{D}_k\}_{k=1}^{K}$. When local data distributions significantly differ across clients, local network models reach a desirable optimum point; however, model aggregation suffers from the model aggregation since the global model does not reach a global minimum. The limitation causes many problems, such as slow convergence, longer communication rounds, negative impact on learned local features. To address these limitations, various model aggregation strategies have been presented, such as SCAFFOLD (Karimireddy et al., 2020), MOON (Q. Li et al., 2021), FEDDC (Gao et al., 2022). Therefore, developing several model aggregation methods is quite popular in the federated learning field. In addition, recent research has mostly focused on developing algorithms that improve representation alignment, decrease client model shift, and present global model consistency across heterogeneous datasets through regularization, model contrastive learning, and representation-matching mechanisms. These research directions are the fundamental parts of the thesis, and mainly focus on developing algorithms to handle these limitations by representation learning approaches.

2.3.2 Federated Averaging (FedAvg)

Federated Averaging (FedAvg) is a well-known, pioneering model aggregation algorithm in FL. FedAvg is divided into two key steps:

1. **Local Update:** Each participating client initializes its local network model with the global network model parameters w_t and performs several epochs of stochastic gradient descent (SGD) on its local data, obtaining $w_t^{(k)}$ (new, learned parameters).

2. **Global Aggregation:** The central server collects each participating local network model updates (weights) and computes new weights based on average weight calculation, like in the equation 2.15.

In short, FedAvg is a simple model aggregation algorithm to implement, but a powerful method to achieve scalable, distributed learning without data collection on a single server. However, the main obstacle here is that FedAvg assumes each client has the same data distribution.

2.3.3 Impact of Data Heterogeneity and Imbalance in Federated Learning

In federated learning (FL), participating clients typically have data that are **non-independent and identically distributed (non-i.i.d.)** and frequently **imbalanced** in both class and quantity. Unlike centralised training, where the model is optimised on a single server, mixed datasets, federated frameworks operate on distributed local datasets that may vary significantly in statistical heterogeneity, scale, and feature distribution. Although many model aggregation methods present novel algorithms to set the global network model parameters, they cannot address the main problem, data heterogeneity.

2.3.4 Categories of Data Heterogeneity

In the real world, data distributions are mostly different. Since the global model cannot access the local datasets, the difference between local data distributions leads to a data heterogeneity problem. Therefore, data heterogeneity is one of the major challenges in Federated Learning (FL), as local datasets commonly differ across clients (Kairouz et al., 2021), (Jimenez-Gutierrez et al., 2025). In a standard FL framework setup, the central server dataset is partitioned across K clients, where each client k maintains a private local dataset

$$\mathcal{D}_k = \{(x_i^{(k)}, y_i^{(k)})\}_{i=1}^{n_k}. \tag{2.16}$$

Under perfect conditions, each local data sample is independently and identically distributed with a joint distribution $P(x, y)$:

$$(x_i^{(k)}, y_i^{(k)}) \sim P(x, y), \quad \forall k \in \{1, \ldots, K\}, \tag{2.17}$$

The equation shows an ideal condition as an i.i.d. learning environment. However, many real-world federated systems rarely supports this assumption. As an alternative, each client has own local data distribution $P_k(x, y)$, shaped by factors such as class imbalance, various source environments, several device usages, or institutional aspects. Therefore,

$$P_k(x, y) \neq P_{k'}(x, y) \quad \text{for some } k \neq k', \tag{2.18}$$

leading to a fundamentally **non-i.i.d. learning scenario**. This distributional difference can arise several versions of data heterogeneity, and we mainly categorise as follows:

- **Feature distribution skew (covariate shift):** Clients have diverse, distinct input datasets. For example, two client version shows in the following equation that their probability data distributions are not same or similar:

  $$P_k(x) \neq P_{k'}(x). \tag{2.19}$$

  Thus, heterogeneous distributed clients follow diverse data variation distributions. This occurs when clients operate in distinct environments such as limited computation, low bandwidth connection network. In a different scenario, each client has specific dataset contains different labels, image samples collected from several sensors that cause illumination, scale, shape. For instance, in a medical imaging based federated frameworks, many hospitals commonly use different imaging devices (MRI, X-Ray) or these device settings. Also, in these hospitals, collected data depends on the patents clinical records, age, gender. Thus, these challenges lead to data heterogeneity problem based on their feature distribution skew or covariate shift among clients.

- **Label distribution skew:** In real world, mostly client datasets contain different portion of class labels:

  $$P_k(y) \neq P_{k'}(y). \tag{2.20}$$

  Label distribution skew is one of the most common variants of data heterogeneity in FL. Since clients have samples from only a subset of classes (or sometimes a single class), leading to label imbalance. Thus, some clients can only access a few labels of classes. For example, some clients train on samples from cat and dog classes. some clients can train on apple and orange samples. In addition, class imbalace causes label skewness. Even each client has same number of class dataset, the quantity of each labels may differ. These differences are major obstacles of federated model aggregation and finding global minimum for the global network model.

- **Conditional distribution skew:** When clients have similar label proportions, their class-

conditional feature probability distributions may differ:

$$P_k(x \mid y) \neq P_{k'}(x \mid y). \tag{2.21}$$

For example, images corresponding to the same class may vary across institutions due to geographic differences, weather condition, annotation biases, or image storing conditions. Therefore, this mismatch poses several challenges for decision boundaries in representation space across clients.

To sum up, these heterogeneous data variations individually harm the key assumptions of traditional distributed optimisation and can lead to client data shift, unstable training behaviour, slow convergence and limited global model generalisation. In this respect, addressing non-i.i.d. data scenarios still remains a core research problem in federated learning. Thus, this thesis mainly aims to present novel methods to address these obstacles in various federated settings.

2.3.5 Dirichlet Distribution for Simulating Data Heterogeneity

In order to evaluate federated learning algorithms under various levels of data heterogeneity, researchers commonly use the **Dirichlet distribution** to partition datasets across many clients (X. Li et al., 2023), (Oh et al., 2021). The Dirichlet distribution provides a simple but effective and flexible way to control the degree of (non-i.i.d.)ness through a coefficient α.

Theoretically, let $\mathbf{p}_c = [p_{c,1}, p_{c,2}, \ldots, p_{c,K}]$ represent the distribution of class c across K clients, where $p_{c,k}$ denotes the proportion of samples of each class c assigned to client k. For each class $c \in \{1, \ldots, C\}$, the distribution is:

$$\mathbf{p}_c \sim \text{Dirichlet}_K(\alpha) \tag{2.22}$$

where $\text{Dirichlet}_K(\alpha)$ is a K-dimensional Dirichlet distribution parameterized by the value $\alpha > 0$. Thus, each client k receives a fraction $p_{c,k}$ of the data belonging to class c. The total number of dataset of client k is:

$$\mathcal{D}_k = \bigcup_{c=1}^{C} \mathcal{D}_c^{(k)}, \quad \text{where} \quad |\mathcal{D}_c^{(k)}| = p_{c,k} \cdot |\mathcal{D}_c|. \tag{2.23}$$

This Dirichlet distribution cannot guarantee the total size of the training data. The typical Dirichlet distribution simply results in class imbalance among clients.

2.3.6 Effect of the Parameter

The α parameter play critical role in data settings and controls how uniformly data are distributed among clients:

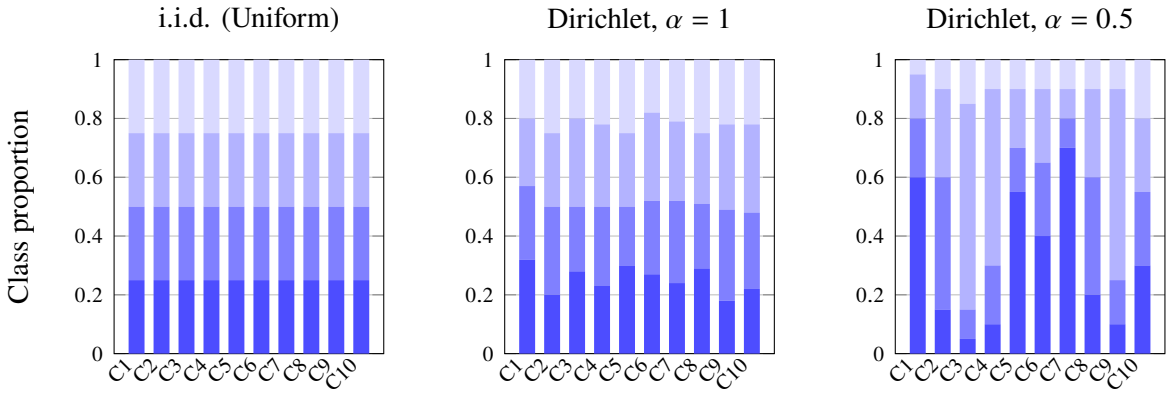


Figure 2.7: Client distributions for i.i.d., Dirichlet $\alpha = 1$, and Dirichlet $\alpha = 0.5$. Smaller bars and increased spacing improve readability. Each class uses blue-shaded colour variations.

- **Higher** α values (e.g., $\alpha > 10$): Where α values are higher, proportions $\mathbf{p}_c$ are nearly uniform across clients. Thus, each client obtains an approximately equal sample size of each class. Mostly, this setting approximates an i.i.d. distribution or uniform.

- **Moderate** α (e.g., $\alpha \approx 1$): When the α parameter is around 1, the class proportions vary very slightly across clients, creating a realistic but moderately heterogeneous data scenario.

- **Lower** α (e.g., $\alpha < 0.5$): When experimenting with smaller α values like 0.3, the sampled proportions are highly skewed. Thus, each class is divided into a few clients, and many clients lack samples of specific classes. Therefore, this data distribution represents substantial statistical data heterogeneity (a.k.a. (non-i.i.d.)ness).

In a theoretical perspective, when α is a smaller value like 0.3, the Dirichlet distribution yields proportion vectors $\mathbf{p}_c$. Thus, these vectors exhibit a concentration of data distribution probability on a few components. As $\alpha \to 0$, each class is effectively localised to a single client, which represents an extremely sparse distribution.

Figure 2.7 illustrates the impact of different α values on the distribution of data among clients. The figure represents three different data distributions: uniform (i.i.d.), moderated data heterogenity and high heterogeneity data distribution across ten clients. Typically, the total training sample size may differ across clients; however, we illustrate the same training size in plots to show the difference between data homogeneity and heterogeneity among clients' datasets clearly. For higher α values like 1 or higher than 1, proportions are almost similar across clients, whereas smaller α values like 0.5 yield sparser and uneven data distributions. This change with the parameter makes the Dirichlet distribution notably suitable for simulating heterogeneous data distribution in many experimental studies.

As a summary, the Dirichlet distribution provides a reliable and adjustable data scenario for simulating non-i.i.d. or i.i.d. data across clients in federated learning experiments. By adjusting the α parameter, we simulate varying ranges of statistical data heterogeneity from nearly i.i.d.

(large α values) to extremely skewed (smaller α values) partitions. Therefore, this approach enables the definition of several data distributions and compares state-of-the-art approaches.

2.4 Limitations

2.4.1 Limitations of Existing Self-Supervised Learning Methods

Self-supervised learning (SSL) is a paradigm for high-quality representation learning without human annotations. Self-supervised contrastive methods such as SimCLR, MoCo, MoCo-v3 (X. Chen et al., 2021), followed by self-supervised non-contrastive approaches such as BYOL and SimSiam (X. Chen and He, 2021). Also, representation redundancy-reduction based methods such as Barlow Twins (Zbontar et al., 2021) and VICReg (Bardes et al., 2022), have demonstrated that self-supervised pre-training surpasses supervised learning baselines on a range of downstream tasks. In the computer vision area, these methods typically learn an encoder f_θ that maps an input image x into a representation $z = f_\theta(x)$, trained by enforcing invariances across augmentations of the same image. In many cases, discriminating between different images in a large batch or memory bank.

Despite these methods' success, existing SSL methods still have important limitations when deployed beyond the standard centralised, large-scale, and curated benchmarking setups (e.g., ImageNet). Thus, these limitations concern their reliance on data augmentations related to data type, data sensitivity to optimisation and data heterogeneity. In addition to these limitations, computational costs, hyperparameter tuning and architectural network model choices. This section reviews key limitations of the SSL literature and highlights the main challenges that motivate more robust, distribution-aware representation learning methods.

Reliance on Augmentations and Handcrafted Invariances

Most modern SSL methods are built around the instance discrimination paradigm: two augmented views of the same image must be mapped to similar representations, while views from different images should be pushed apart. Thus, SimCLR shows that the selection of data augmentations (e.g., random crops, colour jitter, blur) is critical for the network model performance; inadequate data augmentation pipelines lead to trivial and insignificant views or inadequate generalisation. Likely, MoCo relies on data augmentations combined with a momentum-updated encoder and a large memory queue of negative examples.

Therefore, this reliance on handcrafted invariances introduces several issues:

- Data augmentations explicitly encode assumptions about which transformations preserve the semantic content of image samples. When these assumptions are violated (e.g., in remote sensing or biometric images), the learned invariances can be inappropriate or even harmful.

- SSL performance is highly sensitive to data augmentation strategy, the accuracy scores mostly requiring dataset-specific tuning. Thus, the decision of data augmentations limits robustness and reproducibility.
- In heterogeneous settings, different clients may require various data augmentation strategies, which current SSL frameworks perform poorly.

As a result, existing SSL frameworks frequently lack mechanisms for adapting data augmentation on specific datasets or learning invariant representations under complex, multi-domain conditions.

Limitations of Contrastive Objectives and Negative Sampling

Contrastive representation learning methods such as SimCLR and MoCo optimise InfoNCE-like losses that require a large number of negative samples to avoid misleading positive pairs and increasing mutual information (T. Chen, Kornblith, Norouzi, and Hinton, 2020), (K. He et al., 2020). While effective on centralised datasets with large batch sizes or memory queues, these methods demonstrate several limitations:

- **Negative sampling bias:** Negatives are typically assumed to be semantically irrelevant; however, in practice, many negative pairs may share similar class or common characteristics with the anchor. Thus, it leads to the *false negatives* problem that harms learning the semantic patterns and degrades downstream task performance.
- **Batch size and memory dependence:** Like SimCLR, many contrastive learning methods require unnecessarily large batches to perform well, which has a bad impacts applicability to resource-limited environments. Also, requiring high-resolution image datasets limits the learning of robust representations on many small-scale datasets.
- **Sensitivity to temperature and margins:** The temperature parameter in the softmax and the relative weighting of positive versus negative terms significantly affect the geometry of the representation space, often requiring careful tuning.

In summary, these limitations have motivated non-contrastive methods such as BYOL and SimSiam by focusing on representation learning, which avoids explicit negatives. In addition to these methods, many redundancy-reduction frameworks, such as Barlow Twins and VICReg shows the quality of learned representations directly affects network models' generalisation and stability. However, these different research directions introduce their own challenges and limitations, which are discussed next section.

Risk of Collapse and Fragility of Non-Contrastive Methods

Self-supervised non-contrastive methods demonstrate that it is possible to learn representative features without negative pairs by designing asymmetric architectures, stop-gradient operations, and predictor networks. Likely, Barlow Twins and VICReg penalize representation redundancy and enforce variance constraints across feature dimensions.

Although these methods alleviate limitations of contrastive representation learning, they remain weak with the following obstacles:

- A brief explanation of why representational collapse is not avoided; small layer changes to architecture (e.g., removing the predicting parts), objective weighting, poor feature space separation or training on false positive or negative samples can lead to representation collapse.
- Many non-contrastive methods implicitly rely on architectural asymmetry or specific training tricks (e.g., momentum encoders, specified projection heads, batch normalisation). Thus, these changes make the separation of features harder. Also, adaptation or extension to new network models, unseen but similar datasets or modalities are extremely difficult.
- In imbalanced or highly complex datasets, the variance and covariance penalties may not be sufficient when data diversity is low. Thus, it results poor and biased embedding space, and spreads features.

Although self-supervised non-contrastive learning reduces the dependence on negatives, the learning paradigm induces high feature variance, representation collapse, and poor separation in space. Thus, the network models struggle in various data regimes, yielding suboptimal performance and reduced accuracy.

Scalability, Compute, and Data Assumptions

State-of-the-art SSL methods typically assume: (i) large, high-quality unlabeled datasets are available for any task, (ii) substantial computing resources (e.g. multi-GPU training, long schedules), and (iii) stable, centralised training pipelines. For example, SimCLR and MoCo were originally evaluated on large-scale ImageNet data, with hundreds of epochs and high computational cost (T. Chen, Kornblith, Norouzi, and Hinton, 2020), (K. He et al., 2020). In addition, more efficient and recent methods, such as DINOv3 or VICReg, still rely on substantial compute and memory space.

However, in the real world, these assumptions are rarely appropriate. Since data is mostly unlabelled, many models remain highly dependent on labelled samples. Storage of large datasets in one server is inefficient, and transferring the whole data to many other places raises data accessibility concerns. Processing large-scale datasets demands longer training epochs and more epochs to reach the global minimum or optimal performance. Moreover, the real world

is mostly unsupervised. Learning human annotations is not only a problem, but also, data is noisier, low resolution, diverse, and distributed across many devices. Existing self-supervised or unsupervised representation learning approaches suffer from these limitations. In addition, these models have limited adaptability under these settings.

Limitations under Heterogeneous and Federated Settings

Most SSL approaches rely on centralised training under the assumption of similar or identical data distribution. When extending SSL to federated heterogeneous scenarios, various additional limitations arise:

- **non-i.i.d. client data distributions:** Clients mostly have different domains or share the same dataset but diverse labels, in addition to variations in data augmentations, feature distributions, and invariances. SSL objectives assume that a single underlying data distribution may then encourage incompatible feature spaces across clients.
- **Limited cross-sample relationships:** Contrastive learning methods rely on large sets of negatives and diverse positives, which are available in centralised settings; however, large negatives are not possible by distributing the datasets among clients. Thus, this scarcity of cross-sample relationships can lead to a smaller batch size and potential representational collapse.
- **Aggregation-induced distortion:** Aggregating local network model parameters trained on heterogeneous datasets can distort the embedding space established during training locally.

Although recent studies investigate federated self-supervised or contrastive representation learning, they commonly adapt centralised methods to decentralised learning with minimal changes. Also, the lack of a comprehensive analysis of representation alignment and feature collapse under client heterogeneity is still an important research gap in the field.

2.4.2 Limitations of Existing Federated Learning Methods

Despite considerable progress over federated learning (FL), existing methods still face fundamental limitations when applied in realistic, heterogeneous real world environments. Traditional FL algorithms, such as FedAvg, and their numerous variants, have demonstrated that it is possible to train high-capacity models across distributed clients without centralising raw data. However, there are still major challenges related to statistical data heterogeneity, communication efficiency, limited supervision, representation quality, and embedding robustness. In this section briefly summarise the main limitations in the literature and motivate the need for more robust, representative FL frameworks.

Statistical Data Heterogeneity

A main assumption in many distributed learning methods is that local data are independent and identically distributed (i.i.d.) across clients. However, this assumption is rarely correct: many clients maintain datasets from diverse marginal and conditional distributions, reflecting highly non-i.i.d. data partitions (Jimenez-Gutierrez et al., 2025; Kairouz et al., 2021). Under data heterogeneity, local stochastic gradients become biased with respect to the global objective, and simple model averaging, as in FedAvg, can cause *client drift*: local models move towards client data-specific optima that are misaligned with the global model optima. Several research studies have proposed regularisation-based approaches to mitigate this client drift. For example, FedProx, which penalises from the current local model weights previous global model parameters, and FedDyn (Acar et al., 2021), which adds dynamic gradient correction terms to the model aggregation step.

Control-variate methods such as SCAFFOLD (Karimireddy et al., 2020) aim to reduce client drift by following gradient corrections of local models. While these approaches improve stability and model convergence in a small part of data scenarios. These methods frequently rely on additional memory space, extra calculation between previous global model or tuning of regularization hyperparameters. Additionally, they mostly operate at the *parameter-gradient level* and do not clearly address the quality or alignment of learned representations.

Communication and Systems Constraints

Another major limitation of existing FL methods arises from communication and systems constraints. Standard FL algorithms require iterative exchange of local-global model parameters or many updates between the central server and clients over many communication rounds. In cross-device FL, clients suffer from limited bandwidth network connections and have limited computing capabilities. Thus, many methods attempt to reduce communication cost through partial participation, update weight compression, or fewer local epochs. However, these approaches can cause optimisation instability and fairness issues under non-i.i.d. data. Compressing the local network model approaches reduces the size of model updates. However, the compression may degrade convergence when combined with already noisy and biased local gradients. Also, it may cause an overfitting problem. Similar to the overfitting problem, aggressive local training (many local epochs, such as 1000 epochs) can cause client drift, while communication slows down convergence. Thus, existing FL methods still struggle to successfully achieve a balance between communication efficiency and optimal performance under realistic data constraints.

Label Scarcity and Limited Supervision

A further limitation is the dependence on labelled data. Many FL algorithms assume that each participating client possesses enough samples with labels to perform supervised training with

standard cross-entropy losses. In the real world, data are frequently collected from many devices or sensors, and remain largely unlabeled. Under these conditions, supervised FL methods rely on annotated data. Semi-supervised and self-supervised FL approaches have been proposed (Jeong et al., 2021), (Liang et al., 2020), (Y. Liu, Sun, et al., 2021) to train on distributed unlabelled data; however, these methods still exhibit suboptimal performance for supervised representation learning. Also, diverse and limited data availability and evaluation on simplified benchmarks highlight the need for novel unsupervised FL methods. Specifically, contrastive and metric-learning-based FL methods remain extremely sensitive to the availability of positive and negative samples. In addition, label sparsity and unbalanced class coverage per client directly limit the quality of the learned embeddings.

Representation Quality in Non-i.i.d.

Most traditional FL algorithms are designed from an optimisation perspective, treating the local model training as a black box and focusing on minimising a global model loss. However, recent work on federated representation learning, contrastive learning, and metric learning has shown that non-i.i.d. data can significantly degrade the learned embedding space. Clients trained on label-skewed distributions or covariate shift among local datasets may learn highly specific, incompatible features. Thus, global model aggregation aims to calculate these divergent representations, which can lead to several difficulties:

- **Representation misalignment:** Features corresponding to the same class occupy different areas in the embedding space across clients, resulting in insufficient cross-client consistency after model aggregation.
- **Increased intra-class variance:** Limited orientation to intra-class variance on each client leads to distorted or overfitted local feature clusters, and the aggregated representation shows high intra-class variance.
- **Reduced inter-class variance:** Label-skewed data distributions and limited negative samples per client expand the local representation boundaries, causing the global model representation space to overlap many class margins.
- **Cross-client representational collapse:** In some cases, clients have a small set of classes (one class per client or two classes per client) or highly imbalanced samples per class (a single class dominates the client dataset). Additionally, local contrastive or metric learning objectives can lead to degenerate or highly predictable embeddings.

Recent methods, such as MOON (Q. Li et al., 2021), FedCL and FedProto aim to align representations or prototypes across clients by introducing additional contrastive or clustering-based regularisation. While these approaches improve model robustness to data heterogeneity,

they frequently require extra memory (for storing current global model or previous global model representations), detailed weighting of auxiliary losses. However, they do not fully prevent representational collapse or feature misalignments under severe non-i.i.d. data regimes. Moreover, many algorithms lack a deep analysis of representation quality (e.g., intra-class, inter-class distances) beyond typical accuracy metrics.

Personalisation, Fairness, and Generalisation Trade-offs

Another challenge in FL lies in global performance, personalisation, and fairness. Personalised FL methods (Y. Deng et al., 2020), (Fallah et al., 2020), (Sattler et al., 2020) or (Tamirisa et al., 2024) aim to tailor local models to individual clients by learning client-specific projection heads, fully model parameter, adaptation multilayer, or meta-learned initialisations. While these approaches significantly improve local model performance. Clustered FL partitions clients into groups with similar data distributions, training separate models per cluster, which complicates deployment and evaluation in dynamic environments.

Moreover, existing methods rarely provide guarantees of fairness across heterogeneous clients. Clients with small-scale, limited or highly label-skewed datasets may receive worse models. Domain shifts can produce inefficient or inconsistent predictions for clients, and current algorithms frequently prioritise global model aggregation, personalising each client instead of achieving optimal performance under many data regimes.

Robustness, Accessibility, and Security Limitations

FL is usually motivated by data-preserving considerations. However, in the practical world, these methods face multiple robustness and accessibility challenges. Typical FL methods keep the local dataset private, without accessing the raw data. However, updating global model parameters relies on the calculation of each local model parameter. Each communication round, these local model parameters are shared with the global model for model aggregation. This communication round can cause privacy and security concerns by sharing model weights and gradients. Some network models are capable of reproducing data samples from their gradients or weights. It is a significant problem to work as a future direction in the FL field.

In addition, from a robustness perspective, FL frameworks are vulnerable to malicious or unreliable clients. Byzantine-robust aggregation methods (Lu et al., 2024) and anomaly detection algorithms have been proposed to address these limitations; however, many of these methods mostly rely on simplified threat models or assumptions of no risk or attacks during training.

Benchmarking and Evaluation Gaps

In the end, there are significant limitations in FL or SSL in how existing methods are evaluated. Many FL algorithms are evaluated on a small subset of vision benchmarks or text-based bench-

marks using synthetic non-i.i.d. data partitions (e.g., Dirichlet splits). Representation quality is usually assessed indirectly through downstream task accuracy, without deeper analysis of embeddings, representational collapse, or cross-client alignment. These limitations make it difficult to compare methods in terms of their ability to produce robust, transferable, and well-structured representations in (de)centralised learning paradigms.

In summary, self-supervised learning has significantly advanced visual representations, current SSL approaches remain limited in several ways, such as distributed data settings and representation learning under limited data scenarios. These methods typically rely on manually customised data augmentations and large-scale curated datasets, limiting their applicability when data are scarce, unlabelled, or domain-specific. Thus, model performance is usually sensitive to architectural network model design (deciding how many layers, normalisation, etc.), setting up hyperparameters, and batch-size conditions, leading to instability and reduced reproducibility. In addition, most SSL methods provide limited knowledge for explicitly controlling the embedding space, making it difficult to guarantee discriminative patterns under data distribution shift. Contrastive and transformer-based representation learning methods have high computational and memory demands, which limit their deployment in resource-limited or cutting-edge environments.

Following the limitations in SSL, existing federated learning methods have established a solid foundation for distributed training under data-preserving constraints; FL methods face substantial limitations when there is data heterogeneity, label scarcity, and representation misalignments. Thus, these challenges are critical for visual representation learning and metric-based objectives, which depend on balanced class distributions and consistent feature spaces.

Motivated by these limitations, this thesis investigates novel methods and algorithms for robust and efficient visual representation learning in both centralised limited-data settings and decentralised learning environments. The proposed approaches aim to reduce reliance on large-scale datasets and augmentation tuning, improve robustness of models under data heterogeneity and covariate shift. In addition to these contributions, this research presents solutions where data accessibility is necessary, data-efficient learning for real-world applications is essential, and non-i.i.d. conditions are significant.

Chapter 3

Enhancing Self-Supervised Learning Through Batches

"Self-supervised learning is the dark matter of intelligence."
— Yann LeCun

The majority of this chapter is derived from our published work at the International Joint Conference on Neural Networks (IJCNN), entitled The Bad Batches: Enhancing Self-Supervised Learning in Image Classification Through Representative Batch Curation (Goksu, O. and Pugeault, N., 2024)[1].

Learning effective visual representations is crucial to the success of image understanding tasks, including classification, detection, and segmentation. However, acquiring high-quality labelled data remains a significant bottleneck in many practical settings.

Unsupervised representation learning, most notably self-supervised learning (SSL), has emerged as a powerful paradigm for extracting robust representations from unlabeled data. By leveraging carefully designed pretext tasks, SSL methods learn to capture the underlying structure and statistical regularities of the data without relying on manual annotations. SSL methods, a subclass of unsupervised learning, have demonstrated remarkable performance across a wide range of downstream tasks.

Unlike traditional unsupervised techniques, such as clustering or generative modelling (e.g., GANs), which often focus on modelling the data distribution or grouping similar samples, self-supervised approaches learn transferable representations by solving proxy tasks that do not require manual annotations. These methods typically involve two stages: (i) pretraining, where a network is optimised on a self-defined pretext task to capture semantic or structural properties of the input data, and (ii) fine-tuning, where the learned representations are adapted to a specific downstream task using a limited amount of labelled data.

The flexibility of SSL enables it to scale across domains and tasks, often rivalling or surpassing

[1] https://arxiv.org/pdf/2403.19579

fully supervised counterparts when labelled data is scarce (Ericsson et al., 2022), (Jaiswal et al., 2020). Motivated by these capabilities, this chapter focuses on the design and application of a self-supervised learning method in scenarios with limited labelled training data, examining both pretraining and fine-tuning stages. In this work, we aim to answer the question *" **RQ1;** How can we improve self-supervised contrastive learning under limited data conditions, where augmentations may introduce misleading (false positive/negative) views?"*. We hypothesise that data augmentations, while essential for contrastive learning, may occasionally generate weak or unstable views of the same sample, as illustrated in Figure 3.1a (Robinson et al., 2021; Tian et al., 2020). These views may lead to false positive or false negative pairings during training, ultimately degrading the robustness of the learned representations.

We assume that batches dominated by such weak augmentations (referred to as **bad batches**) introduce noise into the contrastive objective and hinder convergence. In this work, we propose to detect and discard these detrimental batches using a distributional similarity-based criterion, Frechet ResNet Distance (FRD). By selectively focusing on representative views, those that preserve the semantic content of the original image, self-supervised learning can yield more robust and generalizable features, even under limited data conditions. We conduct a comprehensive evaluation of several state-of-the-art self-supervised learning algorithms, analysing how their performance degrades under varying levels of attribute overlap among data instances.

This analysis reveals their sensitivity to false positives and false negatives, underscoring the need for more robust techniques that can effectively handle such cases. In this chapter, we detail the methodology, experimental setup, and results demonstrating its effectiveness. The code is publicly available.[2]

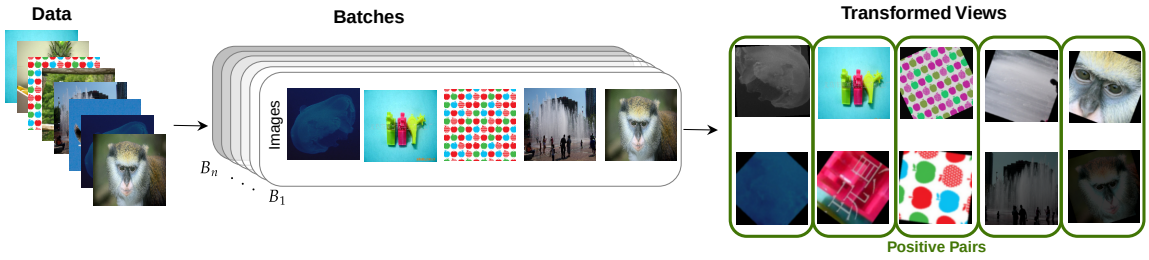
3.1 Introduction

The pursuit of learning robust representations without human supervision is a longstanding challenge. The recent advancements in self-supervised contrastive learning approaches have demonstrated high performance across various representation learning challenges. However, current methods depend on the random transformation of training examples, resulting in some cases of unrepresentative positive pairs that can have a large impact on learning. This limitation not only impedes the convergence of the learning process but the robustness of the learnt representation as well as requiring larger batch sizes to improve robustness to such bad batches. This paper attempts to alleviate the influence of false positive and false negative pairs by employing pairwise similarity calculations through the Fréchet ResNet Distance (FRD), thereby obtaining robust representations from unlabelled data. The effectiveness of the proposed method is substantiated by empirical results, where a linear classifier trained on self-supervised contrastive representations achieved an impressive 87.74% top-1 accuracy on STL10 and 99.31% on the

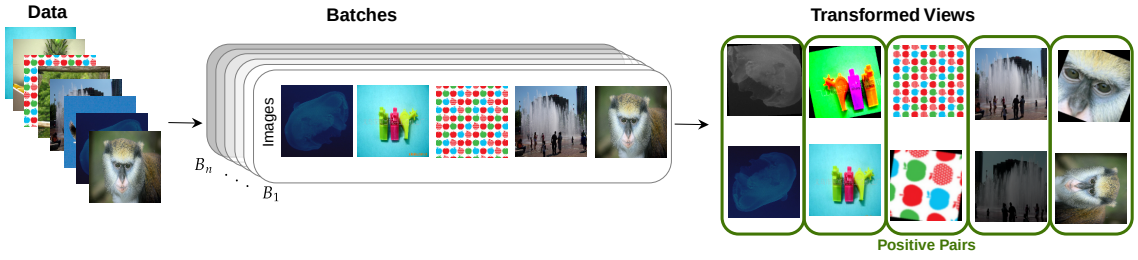
[2]https://github.com/ozgugoksu/Self-Supervised-Learning-/tree/main/Contrastive/SimCLR

Flower102 dataset. These results emphasize the potential of the proposed approach in pushing the boundaries of the state-of-the-art in self-supervised contrastive learning, particularly for image classification tasks.

Despite the rapid progress of supervised deep learning, the availability of labelled data remains a limitation: transferring the success of deep learning approaches to applications and domains where data and annotations are scarce is a challenge. This has led to the popularity of approaches for unsupervised and self-supervised pre-training: training a deep network to solve a *pretext task* on a large dataset of usually unlabelled examples, before fine-tuning it for the desired downstream task on a typically much smaller, but correctly labelled dataset. This approach has been successful in allowing the use of deep models for tasks where limited labelled data is available. One successful example of self-supervised learning is *contrastive learning*, such as SimCLR (T. Chen, Kornblith, Swersky, et al., 2020), where the pretext task is to optimise a latent space such that augmented versions of the same training example are more similar to each other than to other training examples. Contrastive learning has shown excellent performance (Caron et al., 2021), (Hu et al., 2021), (T. Chen, Kornblith, Norouzi, & Hinton, 2020) but requires large datasets, large batch sizes and long training times. In this article, we propose the hypothesis



(a) Existing Methods



(b) Our Method

Figure 3.1: Existing self-supervised contrastive methods mainly rely on various data augmentations to increase diversity; however, this causes weak transformed views of original images. Our method aims to eliminate weak augmented views, such as darker images as a similar pair, and insufficient colour changes.

that contrastive learning is negatively impacted by a relatively small proportion of training examples and augmentations. This is illustrated in Figure 3.1, which shows examples of positive

pairs produced by existing methods. In those examples, we observe that data augmentation has degraded essential information in the image, leading to the creation of *false positive*, which could hinder the convergence of the contrastive objective. We argue that discarding such "**bad batches**" would converge more effectively and efficiently, thereby necessitating smaller batch sizes. While self-supervised contrastive learning has shown promise in representation learning, its reliance on randomly-formed batches containing numerous false positives and negatives presents a major obstacle.

In this work, we present a simple yet effective approach, which involves the evaluation of each batch's representations by measuring their FRD (Fréchet ResNet Distance). FRD can provide a criterion to detect and reject "bad batches" in unsupervised representation learning. We assume that bad batches are characterised by an abundance of false positives and negatives, whereas good batches consist of images with semantically similar views. This study focuses on learning robust representations through the utilisation of good batches, which consist of appropriate views of the original images. Conversely, darker, blurry, or randomly cropped and inappropriate views are intentionally excluded to enhance the quality of the learned representations. In summary, our major contributions to this work are

- a criterion to evaluate augmented batches by Fréchet ResNet Distance (FRD) calculation,
- a regularised loss function to reduce the required batch size for training self-supervised contrastive models,
- a robust representation learning method by avoiding false positives in batches.

This work breaks new ground by proposing a novel approach for evaluating batch quality in self-supervised contrastive learning (SSCL). In traditional SSCL, batches are typically randomly formed, which can lead to the inclusion of distorted or irrelevant image views, thereby introducing false positives and negatives. By implementing a method based on (FRD), we can identify and remove batches that are likely to contain such misleading samples. Our process ensures that only batches with semantically similar views of the original images, referred to as "good batches", are utilised for representation learning. By excluding darker, blurry, or inappropriately cropped views, which are considered irrelevant, the quality of the learned representations is significantly enhanced. This breakthrough has the potential to significantly accelerate the development of efficient and generalizable self-supervised models.

3.1.1 Related Work

Learning representations without human supervision has long been an ongoing challenge due to enhancing generalization and extracting relevant features from raw data. Self-supervised learning, as opposed to traditional supervised approaches, learns features from unlabelled data and captures representative features as well as the underlying data distribution. Self-supervised

learning approaches are divided into two subcategories: generative and discriminative (Ozbulak et al., 2023). Generative models such as Generative Adversarial Networks (GANs) and autoencoders are capable of generating realistic, high-quality images from the image dataset. However, they might be insufficient to learn representations, since achieving interpretable representations in the latent space and hard-to-learn informative features are challenging (T. Chen, Kornblith, Norouzi, & Hinton, 2020).

Unlike generative algorithms, discriminative models learn representations through pretext tasks, which focus on understanding the content of images to learn representations instead of generating an image. There are many pretext tasks such as image rotation prediction (Johnson et al., 2020), jigsaw puzzle solving (Noroozi & Favaro, 2016), image colourization (Larsson et al., 2017) and contrastive learning (T. Chen, Kornblith, Norouzi, & Hinton, 2020), (Jaiswal et al., 2020), which are designed to replace supervision and extract meaningful and informative representations from raw data, which can later be used for downstream tasks. Contrastive learning (CL) is a self-supervised learning methodology that is designed to determine a representation space wherein similar instances (positive) are drawn into closeness while dissimilar instances (negative) are pushed apart. Contrastive learning (CL) facilitates mutual information (MI) among views while preserving task-relevant information. MI evaluates the statistical dependency of two random variables. Regarding CL, the two views of the same image are utilised as random variables. CL attempts to decrease the MI between the two views by maximising the similarity (and minimising the dissimilarity) between positive pairs while minimising the similarity between negative pairs. The fundamental assumption underlying CL is that obliging the model to obtain representations emphasising task-relevant information enhances the similarity among positive pairs, facilitating a more straightforward differentiation between positive and negative pairs. This objective is realised by being able to consistently map positive pairs onto proximate points within the feature space, thereby enhancing their ability to distinguish from negative pairs.

Positive pairs encompass images featuring similar views of the original raw data, including various transformations such as rotations or cropping applied to the same image. Conversely, negative pairs are deliberately chosen to be dissimilar, typically involving random combinations from the dataset. Without labels, recent improvements rely on instance discrimination problems in which positive pairs are treated as two augmented versions of the same images while considering the remainder as negatives (T. Chen, Kornblith, Norouzi, & Hinton, 2020), (K. He et al., 2020). As an illustration, the initial row of the figure 3.1 shows two transformed views of the same image, captured from distinct angles. CL aims to extract features that are present in both views, like the object's shape and colour. This shared information is generally beneficial for recognising the object in different contexts. However, some details specific to one angle, like a unique marking or texture, might be discarded as noise during the simplification process in figure 3.1. This discarded information could potentially be relevant for a specific task, like identifying individual instances of the object. This process ensures that the positive pairs share the same

semantic content or meaning. The model can generalise well to downstream tasks by learning to recognise these shared features.

Existing self-supervised contrastive learning methodologies predominantly generate batches through random transformations like cropping and colour jittering. The randomness in batches introduces many weak transformed views (false positives or negatives) and hinders the model's ability to learn from relevant image pairs in figure 3.1. Moreover, the requirement to fine-tune transformations for each dataset introduces unnecessary complexity and undermines performance on subsequent tasks. Several methods eliminate false negative impact on representation; they require larger batch sizes, such as 4,096 (T. Chen, Kornblith, Norouzi, & Hinton, 2020), and 8,192 (K. He et al., 2020), (Caron et al., 2021). A larger batch provides a more diverse set of negative examples, making it harder for the model to distinguish them based on spurious correlations or artefacts of the data augmentation process. This encourages the model to focus on learning more generalizable features. However, it's important to note that training with large batches requires more memory, which can be a limiting factor for smaller machines or datasets. Many approaches to tackle this problem have been offered to eliminate weak data augmentations in batches. GeoDA (Cosentino et al., 2022) method focuses on shape-preserving geometric augmentations on images, and (Mishra et al., 2022) proposes context-aware augmentation. (Fawzi et al., 2016) aims for effective representation learning by adaptive augmentations based on model uncertainty. The Patch Curation method, as outlined in (Welle et al., 2021), introduces a batch curation algorithm through the transformation of images into patches. This approach determines batches by assessing their Euclidean distance. Not only, new data augmentation more specific to use datasets, but it is also not representative.

Previous investigations into batch behaviours have predominantly concentrated on novel data augmentation methods (Kurtuluş et al., 2023), hard negative mining (W. Zhu et al., 2023), (Robinson et al., 2021), or the utilisation of larger batch sizes for self-supervised contrastive learning. However, batch evaluation by measuring pairwise similarity is not only effective in avoiding weak representation learning on downstream tasks but also benefits from a smaller batch size and sufficient data augmentation tunings. Traditional self-supervised learning leverages large-scale data and pretext tasks like image inpainting or contrastive learning to induce informative data representations. While effective, these methods often require significant computational resources. Also, models learn less discriminative features. We propose a novel batch curation approach that significantly improves representation learning efficiency, requiring neither extensive datasets nor large batch sizes. This enables effective self-supervised learning in resource-constrained scenarios. Larger batches can mitigate random noise in gradients, leading to smoother optimisation. However, excessively large sizes can worsen vanishing gradients, hindering efficient network updates. Increased batch size introduces a wider variety of negative examples for contrasting, potentially enriching representations. However, diminishing marginal returns and potential overfitting to strong correlations within large batches must be balanced.

Certain loss functions, like Normalised Temperature-scaled Cross-Entropy (NT-Xent), rely on statistical properties of the data distribution, benefiting from larger batches for accurate estimation. However, over-reliance on larger batches can make performance sensitive to specific data characteristics.

SwAV (Caron et al., 2020) is a self-supervised contrastive learning method that aims to enhance representation learning by leveraging massive web-based images for data augmentation. BYOL (Grill et al., 2020) utilises a two-network architecture with a student and teacher network without a memory bank or momentum contrast. These models benefit from representation learning on many downstream tasks; however, these models have larger batch sizes. Like SimCLR (T. Chen, Kornblith, Norouzi, & Hinton, 2020) and SimCLR-v2 (T. Chen, Kornblith, Swersky, et al., 2020), SwAV trains with large batch sizes ranging from 4,096 to 16,384. In contrast, SimCLR-v2 uses even larger batch sizes, exceeding 32,768. Even BYOL has a simple framework, from 256 to 1,024 batch size. However, training larger batches with these models may not consistently be the optimal strategy, and striking the right balance is essential for effective and high-quality representation learning.

SimCLR-v2 aims to improve representation learning by using a larger ResNet and inspired memory mechanisms like MoCo (K. He et al., 2020), which is a dynamic dictionary that stores past data augmentations or representations. However, they have limitations such as memory management, outdated negative samples in batches, and duplicate samples. Current methods, such as (T. Chen, Kornblith, Norouzi, & Hinton, 2020), (T. Chen, Kornblith, Swersky, et al., 2020), (J. Wu et al., 2024), rely on a pairing strategy that designates two augmented views of the original raw data as positive pairs and randomly selects the remaining transformed views as negative pairs. However, the indiscriminate generation of random images can lead to numerous inaccurate augmentations. This has resulted in less discriminative representations and a diminished generalisation capacity for self-supervised contrastive models. These methods (SimCLR, or similar to SimCLR) focus on representation learning in self-supervised contrastive learning; however, we show that these approaches cannot extract robust representations and have poor performance on downstream tasks. Furthermore, current contrastive objectives struggle to suppress gradient noise and to correct false positive pairs arising from stochastic augmentations. Their optimisation behaviour often becomes unstable at smaller batch sizes, since the loss implicitly assumes access to a sufficiently large and diverse set of negative samples, as demonstrated in (C.-Y. Chuang et al., 2022).

3.2 Methodology

The self-supervised model employs a predetermined method for curating random batches without evaluating individual batches. Fréchet ResNet Distance (FRD) scores are calculated for each batch before the activation of the batch curation algorithm. The rationale behind applying the

batch curation method in the early stages of learning is to leverage the advantages of features acquired during the initial epochs. As a subsequent step, we seek to enhance the model's robustness to false positives or noise within the batch. To achieve this, we introduce regularisation to the contrastive loss by incorporating the Huber loss (Barron, 2019) with an associated coefficient.

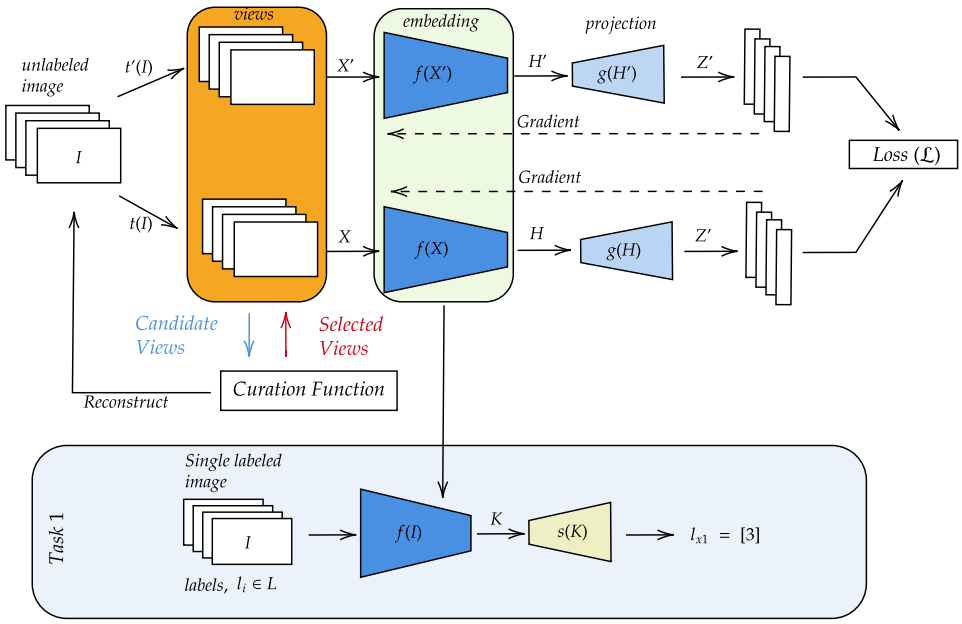


Figure 3.2: Our presented framework for batch curation in self-supervised contrastive learning. Task 1 illustrates image classification as a downstream task. The batch curation part mainly decides which batches are used to update gradients.

3.2.1 Contrastive Learning

Following the contrastive learning framework (as proposed by, e.g., (T. Chen, Kornblith, Norouzi, & Hinton, 2020)), we have a training set of images $x \in \mathcal{D}$, a family of transformations $\mathcal{T}$, a backbone network f such that $h = f(x)$ is a latent encoding of x and a projection head g such that $z = g(h)$ is a projection of h, as illustrated in Figure 3.2. Given a batch of N training samples $B = \{x_1, x_2, \ldots, x_N\} \subset \mathcal{D}$, we generate two stochastic augmentations for each sample by drawing transformations $t \sim \mathcal{T}$. This produces two augmented batches, B_1 and B_2, yielding a total of $2N$ transformed views.

For each original image x_i, the corresponding augmented views in B_1 and B_2 form a *positive pair*, since they originate from the same underlying sample. Conversely, any pair of representations (z_i, z_j) derived from different images $(i \neq j)$ is considered a *negative pair*. Contrastive learning approaches in the literature predominantly employ the Normalised Temperature-scaled Cross-Entropy (NT-Xent) loss (T. Chen, Kornblith, Norouzi, & Hinton, 2020), (X. Chen & He, 2021). Typically, the projection vectors obtained from B_1 and B_2 are concatenated into a single tensor, yielding a unified set of $2N$ representations $\{z_1, z_2, \ldots, z_{2N}\}$. Then, if z_i, z_j is a positive

pair of projections, the NT-Xent loss over those $2N$ projections is:

$$\mathcal{L}_{i,j} = -\log \frac{exp(sim(z_i, z_j)/\tau)}{\sum_{k=1}^{2N} 1_{[k \neq i]} exp(sim(z_i, z_k)/\tau)} \tag{3.1}$$

The function $sim(.,.)$ represents the cosine similarity, and $2N-1$ denotes the number of negative (dissimilar) samples since z_i, z_j are positive (similar) samples transformed in the same image and τ is the temperature value. The contrastive learning literature has shown that minimising this loss produces a latent space $\mathcal{H}$ that can allow solving many downstream tasks by just training a new projection head, requiring limited sets of labelled examples. In this article, we argue that contrastive learning can be severely affected by just a few instances of poor data augmentation. In the following, we describe two approaches to reduce the impact of those: Huber loss regularisation and Fréchet batch curation.

3.2.2 Huber Loss Regularization

Contrastive learning ensures that projections of the same image (positive pairs) are more similar than projections of other samples. We hypothesise that training converges more reliably when the projections z_i and z_j, corresponding to different augmentations of the same image I, are encouraged to be similar, thereby promoting augmentation-invariant representations. To achieve this, we employ the Huber loss, which behaves quadratically for small differences between z_i and z_j, encouraging the model to align their representations closely, while transitioning to a linear penalty for larger discrepancies. This can be enforced by adding a regularisation term to the loss. When there is considerable dissimilarity between positive pairs, the Huber loss imposes a penalty on the loss value, encouraging the model to bring the representations closer together in the latent space. Due to its proven effectiveness in handling outliers, the Huber loss ensures enhanced resilience and stability throughout the learning process. In practice, for a positive pair z_i, z_j, we have

$$l_{i,j} = \begin{cases} 0.5\left(z_i - z_j\right)^2, & \text{if } \left|z_i - z_j\right| < \delta \\ \delta\ \left(\left|z_i - z_j\right| - 0.5\,\delta\right), & \text{otherwise} \end{cases} \tag{3.2}$$

where the parameter δ determines the point at which the loss function transitions from quadratic to absolute, set to 1.0 in our experiments. Over the two transformed versions B_1 and B_2 of a batch B of N images, the Huber loss is defined as the average loss over all N positive pairs in B_1 and B_2:

$$\mathcal{L}_{huber} = \text{mean}(l_1, ..., l_N). \tag{3.3}$$

Including this term, the new regularised objective function becomes

$$\mathcal{L} = \mathcal{L}_{nt\text{-}xent} + \lambda \mathcal{L}_{huber} \tag{3.4}$$

where λ adjusts how strongly the Huber loss affects the gradients, thereby tuning the penalty applied throughout training. Notably, our approach distinguishes itself by not necessitating the utilisation of a memory bank, a larger batch size, extended training periods for effective generalisation, or a huge dataset like ImageNet with a substantial number of samples.

3.2.3 Fréchet Batch Curation

In this paper, we mainly focus on two problems: avoiding learning false positive and negative views, and updating the gradient with each randomly curated batch. In our batching strategy, the FRD score is used to assess the similarity between the distributions of augmented images within a batch. If the FRD score for a given batch falls below a predefined threshold, the batch is accepted and its gradients are used to update the model parameters. Otherwise, the batch is discarded and regenerated with new augmentations. The threshold is computed as the mean FRD score across all batches at 5^{th} epoch. We estimate this threshold after several epochs rather than at the beginning of training, since the random initialisation of model weights leads to unstable and unreliable FRD estimates in the early stages.

During the initial five epochs, we employ algorithms without batch curation to facilitate the learning of representations. In our experiments, empirical testing of various starting epochs reveals that the 5^{th} epoch is optimal as the starting point for curating the batch algorithm.

3.2.4 Concentration of FRD and Threshold Selection

The Fréchet distance calculates the difference between two normal distributions. In the field of generative learning, the Fréchet Inception Distance (FID) is commonly used to measure the similarity between real and fake images in the generative learning literature. We extract representations in the latent space, including the projection head, as depicted in Figure 3.2, for (FRD) scores. Let us consider a single batch B_k, representing the k^{th} randomly sampled batch of size N. Denote by $B_1^{(k)}$ and $B_2^{(k)}$ the two independently augmented versions of B_k, and let representations of $B_1^{(k)}$,

$$\mathbf{z}_1^{(k)}, \mathbf{z}_2^{(k)}, \ldots, \mathbf{z}_N^{(k)} \tag{3.5}$$

be their feature embeddings from the ResNet18 architecture. Define the mean and covariance matrix of the embeddings for $z_i^{(k)}$ as

$$\mu_1^{(k)} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{z}_i^{(k)}, \qquad \Sigma_1^{(k)} = \frac{1}{N}\sum_{i=1}^{N}\big(\mathbf{z}_i^{(k)} - \mu_1^{(k)}\big)\big(\mathbf{z}_i^{(k)} - \mu_1^{(k)}\big)^{\top}. \tag{3.6}$$

and similarly $\mu_2^{(k)}, \Sigma_2^{(k)}$. The Fréchet ResNet Distance between batches $(B_1^{(k)}, B_2^{(k)})$ is calculated as

$$\mathrm{FRD}\big(B_1^{(k)}, B_2^{(k)}\big) = \|\mu_1^{(k)} - \mu_2^{(k)}\|^2 + \mathrm{tr}\Big(\Sigma_1^{(k)} + \Sigma_2^{(k)} - 2\big(\Sigma_1^{(k)}\Sigma_2^{(k)}\big)^{1/2}\Big). \tag{3.7}$$

where $\mathrm{tr}(\cdot)$ denotes the trace operator, which returns the sum of the diagonal elements of a square matrix. Consequently, for a given epoch, we compute the empirical mean and standard deviation of the FRD scores across the M batches:

$$\begin{aligned} \widehat{\mu}_{\mathrm{FRD}} &= \frac{1}{M}\sum_{k=1}^{M} \mathrm{FRD}\big(B_1^{(k)}, B_2^{(k)}\big), \\ \widehat{\sigma}_{\mathrm{FRD}} &= \sqrt{\frac{1}{M}\sum_{k=1}^{M}\Big(\mathrm{FRD}\big(B_1^{(k)}, B_2^{(k)}\big) - \widehat{\mu}_{\mathrm{FRD}}\Big)^2}. \end{aligned} \tag{3.8}$$

concentrate around their population counterparts. We set our curation threshold as

$$\tau = \widehat{\mu}_{\mathrm{FRD}} + \kappa\,\widehat{\sigma}_{\mathrm{FRD}}, \qquad \kappa \approx 1. \tag{3.9}$$

We define the threshold as the average similarity across the dataset, which serves as a lower bound for acceptable similarity scores. This choice is motivated by the observation that significantly lower similarity values are often associated with a high incidence of false positives or false negatives, both of which can be detrimental to representation learning. By enforcing this lower bound, we aim to filter out low-quality pairings within a batch, thereby reducing the risk of harmful noise and preserving the integrity of learned representations.

Our presented batch curation framework in figure 3.2 for self-supervised learning follows default parameter training for the first 5 epochs. Algorithm 1 outlines the batch curation procedure. In this process, we estimate the threshold by computing the mean FRD score of all batches at epoch 5, which is then used to filter batches in the remaining epochs. Using an average threshold value for batch evaluation can provide balance in data distributions instead of using the maximum or minimum in a dataset. The threshold value is 0.56 for our experiments. The FRD score shows the similarity of views, and smaller scores represent highly similar views; otherwise, data augmentation can transform hardly similar views. Proposed FRD batch curation discards a candidate batch B_k if $\mathrm{FRD}(\mathrm{B}_1^{(\mathrm{k})}, \mathrm{B}_2^{(\mathrm{k})}) > \tau$.

The process uses similar, better-augmented views as a good batch for representation learning. In this study, the average value, referred to as the threshold, was employed in the evaluation of various representations. The rationale behind utilising trained representations for threshold calculation is grounded in the pursuit of obtaining quantifiable and representative features. This ensures that the threshold, a pivotal parameter in the evaluation process, is derived from a set of features that are not only informative but also reflective of the model's learning capabilities. By focusing on the 5^{th} epoch, which has been identified as the epoch of peak performance, we aim to capture and leverage the most effective and discriminative features embedded in the trained representations. In essence, this approach seeks to enhance the interpretability and reliability of the threshold by basing it on features extracted from representations at a specific epoch that has

proven to yield optimal results.

Input: Training dataset $\mathcal{D}$, model f_θ, batch size N, number of batches M, threshold parameter $\kappa \approx 1$

Output: Updated model parameters θ

Initialisation: Randomly initialise model parameters θ.

Compute FRD statistics at epoch 5 to obtain $\widehat{\mu}_{\text{FRD}}$ and $\widehat{\sigma}_{\text{FRD}}$.

Set threshold $\tau = \widehat{\mu}_{\text{FRD}} + \kappa\widehat{\sigma}_{\text{FRD}}$.

for *each training epoch* **do**

  for $k = 1$ **to** M **do**

    Sample a batch $B_k \subset \mathcal{D}$ of size N.

    Generate two augmented versions $B_1^{(k)}$ and $B_2^{(k)}$ using random transformations $t \sim \mathcal{T}$.

    Compute the projected representations $\{\mathbf{z}_i^{(k)}\}_{i=1}^N$ and $\{\mathbf{z}_j^{(k)}\}_{j=1}^N$.

    Compute batch statistics $(\mu_1^{(k)}, \Sigma_1^{(k)})$ and $(\mu_2^{(k)}, \Sigma_2^{(k)})$.

    Compute $\text{FRD}^{(k)} = \text{FRD}(B_1^{(k)}, B_2^{(k)})$.

    if $\text{FRD}\big(B_1^{(k)}, B_2^{(k)}\big) \leq \tau$ **then**

      Compute the similarity loss $\mathcal{L}_{\text{Huber}}(\mathbf{z_i}, \mathbf{z_j})$.

      Update parameters $\theta \leftarrow \theta - \eta\nabla_\theta\mathcal{L}$.

    else

      Regenerate the batch B_k with new augmentations and repeat.

    end

  end

end

Algorithm 1: FRD-Based Batch Curation and Training Procedure

3.3 Evaluation

To validate our hypotheses and the proposed approach, we compared performance with baseline contrastive learning approaches on multiple datasets. Moreover, we systematically explore the impact of the Huber loss, akin to ℓ^1 and ℓ^2 losses, by conducting linear evaluations and transfer learning assessments. These evaluations are carried out across widely recognized datasets, providing a comprehensive investigation into the performance and capabilities of our proposed approach relative to existing methods.

Datasets and Metrics: Self-supervised approaches generally use ImageNet for pre-training with large batch sizes such as 4,096 or 8,192 to improve robustness to outliers at the cost of memory. We argue that our improved approach can perform with smaller batch sizes and datasets. Accordingly, we tested models trained on ImageNet and CIFAR10 for self-supervised learning and tested the learnt representations on a range of downstream tasks such as CIFAR100 (T. Chen,

Kornblith, Norouzi, & Hinton, 2020) (resolution 32×32, 100 classes), STL10 (Coates et al., 2011) (resolution 96×96, 10 classes), ImageNet(J. Deng et al., 2009) (resolution 224×224, 1000 classes), Caltech101 (Fei-Fei et al., 2004) (resolution 224×224, 101 classes), Flower102 (Nilsback & Zisserman, 2008) (resolution 224×224, 102 classes), and MNIST (L. Deng, 2012) (resolution 28×28, 10 classes).

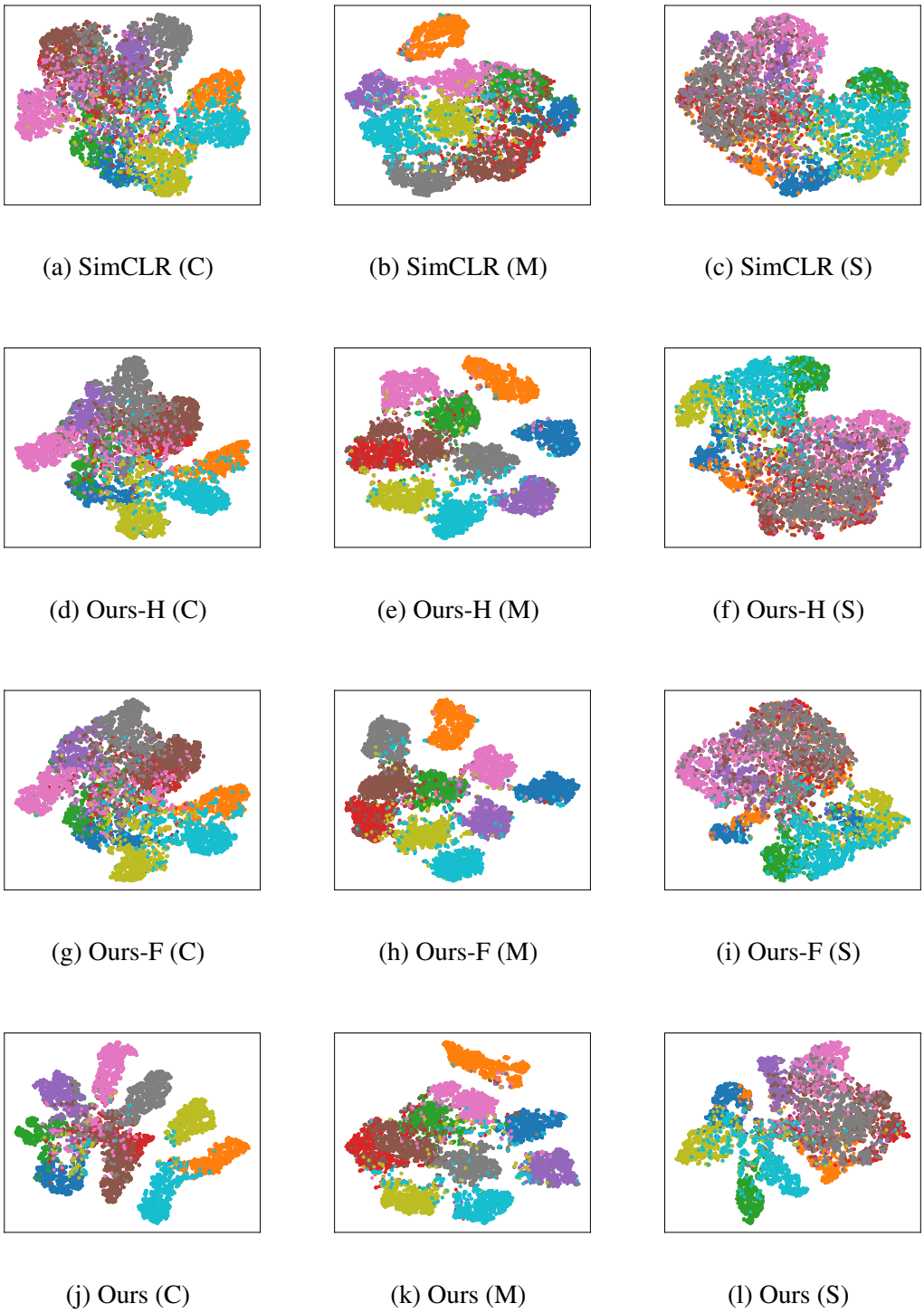
Small-scale Training: We use random cropping and resizing, colour jittering, random horizontal flipping, and random greyscale for CIFAR10 as transformations to create pairs. We use a Resnet-50 architecture with a smaller convolution kernel size (3×3) as a base encoder and a non-linear projection head with two linear layers with batch normalisation to project representations to a 128-dimensional latent space. Furthermore, we use a lambda warm-up scheduler for 30 epochs, and following that we continue the cosine decay scheduler without restarting for training. The only difference between CIFAR10 and ImageNet is that for the latter, the first layer of the encoder network has a kernel size (7×7), to ensure that the latent space is of the same dimension. The choice of different kernel sizes for extracting representative features is motivated by the disparity in resolution between the ImageNet and CIFAR10 datasets.

Loss Function and Batch Size: We enhance the contrastive loss with a regularising Huber loss, with a δ parameter to tune the impact of sensitivity. Outliers can cause significant errors, and Huber loss can address the error problem by transitioning from a quadratic loss, which is similar to ℓ^2 loss for small errors and absolute, ℓ^1 loss for large errors. In our case, we assume that false positives and negatives can be considered outliers and noise within the representation space. The δ parameter, which is set to 1.0 in our experiments, can control these changes. We hypothesise that some unusual combination of image transformation and training images can produce misleading examples (as illustrated in Figure 3.1) that hinder the convergence and performance of contrastive learning.

3.4 Results

Figure 3.3 illustrates t-SNE plots for three distinct datasets. We assess representation quality using linear evaluation accuracy and t-SNE embeddings. Under these metrics, the regularised version of our method consistently outperforms SimCLR on MNIST and CIFAR10. Furthermore, combining FRD with the Huber loss yields the strongest results across all three datasets, producing the highest linear probe accuracy and the clearest separation of classes in the learned feature space. The relationship between FRD and the Huber loss arises from their complementary functions during training. FRD acts at the batch-selection level by filtering out weak or unstable augmentations that would otherwise introduce noise into the gradient updates. In contrast, the Huber loss operates at the representation level by reducing the influence of extreme discrepancies between positive pairs, thereby mitigating the effects of both false positives and false negatives within a batch. Together, these mechanisms lead to more stable optimisation and more coherent

feature representations, which is reflected in the improved class separation observed in the t-SNE visualisations.



(a) SimCLR (C) (b) SimCLR (M) (c) SimCLR (S)

(d) Ours-H (C) (e) Ours-H (M) (f) Ours-H (S)

(g) Ours-F (C) (h) Ours-F (M) (i) Ours-F (S)

(j) Ours (C) (k) Ours (M) (l) Ours (S)

Figure 3.3: The illustration shows the impact of representation learning by our methods and SimCLR with only 30 epochs of fine-tuning on several datasets. C represents CIFAR10, M represents MNIST, and S is for STL10 datasets. Ours-H has trained models only with Huber loss, and Ours-F represents FRD batch curation without Huber loss. Ours is a combination of Huber loss and FRD.

Table 3.1 provides a comparison of our method against other contrastive learning approaches, comparing the number of parameters, the backbone network model and performance. All models are trained and tested on ImageNet. We note that the proposed method achieves superior performance with a significantly lower number of parameters, achieving an accuracy of 83.94% when we compare the baseline SimCLR with 62.5% in this table 3.1. This result supports our hypothesis that bad batches are a large hindrance to contrastive learning approaches.

To verify whether the proposed approach allows for better performance in low-memory, low-data situations, we compared performance pre-training and testing on CIFAR10. Table 3.2 presents results obtained through varying epochs of training and batch sizes with the ResNet-50 backbone network. The proposed approach achieves the best top-1 accuracy of 99.67%, with a batch size of only 128 and only 200 epochs of pre-training. It is worth noting that DINO ViT-S (Caron et al., 2021) achieved comparable performance on CIFAR10 with 99.1%, but it is important to also note the larger number of epochs required (800 versus 200) and substantial parameterisation of vision transformers (21 million trainable parameters, DINO ViT-S), which may pose memory consumption challenges during self-supervised training. The results also show that the achieved performance is comparable to the one achieved with a fully supervised training of the network.

One question is whether it is the Huber regularisation or the Fréchet batch curation that provides the main improvement. To elucidate this, table 3.3 provides an ablation study comparing regularisation with L_1, L_2 or Huber loss, with and without curation, against standard SimCLR and PatchBatch (Welle et al., 2021), another batch curation approach. The results show that both components lead to a large improvement in performance and that Huber loss regularisation performs better than L_1 or L_2. This confirms that the Huber loss offers robustness to noise by modulating sensitivity to outliers through the parameter δ, where $\delta = 1.0$ proved effective. We assume that outliers might be false positives and negatives in our experiments.

Method	Backbone Network	Param (M)	Top-1
SimCLR (T. Chen, Kornblith, Norouzi, & Hinton, 2020)	ResNet-50	25	62.5
SimCLR (T. Chen, Kornblith, Norouzi, & Hinton, 2020)	ResNet-50 (2×)	94	74.2
SimCLR (T. Chen, Kornblith, Norouzi, & Hinton, 2020)	ResNet-50 (4×)	375	76.5
AMDIM (Bachman et al., 2019)	Custom ResNet (2×)	626	68.1
BYOL (Grill et al., 2020)	ResNet-200 (2×)	250	79.6
DINO (Caron et al., 2021)	VIT-S	21	77.0
DINO (Caron et al., 2021)	VIT-B	85	78.2
BINGO (Xu et al., 2022)	ResNet-34	23.1	66.1
SMoG (Pang et al., 2022)	ResNet-50 (4×)	375	79.0
VicRegL (Bardes et al., 2022)	ConvNext-S	50	75.9
VicRegL (Bardes et al., 2022)	ConvNext-B	85	77.1
Ours	ResNet-50	27.9	**83.94**

Table 3.1: Top-1 accuracy results on several representation learning methods with 200 epochs, 128 batch size, and non-linear projection head on ImageNet dataset. (2×) means the kernel size.

A key advantage of our proposed method lies in its ability to extract robust features from a small dataset. Table 3.3 highlights the performance of self-supervised batch curation methods pre-trained on CIFAR10 and tested on the same dataset. In contrast to other algorithms, our approach allows the model to acquire a resilient representation without necessitating extensive epoch training, and data augmentation configuration for each specific dataset. The Patch Curation algorithm (Welle et al., 2021), utilizing Euclidean distance for batch selection, is surpassed by our method, particularly in cases where adjacent transformations construct patches without intersection.

This implies that the Patch Curation algorithm may inadequately address false positives or negatives by relying solely on distance measurements within a batch. Conversely, our FRD batch curation method leverages pairwise data distribution analysis to effectively eliminate false negatives and false positives in representations. Eventually, contrastive learning is often used for transfer learning, where a large, unlabelled dataset is available for pre-training, and a limited dataset is available for the downstream task. This raises the question of how well the proposed approach performs in this scenario.

Method	Batch Size	Epochs	Top-1
Supervised	128	200	**99.87**
CaCo (X. Wang et al., 2023)	128	200	92.6
ReSSL (Zheng et al., 2021)	256	200	89.37
SimSiam (X. Chen & He, 2021)	128	200	70.0
SimCLR (T. Chen, Kornblith, Norouzi, & Hinton, 2020)	128	200	62.5
SimSiam(X. Chen & He, 2021)	512	800	91.8
SimCLR (T. Chen, Kornblith, Norouzi, & Hinton, 2020)	256	200	87.5
DINO (Caron et al., 2021)	1024	800	99.1
Mixed Barlow Twins (Bandara et al., 2025)	128	2000	92.58
Ours	128	200	**99.67**

Table 3.2: These are top-1 accuracy scores for the linear classifier testing on CIFAR10.

In table 3.4, we compare the performance of our method against SimCLR, both pre-trained on ImageNet, against a fully supervised ResNet-50 trained on the downstream task. In table 3.4, we present the fine-tuning results of our algorithms with an increased batch size (instead of 128). The table demonstrates that our method can learn representations even with larger batch sizes. When compared across various datasets, our algorithm consistently outperforms large-scale image datasets such as Flower102 and small-scale datasets like STL10 and CIFAR10. Notably, even in the case of the greyscale MNIST handwritten dataset, our method exhibits only marginally inferior performance compared to supervised results. This underscores the effectiveness of our approach in achieving superior performance across diverse datasets, demonstrating its potential to enhance generalization ability without the need for an extensive number of samples or large batch sizes.

Method	FRD	Data Augmentation	Loss	Top-1
PatchBatch (Welle et al., 2021)	✗	Intersection Crop	NT-Xent	75.06
PatchBatch (Welle et al., 2021)	✗	Adjacent Crop	NT-Xent	76.31
SimCLR	✗	Adjacent Crop	NT-Xent	65.96
Ours	✗	Adjacent Crop	Reg. NT-Xent (Huber)	73.99
Ours	✗	Adjacent Crop	Reg. NT-Xent (L1)	71.56
Ours	✗	Adjacent Crop	Reg. NT-Xent (L2)	72.87
Ours	✓	Adjacent Crop	Reg. NT-Xent (Huber)	**83.92**
Ours	✓	Adjacent Crop	Reg. NT-Xent (L1)	70.87
Ours	✓	Adjacent Crop	Reg. NT-Xent (L2)	68.49

Table 3.3: These are top-1 accuracy scores for the k-Nearest Neighbour (k-NN) classifier, and the dataset represents self-supervised pertaining which is CIFAR10, testing on CIFAR10.

Method	CIFAR100	STL10	Flower102	Caltech101	MNIST
Supervised	**94.22**	83.26	93.34	86.07	95.56
SimCLR	85.9	-	97.0	**92.1**	-
Ours	91.8	**87.74**	**99.31**	91.42	**96.51**

Table 3.4: Comparison of transfer learning performance of the methods with several image datasets.

Table 3.4 reports the top-1 accuracy scores obtained from linear evaluation. For SimCLR, we use an ImageNet pre-trained ResNet50 encoder, which is then fine-tuned on each target dataset. During self-supervised training, we adopt a batch size of 128; for fine-tuning, we use a batch size of 1024 for CIFAR10, STL10, and MNIST, and a batch size of 256 for Caltech101 and Flower102. The original SimCLR paper does not explicitly specify the batch sizes used for the reported linear evaluation results on these datasets. For comparison, the supervised results reflect models trained directly on the target domain using standard training protocols.

Figure 3.4: Samples from the ImageNet dataset are used to show FRD scores for batches in both

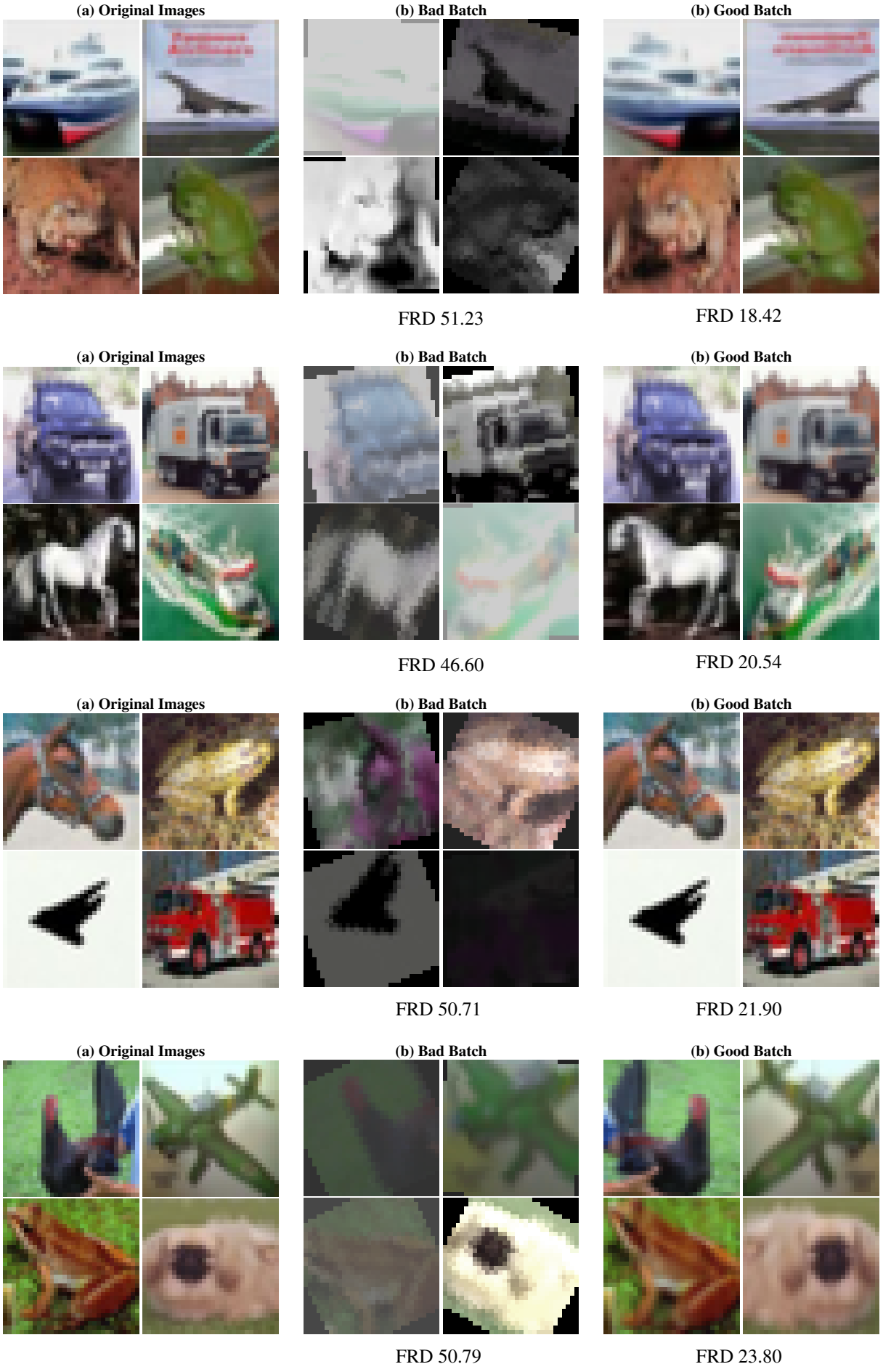


Figure 3.5: Samples from the CIFAR10 dataset are used to show FRD scores for batches in both

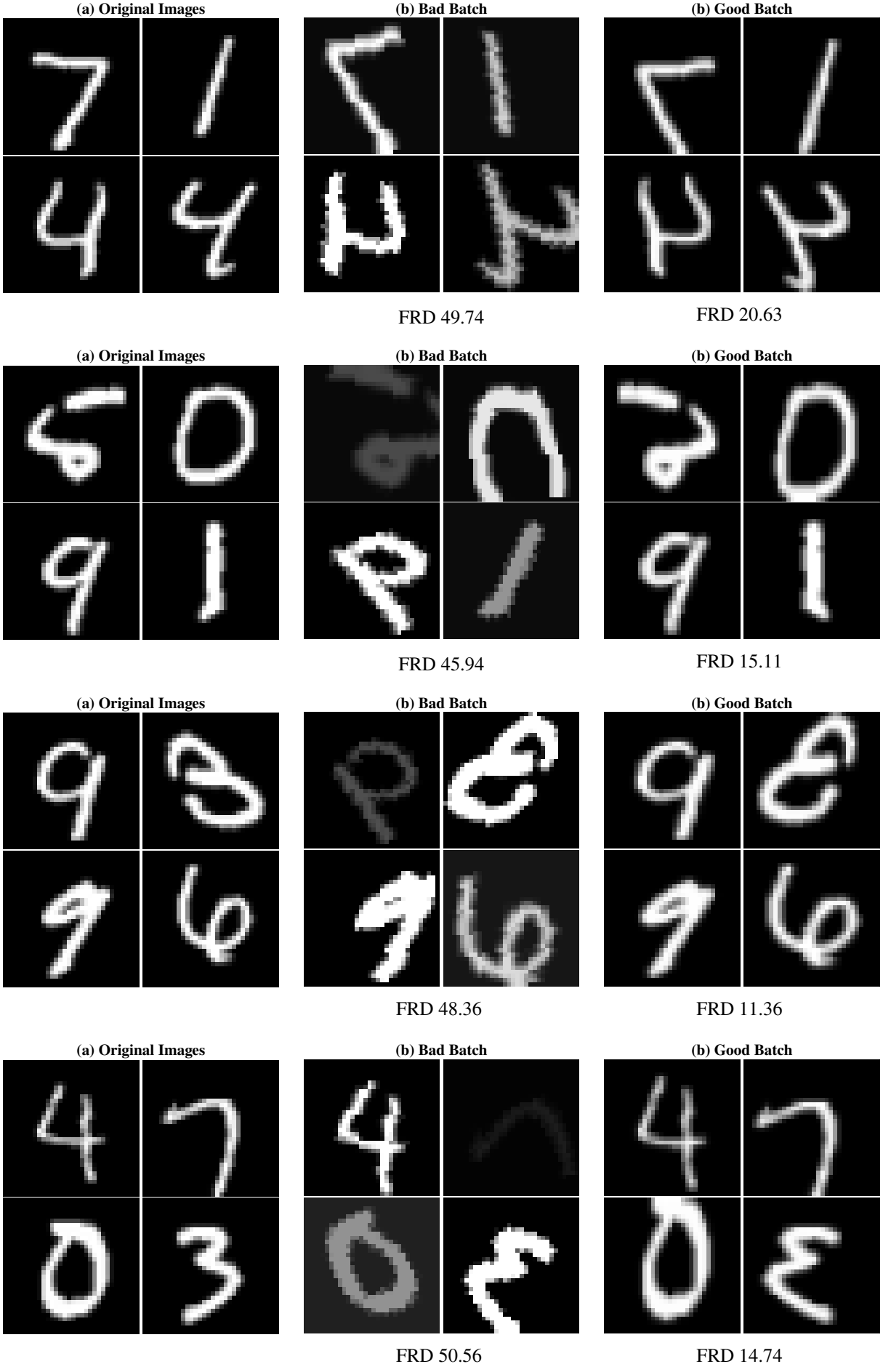


Figure 3.6: Samples from the MNIST dataset are used to show FRD scores for batches in both good and bad conditions.

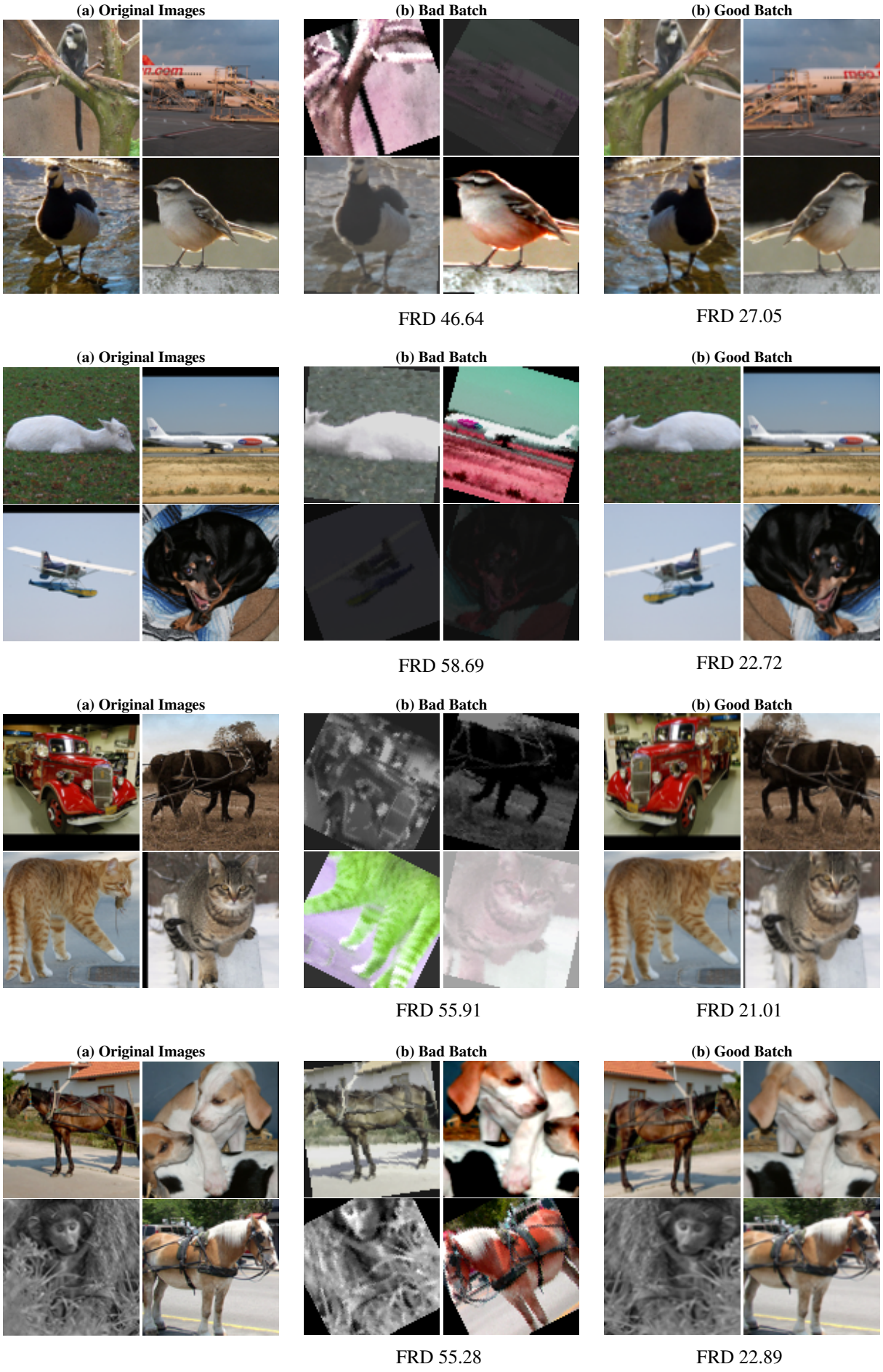


Figure 3.7: Samples from the STL10 dataset are used to show FRD scores for batches in both good and bad conditions.

Figures 3.7, 3.6, 3.5, and 3.4 present the FRD scores alongside the augmented versions of the original images for four datasets, using a batch size of 4. Even with small batch sizes, strongly distorted augmentations may be assigned as positive pairs despite being visually and semantically inconsistent with the original sample. The figures show that such mismatched pairs produce substantially higher FRD scores, reflecting the lack of alignment between their feature distributions. In contrast, semantically aligned positive pairs yield much lower FRD scores. This discrepancy highlights how incorrect positive assignments (**bad batches**) can mislead the representation learning process by enforcing similarity between features that should not be similar.

3.5 Discussion

The experimental results provide several insights into the contribution of each component of the proposed framework and allow us to revisit the hypotheses formulated at the beginning of this chapter.

Impact of the Huber loss. Across all datasets, the Huber loss consistently improves performance relative to the ℓ^2 loss used in standard contrastive learning. This suggests that outliers, instances in which augmented views differ substantially from the underlying semantic content, occur frequently enough to degrade learning when quadratic penalties are applied. The Huber loss reduces the influence of such outliers by transitioning to a linear penalty for large residuals, leading to more stable optimisation and more coherent feature clusters (as evidenced in the t-SNE plots or results at the table 3.3). This empirical behaviour supports our hypothesis that robust similarity constraints are necessary to counteract the effects of false positives and noisy augmentations.

Impact of FRD-based batch curation. FRD plays a complementary role by filtering out batches containing weak or semantically inconsistent augmentations before gradient updates occur. Our results show that batches with high FRD scores consistently correlate with poor alignment between augmented views (bad batches), confirming that FRD serves as a reliable indicator of augmentation quality. The threshold, estimated from early-epoch batch statistics, proves stable across datasets: small variations in the threshold do not significantly alter performance, indicating that the method is not overly sensitive to its precise value.

Transfer learning behaviour. The combination of FRD-based batch curation and Huber loss regularisation yields the most notable gains in transfer learning settings. Models trained with both components transfer more effectively across datasets with different visual characteristics (e.g., MNIST $\rightarrow$ CIFAR10, STL10 $\rightarrow$ Caltech101). This suggests that reducing the influence of

misaligned augmentations, both at the batch and representation levels, encourages the model to learn more general and stable features, rather than overfitting to dataset-specific distortions.

Relation to hypotheses and research questions. These observations collectively support our original hypotheses regarding the importance of both (i) robust similarity constraints and (ii) distribution-based batch curation in self-supervised contrastive learning. Specifically, the results provide partial answers to RQ1 (*How can we improve self-supervised contrastive learning under limited data conditions, where augmentations may introduce misleading (false positive/negative) views?*), showing that high-quality representations can be obtained without large batch sizes, extensive data augmentation configurations, or reliance on ImageNet-scale pretraining. The framework addresses the stated concerns around false positives and augmentation instability and demonstrates that statistically guided batch selection can meaningfully improve representation learning.

Overall, the interplay between FRD and the Huber loss explains the observed gains: FRD prevents unstable batches from influencing the optimisation dynamics, while the Huber loss mitigates residual outliers within accepted batches. Together, these mechanisms produce more discriminative, transferable, and semantically consistent representations across diverse datasets. A key limitation of the proposed method is the additional computational cost introduced by FRD computation and repeated batch curation. Because the model may reject multiple batches before identifying semantically consistent positive pairs, the training process can become slower, particularly in datasets where augmentations are highly variable. Optimising or approximating the FRD step could improve efficiency without losing its stabilising effect.

3.6 Conclusion

Self-supervised learning has demonstrated notable success across various tasks, exhibiting comparable performance to supervised learning methodologies. Numerous state-of-the-art algorithms have employed randomly curated batches in their training processes. However, random batch selection and augmentation may introduce many misleading pairs, including false positives and false negatives within batches. Furthermore, the contrastive loss alone does not adequately address the challenges related to false positives and negatives in batches, instead relying on large batch size to mitigate the issue.

In this study, we described a simple approach to address these limitations, mitigate the impact of false positives and quantify the distributions of similar and dissimilar pairs by evaluating the FRD score. The experiments demonstrate a markedly better convergence and overall performance of the learnt representations on a variety of downstream tasks, while necessitating smaller batch sizes and training epochs.

In conclusion, our findings underscore the importance of good batch selection in self-

supervised contrastive learning, particularly because the occurrence of bad augmentation is unavoidable in practice on large, varied datasets without a lengthy and expensive hyperparameter tuning of the data augmentation. We also note that the approach is generic and could be extended to other contrastive learning approaches.

Chapter 4

Robust Federated Learning in the Face of Covariate Shift

This chapter draws heavily from our paper titled *Hybrid-Regularised Magnitude Pruning for Robust Federated Learning under Covariate Shift*, presented at the International Symposium on Edge Intelligence, Trustworthy and Decentralised Artificial Intelligence (iEDGE 2025) as *Goksu, O. and Pugeault, N. (2025)*. Within this chapter, we introduce a novel attribute-based gender classification dataset called *CelebA-Gender*, the point of the dataset is modelling shifts in within-class distributions. Additionally, we introduce a novel method based on parameter selection to tackle the challenges associated with highly diverse and distributed datasets. By identifying and eliminating redundant parameters that overfit local clients' distributions and impede global convergence, we enhance the stability and performance of federated learning in non-i.i.d. environments. The code is publicly available. [1]

4.1 Introduction

Federated Learning (FL) offers an efficient paradigm for collaboratively training a shared model across multiple users while preserving data in each local device. Thus, FL is particularly well-suited for real-world applications in which data stored across many servers is critical, such as medical imaging, remote sensing, or processing personal data (e.g., faces). The FL process is typically initiated by broadcasting an initialised global model from the server to participating clients. Each client independently updates the model using its local dataset, and the resulting parameters are transmitted to a central server, where local models' parameters are aggregated to update and refine the global model Gao et al., 2022; McMahan et al., 2017; Sharma et al., 2022. In principle, the global model in federated learning is expected to outperform individual local models by leveraging the diversity of client data to achieve improved generalisation. This relies on the assumption that local updates can be aggregated effectively to produce a stable and high-

[1] https://github.com/ozgugoksu/FederatedLearning/

performing global model. However, non-i.i.d. data often induce substantial model drift, resulting in divergent local optima that undermine the stability and effectiveness of global aggregation. This challenge is closely related to **covariate shift**, where the input feature distributions vary across clients while the underlying task remains the same. Effectively addressing covariate shift is therefore central to enhancing the robustness and generalisation of federated learning in heterogeneous environments.

Overcoming these limitations is essential for making FL practical and robust in privacy-sensitive and data-constrained environments. To address this challenge, in this chapter, we explore the question: *"How can covariate shift in federated learning be mitigated to enable robust and generalizable training when clients possess limited and non-overlapping datasets?"*. We hypothesise that an effective strategy involves eliminating irrelevant or non-transferable parameters for each client, allowing models to focus on client-specific patterns while still contributing to a global representation.

In the context of federated learning (FL), non-transferable parameters are model weights that capture patterns specific to a particular client's local data but fail to generalise effectively to other clients' datasets. These parameters often overfit to local cases, such as biases or unique distributions present only in a single client's data. When aggregated into a global model, they can impede convergence and reduce overall performance, as the global model benefits most from parameters that capture knowledge transferable across clients. For instance, in a CIFAR10 classification setting, parameters that specialise in recognising certain vehicle images present only in one client's local dataset may not generalise to images of the same class from other clients, rendering them non-transferable. Identifying and mitigating such parameters allows federated models to focus on representations that are both client-specific and globally relevant.

This perspective assumes that not all weights in a shared model are equally beneficial across heterogeneous data distributions. By identifying and filtering out such redundant parameters, we aim to enhance both local adaptation and global collaboration in highly skewed and non-i.i.d. federated environments.

We propose FedMPR (**F**ederated Learning with **M**agnitude **P**runing and **R**egularization), a novel framework designed to enhance federated learning under data heterogeneity. FedMPR promotes robustness in local models by integrating three key components: (1) magnitude-based pruning to eliminate redundant parameters at the client level; (2) dropout to introduce functional redundancy in decision pathways; and (3) noise injection to regularise model responses and improve resilience to weight perturbations during aggregation. We demonstrate that this framework outperforms standard FL approaches on benchmarks, in particular in datasets with large covariate shifts between clients.

Additionally, we introduce CelebA-Gender, a novel benchmark dataset for heterogeneous FL. Derived from CelebA Z. Liu et al., 2015, it is specifically designed to evaluate FL methods under challenging conditions where inter-client distribution shifts arise not only from class imbalance,

but also from substantial within-class distribution variations. Client data shift in our study arises from attribute-based gender classification, unlike existing literature Caldas et al., 2018, which typically focuses on examining one attribute at a time for binary classification (smiling/not smiling). CelebA-Gender benchmark provides a more complex data distribution, enabling the evaluation of varying levels of attribute overlap in the image content, ranging from high to low overlap scenarios.

Our framework presents several key contributions:

- **FedMPR:** We propose a novel framework for addressing scenarios with significant covariate shifts across clients' data distributions.
- **Data Heterogeneity Scenarios:** We evaluate FL approaches under different levels of covariate shift, testing both low and high shift scenarios, including varying numbers of clients (both limited and large populations), as well as imbalanced and limited data, to assess the adaptability and robustness of FL methods across diverse settings.
- **Novel Classification Dataset:** We introduce CelebA-Gender, a reconstructed gender classification dataset derived from attribute-based labels, designed to facilitate a comprehensive evaluation of our framework and enable comparison with existing methods.

In summary, this chapter presents FedMPR, a robust federated learning framework designed to tackle covariate shifts across clients, along with the *CelebA-Gender* benchmark for low to high numbers of clients. The subsequent sections provide a detailed discussion of related work, formal problem formulation, methodology, and experimental evaluation, highlighting the advantages of our approach in heterogeneous federated learning scenarios.

4.2 Background

4.2.1 Federated Learning

FL frameworks such as FedAvg typically involve three main steps: *broadcasting*, *local training*, and *model aggregation*. After each communication round, the central server broadcasts the current global model to all participating clients. Each client then performs local training on its private data, ensuring that raw data is on-device. Once local updates are completed (e.g., after a fixed number of epochs), clients send their updated model parameters back to the server. The server performs model aggregation, typically by averaging the received parameters, to form a new global model, which is then broadcast in the next round. Following the typical FL approach, the data D is distributed across K clients. Let D_k be the local dataset at client k, with $n_k = |D_k|$ denoting the number of data points at client k. The global objective function in FL is then a

weighted average of the local objectives

$$\min_{w} F(w) = \sum_{k=1}^{K} \frac{n_k}{n} F_k(w) \tag{4.1}$$

where $n = \sum_{k=1}^{K} n_k$ is the total number of data points across all clients and $F_k(w)$ is the local objective function at a client k

$$F_k(w) = \frac{1}{n_k} \sum_{i \in \mathcal{D}_k} f_i(w) \tag{4.2}$$

Current research predominantly relies on FedAvg-based model aggregation and training procedures Oh et al., 2021, Tamirisa et al., 2024, Q. Li et al., 2021 that can broadly be categorised into two main directions: enhancing local training and refining the model aggregation strategy. Our work aligns with the first direction, targeting the local training phase. By improving local optimisation, we aim to reduce divergence across client models, thereby indirectly mitigating challenges typically addressed at the aggregation stage.

4.2.2 Related Work

Federated Parameter Selection

FL approaches encounter several critical challenges, particularly poor convergence on highly heterogeneous data and the lack of solutions for individual clients. To address these issues, parameter selection or decoupling FL approaches introduce tailored models for heterogeneity Oh et al., 2021, Tamirisa et al., 2024, A. Z. Tan et al., 2022, Y. Jia et al., 2024. Among these approaches, some studies Oh et al., 2021, Tamirisa et al., 2024 manually partition the model into personalised and shared parameters. Personalised methods can address the data heterogeneity. When clients have imbalanced or highly distributed data, these algorithms cannot learn robust features. However, they do not provide a joint global model, therefore a significant question remains unresolved: how to effectively eliminate redundant parameters in local models, especially when there are relatively few clients and substantial data heterogeneity among them, while training a joint global model. Covariate shift occurs when local training data across clients is highly diverse, exhibiting little to no overlap in their respective data distributions. Consequently, the global model often fails to generalise effectively, as federated learning typically assumes that the training and test data distributions for each client are identical, a condition rarely met in real-world scenarios.

Regularisation in Federated Learning

FedProxSharma et al., 2022 introduces a proximal term in the loss function that regularises the local model updates. This regularisation term limits the local updates from diverging too far from the global model, addressing issues such as the partial participation of clients in FL. SCAFFOLD

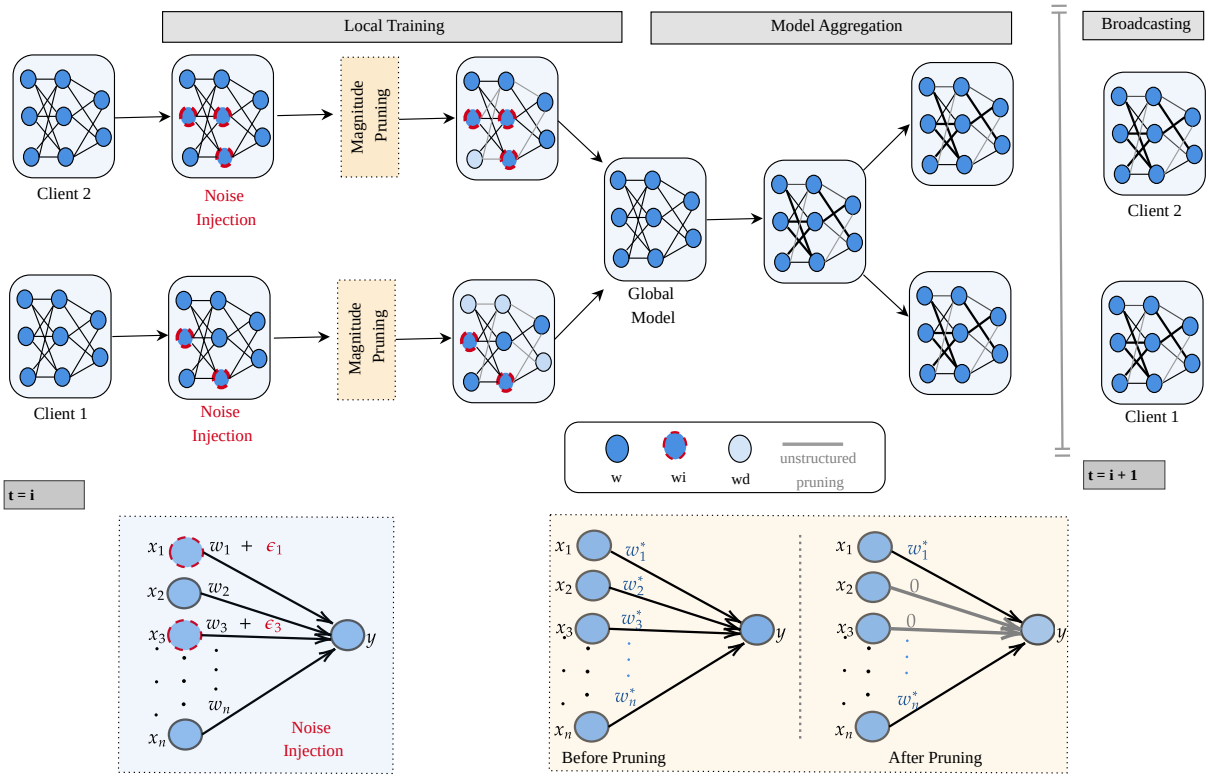


Figure 4.1: FEDMPR framework, each client applies a tailored regularisation that combines dropout during forward passes with Gaussian noise injection inside each basic block. At the ith iteration, the central server broadcasts the global model weights to clients, which then perform local training before model aggregation. Specifically, w denotes the original weights, w_i the weights perturbed by Gaussian noise, and w_d the zeroed (pruned) weights.

Karimireddy et al., 2020 introduces a control variate-based regularisation method to correct for client covariate shifts in non-i.i.d.. The work Lee and Choi, 2024 estimates weight distribution regularisation for each client using the FedAvg. Consequently, the regularisation-based FL approaches D. Wen et al., 2022 may suffer from real-world applications where the classes are completely different per client, leading to instability in distribution estimation. MOON Q. Li et al., 2021 reformulates contrastive loss to leverage the global model, which captures representations from the entire data distribution. It builds on a similar principle to FedProx by constraining local updates concerning the global model. In contrast, our proposed method trains each client model and eliminates redundant parameters, thereby reducing the impact of updated parameters on local training and global model aggregation, leading to better convergence.

4.3 Methodology

4.3.1 FedMPR

Unlike Personalised Federated Learning (PFL), where clients are allowed to fine-tune or adapt a local model to their specific data, our approach adheres to the standard FL framework. In this framework, all clients collaboratively train a single shared global model without performing any local modifications or customisations. Each client only applies the global model as is, without any modifications to enhance it to fit its local data. We reformulate Iterative Magnitude Pruning (IMP) Frankle and Carbin, 2019 (Algorithm 3 summarises our federated iterative magnitude pruning procedure) by introducing a round-wise strategy, enabling progressive scarification across rounds instead of applying it only once per round. Our approach enhances the standard FL framework by mitigating the adverse effects of model aggregation. As illustrated in Figure 4.1 (two-client example), and Algorithm 2 the proposed method introduces three key components:
Pruning: In neural networks, weights with small magnitudes often contribute minimally to the model's output S. Han et al., 2015. While such parameters may have a negligible impact in centralised training, in FL, they can amplify aggregation misalignment under data heterogeneity.
Federated Iterative Magnitude Pruning: In the context of federated learning, we adapt Iterative Magnitude Pruning (IMP) to operate at the level of *communication rounds* rather than during continuous local training. Specifically, in our approach, each client performs local training for a fixed number of epochs using its local data, and the resulting parameters are transmitted to the server for aggregation. After the global model is updated, we apply magnitude pruning to the local model weights before the next communication round. This process is repeated until the *desired sparsity level* is reached, after which no further pruning is applied in subsequent rounds.

Formally, let $\theta_c^{(r)}$ denote the parameters of client c at communication round r, and let $\theta_G^{(r)}$ be the aggregated global model. Our method proceeds as follows:

1. **Local training:** Each client c updates its model on local data for E epochs to obtain $\theta_c^{(r)}$.

2. **Magnitude pruning:** After training, each client applies a pruning mask based on parameter magnitude:

$$\theta_{c,i}^{(r+1)} = \theta_{c,i}^{(r)} \cdot m_{c,i}^{(r)}, \quad m_{c,i}^{(r)} = \begin{cases} 0 & |\theta_{c,i}^{(r)}| \leq \tau^{(r)} \\ 1 & \text{otherwise} \end{cases} \tag{4.3}$$

   where $\tau^{(r)}$ is the pruning threshold at round r.

3. **Global aggregation:** The server aggregates the local models to form the global model $\theta_G^{(r)}$:

$$\theta_G^{(r)} = \frac{1}{C} \sum_{c=1}^{C} \theta_c^{(r)} \tag{4.4}$$

   where C is the total number of clients.

4. **Repeat:** Local training resumes on the pruned model for the next communication round. Once the target sparsity is achieved, no further pruning is performed in subsequent rounds.

This *round-based iterative pruning* ensures that clients progressively remove non-transferable parameters while allowing the global model to aggregate robust, shared representations. By performing pruning only after aggregation, the approach stabilises convergence, mitigates overfitting to local distributions, and reduces communication overhead by transmitting only relevant weights.

Redundancy: Dropout applied during local training induces redundancy by encouraging diverse subnetworks D. Wen et al., 2022, which can improve the robustness of model aggregation under non-i.i.d. conditions by mitigating client-specific overfitting and reducing the variance in local updates.

Robustness: Injecting noise during local training enhances robustness to small weight perturbations introduced during aggregation. Additionally, applying dropout promotes redundancy in neural pathways, further improving the model's tolerance to aggregation noise. Together, these techniques help mitigate client-side overfitting in federated settings.

Input: Number of clients n; total rounds T; pruning frequency f; threshold θ; initial global model w_0; prune percentage p; sparsity β
Output: Final global model w_T
Partition $\mathcal{D}$ into $\{\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_n\}$
for *each client* $c_i \in \{1, \ldots, n\}$ **do**
  Initialize c_i with w_0
  Initialize m_i masks
end
for *each round* $t = 1$ **to** T **do**
  Server broadcasts w_{t-1} to clients $c \in C_t$
  foreach *client* $c \in C_t$ *in parallel* **do**
    Apply mask: $w_c^t \leftarrow m_c \odot w_{t-1}$
    Train w_c^t on $\mathcal{D}_c$
    if *density of* $w_c^t > \beta$ **then**
      $w_c^t \leftarrow \text{PruneWeights}(w_c^t, p)$
    end
  end
  Clients send w_c^t to server
  Server aggregates: $w_t \leftarrow \sum_{c \in C_t} \frac{n_c}{n_{C_t}} w_c^t$
end
return w_T

Algorithm 2: Federated Magnitude Pruning and Regularisation (FEDMPR)

Input: Model parameters W, prune percentage p
Output: Pruned model parameters W'
foreach *parameter tensor* w *in* W **do**
  if w *is weights* **then**
    $T \leftarrow \text{flatten}(w)$
    $A \leftarrow |T|$
    $\tau \leftarrow \text{percentile}(A, p \times 100)$; `// Threshold`
    foreach $a_i \in A$ **do**
      if $a_i > \tau$ **then**
        $mask_i \leftarrow 1$
      else
        $mask_i \leftarrow 0$
      end
    end
    $w' \leftarrow T \odot mask$; `// Element-wise`
    reshape w' to shape of w
    replace w in W with w'
  end
end
return W'

Algorithm 3: PruneWeights: Iterative Magnitude Pruning

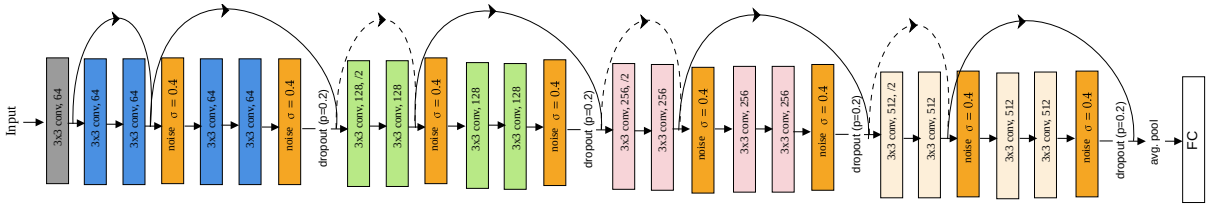


Figure 4.2: ResNet-18 model architecture with regularisation layers integration.

4.3.2 Federated Learning with Varied Data Distributions

Covariate shift, a central form of statistical heterogeneity in FL, arises from variations in input feature distributions across clients. While the Dirichlet distribution Minka, 2000 is commonly used to simulate non-i.i.d. conditions Tamirisa et al., 2024, Oh et al., 2021, it introduces both class imbalance and data overlap. In this work, we explore alternative partitioning strategies to model covariate shifts more explicitly and also consider scenarios with highly similar client distributions.

Dirichlet Distribution: This distribution allows for controlled variation in data partitioning, provides many heterogeneous client data scenarios by α parameter, which controls the imbalance across clients like previous studies Tamirisa et al., 2024, Q. Li et al., 2021. Smaller α values like 0.1 lead to more skewed, imbalanced, non-i.i.d. data partitions. In our experiments, we use 0.1 and 0.5 with 10 and 100 clients.

Low versus High Covariate Shift conditions: One limitation of using the Dirichlet distribution to control the client data distribution is that the training sets are not mutually exclusive, and therefore, multiple clients will see the same training examples. Dirichlet distribution models label scarcity or imbalance across clients, but does not explicitly account for covariate shift, differences in the input feature distributions between clients. In other words, while it simulates heterogeneous labels, it assumes that the underlying feature distribution for each class remains similar across clients. This is unlike any real scenario where each client would hold completely distinct data. Therefore, in addition to the Dirichlet distribution, we experimented with mutually exclusive and inclusive client data in two conditions, denoted as low covariate shift (low-CS) and high covariate shift (high-CS) in the paper. In the low-CS condition, all clients receive an equal number of training examples from each target class, where each example is only allocated to one client. Let D_1, D_2 be client datasets given by

$$D_1 = \{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\} \tag{4.5}$$

$$D_2 = \{(\bar{x_1}, y_1), (\bar{x_2}, y_2), \cdots, (\bar{x_n}, y_n)\} \tag{4.6}$$

where (x_i, y_i) represents a data sample x_i with a corresponding class label y_i. $x_1, \bar{x_1}$ indicate the samples from each client's dataset belong to the same class label. $y_i \in Y$ represents the class labels present, both are equal to the set of class labels. In the high-CS condition, on the other hand, each client receives a subset of classes, for example, in a two-client setup on MNIST, client A would train on examples of the classes $y_1 \in Y_1 = \{0, 1, 2, 3, 4\}$, and client B on $y_2 \in Y_2 = \{5, 6, 7, 8, 9\}$. The class distributions can be represented as follows,

$$D_1 = \{(x_1, y_1), (x_2, y_1), \cdots\} \tag{4.7}$$

$$D_1 = \{(\bar{x_1}, y_2), (\bar{x_2}, y_2), \cdots\} \tag{4.8}$$

client 1 trains on the classes $y_1 \in Y_1 = \{0, 1, 2, 3, 4\}$, and client 2 trains on $y_2 \in Y_2 = \{5, 6, 7, 8, 9\}$.

4.3.3 CelebA-Gender Dataset

In this study, we introduce the CelebA-Gender dataset, designed to facilitate the evaluation of covariate shift ratios in real-world applications. While existing research predominantly relies on datasets like CIFAR10, which consists of entirely distinct classes (e.g., deer, cars), these scenarios do not fully capture the complexities of real-world data distributions. Specifically, in the context of facial recognition, the challenge of partitioning data to represent varying degrees of covariate shift remains largely unexplored. Moreover, different facial attributes exhibit distinct behaviours influencing feature learning, further underscoring the need for a dataset tailored to this nuanced setting.

Our proposed dataset addresses these gaps, offering a platform to study and quantify covariate shifts realistically. We begin our analysis with a dataset containing one attribute and progressively expand to a seven-attribute version. The selection of attributes is carefully optimised to maximise the number of samples in each dataset version, ensuring a balance between high-covariate and low-covariate conditions. This approach minimises the risk of overfitting while maintaining balanced data distributions across the different configurations.

We introduce CelebA-Gender, a novel gender classification dataset derived from CelebA Z. Liu et al., 2015 under Federated Learning evaluation. The CelebA dataset has previously been adapted to the federated learning setting as part of the LEAF benchmark Caldas et al., 2018. In that formulation, the evaluation is restricted to binary classification tasks based on a single facial attribute (i.e., "Smiling" vs. "Not Smiling"). While this setup enables controlled experimentation, it oversimplifies the complexity of real-world scenarios. In contrast, our proposed data partitioning extends the use of CelebA by constructing client datasets based on multiple overlapping attributes, including gender, and images with no prominent attributes. This results in more heterogeneous, non-trivial data distributions across clients, better reflecting the challenges of federated learning under realistic and complex conditions. More details are shown in the appendix section.

4.3.4 Experimental Setup

Model & **Datasets** We use ResNet-18 as the backbone architecture, initialising both global and local models randomly. As shown in Figure 4.2, we modify the standard ResNet-18 by incorporating noise injection (percentage 0.4), layers and dropout (ratio 0.2) within each basic block. For small-scale datasets such as CIFAR10 Krizhevsky, 2009, MNIST LeCun et al., 1998, Fashion-MNIST Xiao et al., 2017, and SVHN Netzer et al., 2011, we reduce the convolutional kernel size within the basic blocks to 3×3 to better fit the lower-resolution inputs. For larger-scale

datasets such as CelebA-Gender. [2] and RAF-DB S. Li et al., 2017, we retain the original kernel sizes.

Training Setup: We set the number of clients to 2 for both low and high covariate shift scenarios, allowing us to evaluate the performance with limited clients exhibiting higher similarity in their data, as well as vice versa. Additionally, we use the Dirichlet distribution with $\alpha = 0.1, 0.5$ for a higher number of clients (10 and 100). We employed Stochastic Gradient Descent (SGD) and SGD momentum is 0.9, a batch size is 128 to train each model, using a learning rate of 2e-2. Each local model was trained for 5 epochs per round, communication round is 100.

4.4 Results

Pruning methods with Lottery Ticket Hypothesis (LTH) perform better on many real-world applications, like PruneFL Y. Jiang et al., 2022, LotteryFL A. Li et al., 2020 and FedSelectTamirisa et al., 2024. Among them, we focus on FedSelect for comparison, as it incorporates a gradient-driven LTH strategy tailored for personalisation closely aligning with the sparsity mechanisms employed in our method. This enables a principled evaluation of selective parameter training within the FL framework. In Table 4.1, we present the performance of FL algorithms on a range of benchmarks, in the low-CS (as evidenced by low FID, CMMD scores between the two local data distributions). As inter-client dissimilarity increases, performance degrades significantly due to the absence of overlap between local data, which challenges standard model aggregation methods. In contrast, pruning-based approaches such as FedSelect with 67.45% and FedMPR with 69.41 % on RAF-DB data, demonstrate improved robustness by extracting more generalizable features from client data. As shown in Table 4.1(b), they achieve notably higher accuracy (around 25 % higher than FedAvg on RAF-DB data), particularly on the CIFAR10 and CelebA-Gender datasets. FedMPR consistently outperforms all baselines across datasets, indicating that leveraging model redundancy enhances performance under both balanced and imbalanced data distributions.

When selecting the backbone model, we performed hyperparameter tuning on several architectures, including SimpleCNN, ResNet18, and ResNet50. As shown in Tables 4.2 and 4.3, ResNet18 achieved the highest accuracy on the CelebA-Gender dataset. ResNet-18 outperforms Simple CNN and ResNet-50 on the CelebA-Gender due to its optimal combination of expressive capacity and optimisation stability. ResNet-50 overfitted for the CelebA-Gender dataset; however, the Simple CNN insufficiently underfitted to capture fine-grained facial features. ResNet-18, which is fairly deep and computationally efficient, converges more reliably and generalises better when client datasets have been significantly skewed and attribute-filtered. Therefore, we adopted ResNet18 as the primary backbone for the remaining experiments.

Figures 4.3 illustrate the impact of the number of samples per class on representation learning

[2] https://anonymous.4open.science/r/Dataset-E468/README.md

Method	CIFAR10	MNIST	FMNIST	SVHN	CelebA-Gender	RAF-DB
Supervised	**91.51**	**99.08**	**94.18**	**93.89**	**98.39**	**77.80**
FedAvg	77.00±0.33	99.01±0.04	90.73±0.31	91.11±0.37	91.18±8.23	43.42±1.78
FedProx	76.45±0.89	98.94±0.16	91.31±0.54	91.07±0.17	72.05±3.86	64.31±1.45
FedDC	74.37±2.97	98.58±0.11	90.04±0.61	90.44±0.43	61.67±7.97	62.43±1.18
FedDyn	77.72±0.65	98.98±0.20	91.13±1.16	91.57±0.10	72.07±4.13	64.39±1.91
SCAFFOLD	76.73±0.28	98.77±0.25	90.87±0.39	90.97±0.05	72.13±4.12	64.51±1.59
FedSelect	76.25±1.17	94.91±2.16	89.78±2.58	90.32±1.16	93.14±3.64	72.74±1.48
FEDMPR	**86.61±1.26**	**99.49±0.07**	**93.10±0.46**	**94.17±0.21**	**96.93± 1.05**	**74.92±1.16**
FID Score	1.68	0.63	0.83	0.52	12.00	7.79
CMMD Score	∼0.00	∼0.00	∼0.00	∼0.00	∼0.00	∼0.00

(a) FL accuracy on *low-CS* condition for 2 clients. Lower FID or CMMD suggests higher data similarity between clients.

Method	CIFAR10	MNIST	FMNIST	SVHN	CelebA-Gender	RAF-DB
Supervised	**91.51**	**99.08**	**94.18**	**93.89**	**98.39**	**77.80**
FedAvg	51.28±1.55	89.84±1.31	80.69±1.30	75.69±3.41	47.82±1.93	45.06±1.39
FedProx	54.71±1.25	87.28±3.04	81.49±1.18	74.90±1.21	71.80±1.92	45.99±2.84
FedDC	41.09±1.55	65.62±9.79	60.11±6.75	59.02±2.47	62.77±6.10	40.55±0.82
FedDyn	49.48±0.72	85.53±2.05	79.33±2.55	69.71±2.77	70.50±3.16	43.12±1.26
SCAFFOLD	49.18±0.93	78.73±5.59	71.06±3.12	59.09±2.26	67.63±7.16	43.11±1.27
FedSelect	61.97±5.96	91.02±4.76	85.01±5.16	81.00±4.39	52.50± 4.40	67.45±4.38
FEDMPR	**75.22±5.35**	**98.99±0.34**	**88.92±2.24**	**88.24±4.67**	**84.23±1.22**	**69.41±5.54**
FID Score	70.69	43.14	47.49	4.52	82.85	9.88
CMMD Score	0.16	0.20	0.10	0.12	1.52	0.17

(b) FL accuracy on *high-CS* condition for 2 clients. Higher FID and CMMD values indicate greater data heterogeneity.

Table 4.1: Comparison of global model accuracy (%) under (a) low and (b) high-CS settings. Average test accuracy over three runs. (±) denotes the standard deviation from the average.

Attributes	Simple-CNN	ResNet18	ResNet50
1	84.45	92.66	89.74
2	84.34	93.74	90.06
3	84.56	**93.95**	91.79
4	85.64	92.33	91.47
5	78.62	89.09	88.98

Table 4.2: CelebA-Gender (ME) accuracy with several models.

Attributes	Simple-CNN	ResNet18	ResNet50
1	90.5	95.25	93.84
2	87.58	94.17	94.92
3	88.77	95.46	95.36
4	90.71	96.44	95.79
5	93.95	**98.49**	96.65

Table 4.3: CelebA-Gender (MI) accuracy with several models.

with two clients under high and low covariate shift settings. As the number of samples per class increases, performance consistently improves in both scenarios, regardless of the degree of distributional shift. Table 4.4 demonstrates the Dirichlet-based heterogeneous data distribution, whereas Table 4.1 summarises the label-skew, overlap, and non-overlap situations used in our evaluation. Clients show significant data heterogeneity when using the Dirichlet partitioning strategy, especially when $\alpha = 0.1$ is used.

As expected, our approaches perform slightly lower in this scenario due to a significant label imbalance among clients. In contrast, the high-CS situation has only two clients and hence represents a simpler distributional change. Increasing the number of clients under strong non-i.i.d. conditions makes the learning problem significantly more difficult, which explains the performance degradation observed in the Dirichlet $\alpha = 0.1$ experiments.

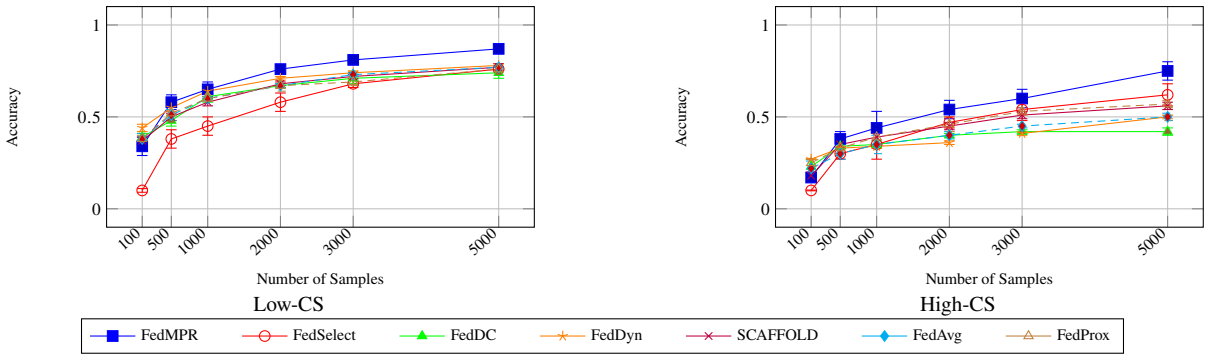


Figure 4.3: Top-1 accuracy on CIFAR10 under covariate shifts with two clients and varying local sample sizes.

Dataset	α	Clients	FedAvg	FedProx	FedDC	SCAFFOLD	FedDyn	FedSelect	FedMPR
CIFAR10	0.1	10	65.86	65.55	67.12	74.35	**74.39**	54.88	72.82
		100	27.95	27.74	30.65	30.1	30.98	18.45	**32.19**
	0.5	10	81.07	81.15	78.04	81.18	81.77	67.78	**88.55**
		100	39.62	40.01	40.91	39.08	41.58	19.16	**49.91**
RAF-DB	0.1	10	41.82	41.36	39.77	40.35	43.68	64.34	**67.73**
		100	17.63	18.29	18.38	18.58	19.04	39.86	**44.85**
	0.5	10	43.87	44.23	49.19	41.53	45.83	73.08	**74.32**
		100	41.66	42.29	34.88	41.29	39.18	42.73	**44.07**
CelebA	0.1	10	56.5	56.46	50.0	57.06	50.7	50.12	**68.85**
		100	47.32	48.06	49.9	49.96	48.1	**50.1**	48.88
	0.5	10	83.76	80.8	55.13	80.62	50.62	85.32	**90.05**
		100	48.68	49.1	53.56	48.42	50.0	**65.8**	49.58
CelebA-Gender	0.1	10	70.2	71.0	52.2	66.19	65.4	**72.9**	65.11
		100	50.8	54.88	49.0	42.6	49.2	49.04	**58.36**
	0.5	10	79.2	**79.8**	57.8	78.8	71.4	77.58	76.5
		100	53.6	**62.8**	53.4	59.2	58.2	51.0	52.76

Table 4.4: Accuracy (%) across datasets for varying $\alpha \in \{0.1, 0.5\}$.

Pruning-based methods, such as FedMPR, demonstrate strong scalability under extreme data imbalance, outperforming baseline approaches in settings with 100 clients and a Dirichlet concentration parameter of $\alpha = 0.1$ (Table 4.4). By promoting sparsity and applying regularisation during local training, these methods not only improve generalisation but also mitigate the adverse effects of model aggregation in non-i.i.d. environments. t-SNE plots in Figures 4.4 demonstrate that magnitude pruning effectively captures robust features across multiple datasets. However, FedSelect, as a personalised FL (PFL) method, exhibits superior performance under high client heterogeneity. RAF-DB is inherently imbalanced, unlike other datasets that originally contain a uniform number of samples per class, without requiring synthetic non-i.i.d. partitioning. Applying a Dirichlet distribution to such balanced datasets introduces severe non-i.i.d. conditions and

significant class imbalance, making representation learning substantially more challenging. Nevertheless, our proposed method demonstrates strong robustness to this type of data heterogeneity, effectively handling both natural and induced imbalances.

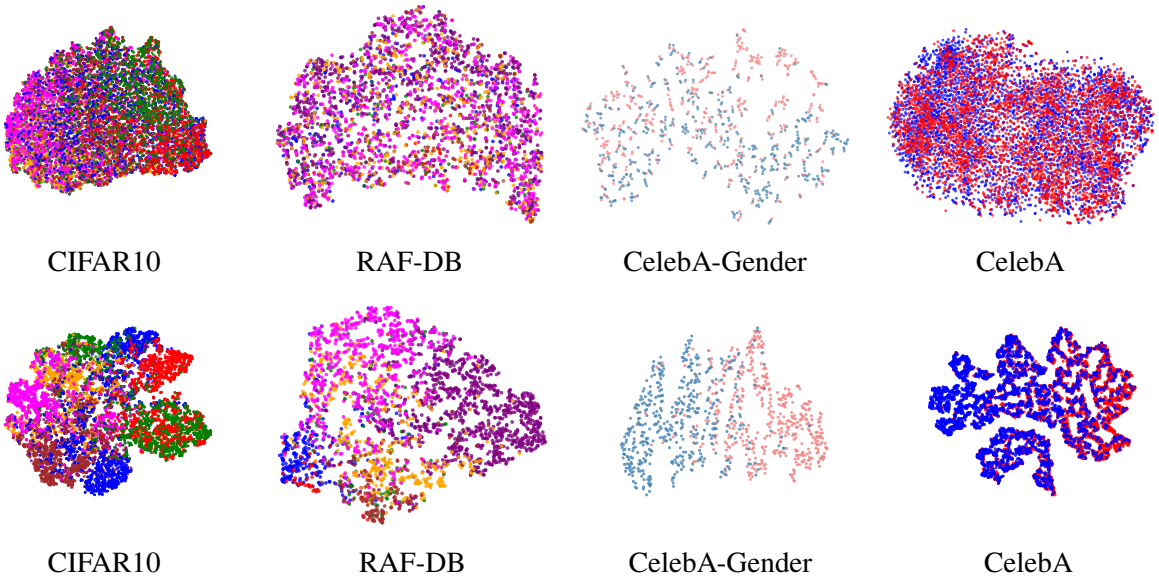


Figure 4.4: t-SNE plots under $\alpha = 0.1$. First row represents FedAvg, second row FedMPR.

The figure 4.5 shows some of the classification results among methods. FedMPR performs well on gender classification when we compare other methods.

4.5 Conclusion

Federated Learning (FL) enables decentralised model training without data sharing but often suffers from covariate shift due to inconsistent local updates. We introduce FedMPR, a pruning-based FL framework that integrates iterative unstructured pruning with dropout and noise injection to enhance robustness under data heterogeneity. FedMPR consistently outperforms state-of-the-art baselines across standard benchmarks.

To facilitate controlled evaluation, we also propose CelebA-Gender, a novel dataset designed to isolate covariate shift by modifying within-class distributions while maintaining class balance. Our findings demonstrate that structured pruning can effectively mitigate distributional shift without requiring explicit personalisation. Moreover, parameter selection plays a crucial role in federated learning, as it directly influences model convergence and generalisation across heterogeneous clients.

Figure 4.5: Classifier performance on CelebA-Gender test data, trained using CelebA-Gender (LC) with 5 attributes. Green indicates correct classification; Red indicates incorrect.

Chapter 5

Federated Metric Learning for Data Heterogeneity

The majority of this chapter is based on our paper entitled *FedQuad: Federated Stochastic Quadruplet Learning to Mitigate Data Heterogeneity*, published at The 3rd IEEE International Conference on Federated Learning Technologies and Applications (FLTA25) as *Goksu, O. and Pugeault, N. (2025)*. In this chapter, we introduce a federated metric learning based approach to address the problem of statistical data heterogeneity in federated learning. Metric learning enables the model to bring similar samples closer in the representation space while pushing dissimilar samples farther apart. Leveraging this property, we apply metric learning within the federated learning framework to mitigate differences in client data distributions. This chapter presents the evaluation of various metric learning algorithms under federated settings and details the proposed methodology developed to enhance model consistency and generalisation across clients. The code is publicly available. [1]

5.1 Introduction

As neural network architectures have grown deeper and more complex over the years, their training has come to require increasingly large and diverse datasets. While deep learning models achieve remarkable performance on large-scale datasets such as ImageNet J. Deng et al., 2009 and LAION Schuhmann et al., 2022, such extensive datasets are not always accessible. In many real-world scenarios, data is distributed across multiple locations, including university servers, research laboratories, and private institutions. This data decentralisation poses a major challenge for the practical deployment and scalability of deep learning models, as data cannot always be centralised due to privacy, security, or regulatory constraints. In addition, data storage in one place and no sharing data concerns are escalating, as many entities are unwilling to share their local data due to confidentiality issues. To address these challenges, *Federated Learning (FL)*

[1] https://github.com/ozgugoksu/FederatedLearning/

has emerged as a promising solution in recent years, enabling collaborative model training Mora et al., 2024, H. Zhu et al., 2021. FedAvg McMahan et al., 2017 is a fundamental algorithm in FL that trains a server across multiple clients. Each client preserves its local training dataset privately, ensuring that raw data is never shared with the server. Instead, clients compute model updates independently and transmit only these updates to the server, while enabling collaborative learning. FL is crucial for various high-impact applications, including medical imaging, remote sensing, and image classification, where decentralisation is paramount.

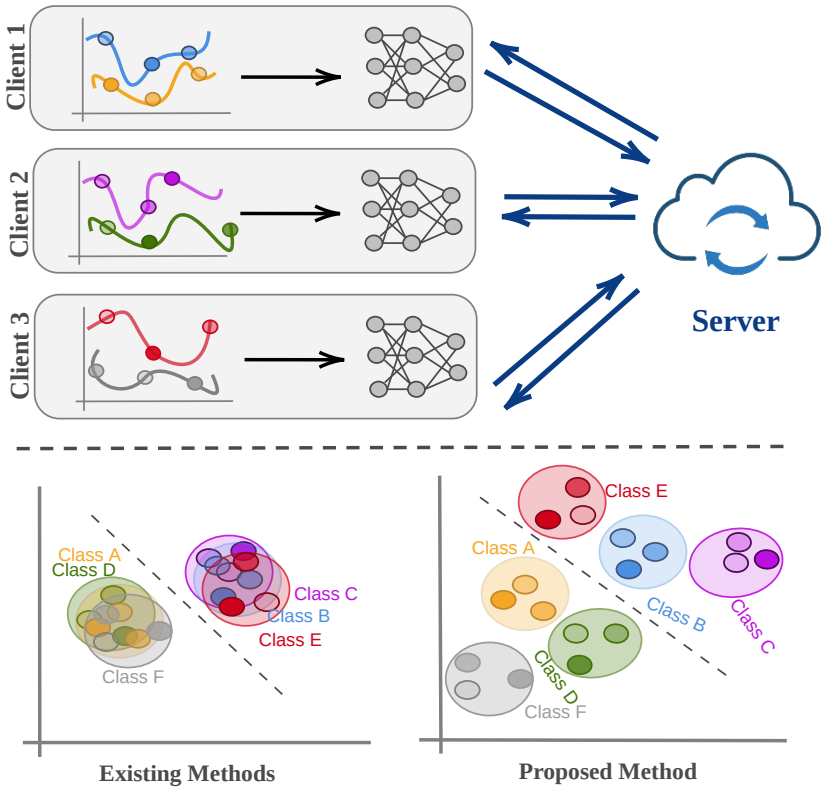


Figure 5.1: The representational collapse problem in FL. While clients may learn well-separated embeddings within their local classifiers, differences in data distributions across clients lead to conflicting feature spaces. After model aggregation, this discrepancy causes the global model's embeddings to collapse, leading to the loss of inter-class separability and discriminative structure.

The statistical heterogeneity of local data distributions remains one of the central challenges in FL. In practical deployments, client datasets are rarely identically distributed, often differing in class proportions, sample sizes, and feature distributions. Such data heterogeneity introduces fundamental difficulties in training a global model that can generalise effectively across clients. Consequently, the performance of FL algorithms can degrade substantially, particularly in settings characterised by limited local data or severe class imbalance, both of which are common in real-world applications.

Representational collapse occurs when the feature space loses its discriminative power, causing embeddings from different classes or clients to become indistinguishable. Figure 5.1 il-

lustrates the impact of data heterogeneity in federated learning. Under heterogeneous conditions, clients are trained on divergent local data distributions, leading to the formation of inconsistent and incompatible feature representations. When these misaligned representations are aggregated on the central server, the global model collapses, failing to preserve class separation. This results in overlapping or trivial embeddings, which ultimately impair generalisation and weaken the model's representational robustness.

There are several studies to tackle data heterogeneity problems in the local training phase. FedProx Sharma et al., 2022 adds the Euclidean distance between the global and server model parameters to the objective function to ensure that local updates do not deviate too much from the global model, making the updates smoother and more stable. SCAFFOLD Karimireddy et al., 2020 presents a stochastic algorithm to solve the client drift problem by gradient dissimilarity using control variables. These approaches demonstrate that leveraging global model knowledge enhances the robustness of local model representations.

Beyond regularisation-based approaches, contrastive learning has emerged as an effective technique for leveraging global knowledge in FL. MOON Q. Li et al., 2021 tackles the data heterogeneity problem by maximising the agreement between local and global model representations, demonstrating how global knowledge enhances local model robustness. FedRCL Seo et al., 2024 relies on a similar perspective to the MOON methodology by ensuring that data points from the same class are sufficiently well-separated in the feature space. This separation preserves diversity, enabling the model to learn more effectively and generalise better.

While aligning local and global representations is essential for effective federated learning, overly aggressive alignment can inadvertently cause representational collapse, where intra-class diversity is diminished within client models. This phenomenon hinders the model's ability to distinguish between minor variations within the same class, finally degrading both local and global performance. Overcoming such collapse requires a careful balance between intra-class variance and inter-class variance, ensuring that learned representations remain both discriminative and diverse across clients. Furthermore, we pose the following research question: *"What is the effect of inter-sample learning distances within each local model on the global model's generalisation capability?"* We hypothesise that if each local model learns to reduce the distance between similar (positive) samples while increasing the distance between dissimilar (negative) samples, the resulting representations will remain well-separated and discriminative even after the global aggregation. Consequently, this process is expected to produce more generalised and robust feature representations across clients, mitigating the adverse effects of data heterogeneity.

We propose FedQuad, a metric learning-based federated learning approach that introduces a novel loss function to mitigate representational collapse under data heterogeneity. Unlike traditional contrastive or triplet-based methods Khosla et al., 2020, Hoffer and Ailon, 2015, FedQuad explicitly models the relative distances between samples by constructing stochastic quadruplets: an anchor, a positive sample (same class), a negative sample (different class),

and a harder negative sample (also from a different class). The loss function simultaneously minimises the distance between the anchor and the positive while maximising the distance between the anchor and both negatives. This formulation encourages the model to learn a feature space in which intra-class samples are tightly clustered, while inter-class samples remain well-separated. By preserving both intra-class diversity and inter-class discriminability, FedQuad provides robust representations that are resilient to aggregation-induced degradation, directly addressing the core challenges of representational collapse in federated learning. To summarise, our major contributions are as follows:

- We propose a novel metric learning-based federated learning framework designed to address representation collapse under data heterogeneity.
- We analyse the impact of intra-class variance and inter-class variance, and their roles in preserving discriminative representations across clients.
- We introduce a novel quadruplet-based loss function that effectively mitigates representational collapse in both local and global models.
- We design an offline stochastic quadruplet sampling strategy per client, tailored to imbalanced and non-i.i.d. data distributions, to ensure robust and diverse training.

Traditional quadruplet loss W. Chen et al., 2017 primarily focuses on minimising the distance between the anchor and the positive sample, while providing only a weak push between the anchor and negative samples. Typically, it enforces a margin between a single negative pair, offering limited guidance on how to handle multiple or harder negatives. As a result, its effectiveness diminishes in highly heterogeneous settings such as FL, where negative samples can vary widely in difficulty. Without explicitly modelling or emphasising strong separation from challenging negatives, the learned representations may lack sufficient discriminative structure, reducing their utility in preserving inter-class boundaries.

Hard negative mining is an important step in contrastive and triplet loss-based algorithms Xuan et al., 2020. In our method, we present a class-separation-aware method to generate quadruplet batches, ensuring that each batch includes at least one correct positive pair and that each of the negative samples is from a different class than the anchor. This construction paradigm enables more relevant and informative optimisation, especially in federated contexts where client data distributions might vary substantially. In addition, our approach outperforms typical contrastive losses in underrepresented (rare) classes, where there are generally insufficient informative negatives.

5.2 Background

5.2.1 Federated Learning

FL frameworks such as FedAvg typically involve three main steps: *broadcasting*, *local training*, and *model aggregation*. After each communication round, the central server broadcasts the current global model to all participating clients. Each client then performs local training on its private data, ensuring keeping raw data on-device. Once local updates are completed (e.g., after a fixed number of epochs), clients send their updated model parameters back to the server. The server performs model aggregation, typically by averaging the received parameters, to form a new global model, which is then broadcast in the next round. Following the typical FL approach, the data D is distributed across K clients. Let D_k be the local dataset at client k, with $n_k = |D_k|$ denoting the number of data points at client k. The global objective function in FL is then a weighted average of the local objectives

$$\min_w F(w) = \sum_{k=1}^{K} \frac{n_k}{n} F_k(w) \tag{5.1}$$

where $n = \sum_{k=1}^{K} n_k$ is the total number of data points across all clients and $F_k(w)$ is the local objective function at a client k

$$F_k(w) = \frac{1}{n_k} \sum_{i \in \mathcal{D}_k} f_i(w) \tag{5.2}$$

To the best of our knowledge, this is the first work to explore the integration and analysis of metric learning losses for representation learning under federated learning settings, particularly during local client updates in the presence of data heterogeneity. Previous research has mostly focused on two complementary ways to address heterogeneity in federated learning: improving local training and optimising global aggregation. This work belongs to the local training improvement.

5.2.2 Metric Learning

Metric learning aims to learn an embedding function that maps input samples into a feature space where the geometric distance between embeddings reflects their semantic similarity. The primary objective is to construct a latent space in which samples from the same class are **closely clustered** (low intra-class variance), while samples from different classes are **well separated** (high inter-class variance).

Formally, given an input sample x and its label y, metric learning seeks to learn an encoder

$f(\cdot)$ that projects x into an embedding vector $z = f(x) \in \mathbb{R}^d$, d dimensional, such that:

$$\|f(x_i) - f(x_j)\|_2^2 \begin{cases} \text{is small,} & \text{if } y_i = y_j, \\ \text{is large,} & \text{if } y_i \neq y_j. \end{cases} \tag{5.3}$$

The goal is to learn a transformation that encodes semantic relations via the Euclidean (or cosine) distance. This ensures that similar samples remain close together, while dissimilar ones are pushed apart.

Two critical principles govern metric learning:

- $\downarrow$ **Intra-class variance:** embeddings belonging to the same class should form tight clusters in the latent space.
- $\uparrow$ **Inter-class variance:** embeddings of different classes should be distant enough to ensure discriminability.

Balancing these two criteria is key to achieving robust, generalizable feature representations.

Contrastive Loss

The *contrastive loss* Khosla et al., 2020 was one of the earliest metric learning objectives, used in Siamese networks Koch et al., 2015 for similarity learning. It operates on pairs of samples (x_i, x_j) and encourages small distances for positive pairs (same class) and large distances for negative pairs (different classes). It is defined as:

$$\mathcal{L}_{\text{contrastive}} = (1 - y_{ij}) \frac{1}{2} D_{ij}^2 + y_{ij} \frac{1}{2} \big[\max(0, m - D_{ij}) \big]^2, \tag{5.4}$$

where $D_{ij} = \|f(x_i) - f(x_j)\|_2$ is the Euclidean distance, $y_{ij} = 0$ for a positive pair and 1 for a negative pair, and $m > 0$ is a margin enforcing inter-class distance. The margin ensures that negative pairs are pushed apart until their distance exceeds m, while positive pairs are pulled closer together.

Triplet Loss

The *triplet loss* Hoffer and Ailon, 2015 extends the contrastive learning paradigm from pairs of samples to triplets, enabling more fine-grained control over the relative distances between instances in the embedding space. Each triplet consists of an *anchor* sample x_a, a *positive* sample x_p from the same class as the anchor, and a *negative* sample x_n drawn from a different class. The objective encourages the representation of the anchor to be closer to that of the positive sample than to the negative one by at least a margin $m > 0$. Formally, the loss is defined as:

$$\mathcal{L}_{\text{triplet}} = \max \big(0, D(x_a, x_p) - D(x_a, x_n) + m\big), \tag{5.5}$$

where $D(\cdot, \cdot)$ denotes a distance or dissimilarity measure—typically the Euclidean distance or one minus the cosine similarity between feature embeddings.

By enforcing this relative distance constraint, the triplet loss simultaneously promotes *low intra-class variance* (reducing distances between samples of the same class) and *high inter-class variance* (increasing distances between samples from different classes). In contrast to binary contrastive losses, which operate on independent pairs, the triplet formulation directly encodes a relational ordering between three instances, yielding richer geometric supervision for embedding learning.

However, the effectiveness of triplet-based training depends heavily on the strategy used to construct triplets. Random sampling often produces a high proportion of *easy triplets* that already satisfy the margin constraint, contributing little to the optimisation and leading to slow convergence. To address this, most implementations adopt *hard* or *semi-hard* mining strategies Xuan et al., 2020 Schroff et al., 2015, which selectively choose negative samples that are challenging yet not overly difficult. Hard negatives, those closer to the anchor than the positive, provide stronger gradients and improve discrimination, whereas semi-hard negatives strike a balance by lying within the margin but maintaining correct class separation.

Recent extensions of the triplet loss have incorporated additional regularisation terms or adaptive margins to enhance stability and generalisation. For instance, adaptive triplet losses Y. Yuan et al., 2020, T. Wang and Isola, 2021 dynamically adjust the margin based on class similarity, while proxy-based variants Movshovitz-Attias et al., 2017 replace explicit negative sampling with class-level representatives to improve scalability. These developments have made triplet-based objectives a cornerstone of modern metric learning and contrastive representation frameworks, serving as the foundation for higher-order formulations such as the *quadruplet loss* and the *n-pair loss*.

Quadruplet Loss

To further improve discriminability, the *quadruplet loss* is presented in the paper, W. Chen et al., 2017, extends the triplet loss by introducing an additional negative sample, resulting in a four-sample tuple (x_i, x_j, x_k, x_l). This formulation simultaneously enforces:

- smaller distance between anchor–positive pairs (x_i, x_j),
- larger distance between anchor–negative (x_i, x_k) and negative–negative pairs (x_l, x_k).

The loss (W. Chen et al., 2017) can be expressed as:

$$\begin{aligned} L_{\text{quad}} &= \sum_{i,j,k}^{N} \left[g\left(x_i, x_j\right)^2 - g\left(x_i, x_k\right)^2 + \alpha_1 \right]_+ \\ &\quad + \sum_{i,j,k,l}^{N} \left[g\left(x_i, x_j\right)^2 - g\left(x_l, x_k\right)^2 + \alpha_2 \right]_+ \\ &\quad s_i = s_j, s_l \neq s_k, s_i \neq s_l, s_i \neq s_k \end{aligned} \tag{5.6}$$

where α_1 and α_2 are margin terms controlling inter-class variance, $[z]_+ = max(z, 0)$ and s_i anchor image. The additional negative n_2 introduces a second repulsion constraint, encouraging global inter-class separation and better class boundary formation.

Compared to the triplet loss, the quadruplet loss not only enforces that an anchor is closer to its positive than to its negative, but also that all negatives are consistently far from all positives, thereby enhancing embedding uniformity across classes.

The trade-off between intra-class compactness and inter-class separation is central to metric learning. Overemphasising compactness may lead to overfitting and reduced generalisation, while excessive separation can cause unstable training or fragmented clusters. A well-designed metric learning objective must therefore balance both, achieving:

Low intra-class variance and **high** inter-class margin.

Recent developments, including centre loss Y. Wen et al., 2016, large-margin softmax W. Liu et al., 2016, and angular-based objectives H. Wang et al., 2018, further refine this balance by explicitly modelling class centres or angular margins in hyperspherical embedding spaces.

In summary, metric learning provides the theoretical foundation for contrastive representation learning. Classical losses such as contrastive, triplet, and quadruplet losses explicitly structure the embedding space by optimising sample distances according to semantic similarity. These principles have inspired modern self-supervised and federated contrastive learning frameworks that build upon pairwise or tuple-wise similarity constraints to learn robust, discriminative, and semantically aligned representations across distributed clients.

Federated Metric Learning

Existing federated metric learning methods, such as FedMetric H. Park et al., 2021, learn embeddings using a metric loss that treats positive and negative pairs differently. However, they often fail to explicitly model the relative similarity between a given positive sample and multiple negative samples, limiting their ability to separate fine-grained structures in the representation space. Additionally, this method relies on proxy-based hypersphere clustering, which oversimplifies the underlying data distribution. Other methods, such as Tian et al., 2022, Gu et al., 2023, and Shao et al., 2023, focus on designing or enhancing the overall metric learning models in a federated setting, rather than explicitly applying metric learning losses (e.g., contrastive, triplet,

or quadruplet loss) during local training on each client.

While metric learning has proven effective in supervised representation learning by minimising intra-class variance and maximising inter-class variance, including Quadruplet loss W. Chen et al., 2017, Triplet loss Hoffer and Ailon, 2015, and supervised contrastive loss Khosla et al., 2020, its integration into FL remains underexplored. Our work is among the first to systematically investigate metric learning losses in the federated setting, focusing on their role in preventing representational collapse under severe data heterogeneity.

Data heterogeneity in Federated Learning

A central challenge in federated learning (FL) is the inherent non-i.i.d. nature of data distributed across clients. In practice, clients often exhibit significant class imbalance or domain-specific biases in their local datasets, which can substantially hinder the convergence speed and generalisation performance of the global model. To solve the challenge, several techniques have been presented Gao et al., 2022, Acar et al., 2021, Fang et al., 2025. Contrastive learning-based approaches have gained considerable attention in federated learning (FL) for their effectiveness in aligning global and local representations, thereby mitigating the impact of client drift and data heterogeneity. For instance, FedProcMu et al., 2023, FedCRL Huang et al., 2024 and FedPCL Y. Tan et al., 2022 reformulate contrastive objectives to reduce the divergence between local and global models, thereby improving alignment. Similarly, relaxed contrastive approaches such as MOON Q. Li et al., 2021, FedRCL Seo et al., 2024, and FedTrip X. Li et al., 2023 employ model-level alignment and distribution-aware aggregation to mitigate the negative impact of data heterogeneity. MOON introduces a model-contrastive loss, which aims to align the current local model with the global model while pushing the current model away from the local model of the previous round. In addition, unsupervised contrastive learning techniques like FedSimCLR Louizos et al., 2024 and FedMoCo Dong and Voiculescu, 2021 have been explored for federated settings, leveraging mutual information maximisation without requiring labelled data. However, these methods fall outside the scope of our work, as our focus lies in supervised image classification tasks where labelled data is available on each client. In contrast to prior methods, our framework neither requires an additional global model for contrastive learning nor depends on global models or prototypes to address deviations in local training.

5.3 Methodogy

Our proposed methodology is built upon the quadruplet loss framework for local training. Unlike the standard quadruplet loss, which contains a negative pair distance term (n_1, n_2) to limit the distance between two negative samples, we advance the focus to modelling the anchor's relation via multiple negative samples chosen from different classes. The negative pair component in

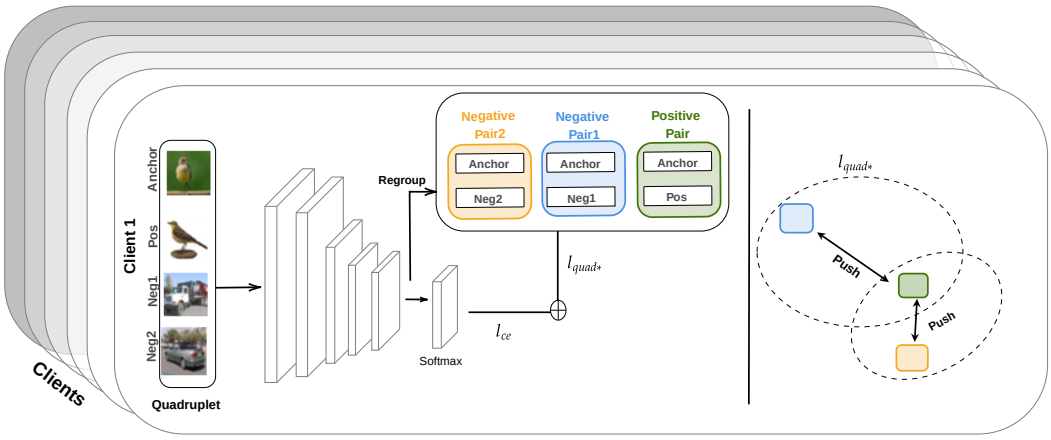


Figure 5.2: Overview of the FedQuad local training framework. Each client minimises loss composed of the cross-entropy loss ℓ_{ce}, computed after the softmax layer, and the proposed quadruplet loss ℓ_{quad*}, applied to the non-normalised embeddings from the encoder. The images are from the CIFAR10 dataset, which explains the low resolution.

traditional quadruplet loss frequently provides a weak push between negative pairs and may not contribute effectively to discriminative representation learning. Rather than explicitly modelling the structure among the negatives, we encourage the positive sample to be well-separated from the entire set of negatives collectively.

As the negative samples originate from different classes, encouraging broader separation from the anchor, rather than focusing solely on hard negatives, can enhance the discriminative capacity of the representations. Our loss function is designed to capture two key objectives: the first component enforces low intra-class variance, ensuring that features from the same class remain close in the embedding space; the second component optimises the distance between classes by simultaneously pushing the anchor away from multiple negatives. The presented loss function facilitates effective feature separation both at the local client models and within the globally aggregated model.

5.3.1 Stochastic Quadruplet Sampling

We designed a stochastic pair sampling strategy to generate quadruplet samples for representation learning. Each data item consists of an anchor image, a positive image sampled from the same class as the anchor, and two negative images drawn from other classes, specifically, classes that differ both from the anchor's class and from each other. In order to enable efficient sampling, the class first constructs a mapping between class labels and sample indices inside the provided subset. During generation samples for each batch, a positive sample is chosen at random from the anchor's class, ensuring that it is not the same as the anchor itself. Negative samples are selected by first identifying all classes present in the subset, eliminating the anchor's class, then sampling two different classes and extracting one instance from each. This class-aware sampling approach

ensures semantically appropriate quadruplets, allowing for the learning of discriminative and robust feature representations, especially under class imbalance or non-i.i.d. limitations in federated learning. $\mathcal{D} = \left\{\left(x_i, y^i\right)\right\}_{i=1}^{N}$ be the dataset with inputs $x_i \in \mathbb{R}^d$ and labels $y^i \in \{0, 1, \ldots, K-1\}$.

- $x^a \sim \mathcal{D}_k$
- $x^p \sim \mathcal{D}_k \setminus \{x^a\}$
- $x^{n1} \sim \mathcal{D}_{k'}$ where $k' \in \{0, \ldots, K-1\} \setminus \{k\}$
- $x^{n2} \sim \mathcal{D}_{k''}$ where $k'' \in \{0, \ldots, K-1\} \setminus \{k, k'\}$

This sampling approach facilitates training with the quadruplet loss, which enforces the following constraints in the embedding space:

- Positive pairs (x^a, x^p) are pulled close together (minimazing intra-class variance),
- Anchor x^a is pushed away from two distinct negatives (x^{n1}, x^{n2}) and enhancing robustness,
- Selecting two distinct negatives encourages higher inter-class variance.

Such structured sampling is particularly beneficial in metric learning and federated learning settings, where it improves generalisation and discriminative ability under data heterogeneity.

As illustrated in Figure 5.2, our local loss comprises two parts. The first part is a typical supervised loss (e.g., cross-entropy loss) denoted as ℓ_{ce}. The second part is our reformulated quadruplet loss denoted as ℓ_{quad*}. We define our loss as

$$\ell = \ell_{\text{ce}} \left(w_i^t; (x^a, y^a)\right) + \beta \ell_{\text{quad*}} \left(wi^t; (x^a, x^p, x^{n1}, x^{n2})\right) \tag{5.7}$$

where β is a hyperparameter to maintain the impact of quadruplet loss. Our local objective is to minimise

$$\begin{aligned} \min_{w_i^t} \; \mathbb{E}_{(x^a, y^a) \sim D^i} \Big[& \ell_{\text{ce}}(w_i^t; (x^a, y^a)) \\ & + \beta \, \ell_{\text{quad*}}(w_i^t; (x^a, x^p, x^{n1}, x^{n2})) \Big] \end{aligned} \tag{5.8}$$

where m_1 and m_2 are the values of margins in the two terms and y^a refers to the class label of image x^a, z_a is representation. The margins define the minimum desired difference between the distance from the anchor to the negatives. Our proposed loss is defined as;

$$\begin{aligned} \ell_{\text{quad*}} = & \left[\|f(x^a) - f(x^p)\|_2 - \left\| f(x^a) - f\left(x^{n1}\right) \right\|_2 + m_1 \right]_+ \\ & + \left[\|f(x^a) - f(x^p)\|_2 - \left\| f(x^a) - f\left(x^{n2}\right) \right\|_2 + m_2 \right]_+ \end{aligned} \tag{5.9}$$

where $[x]_+ = max(0, x)$, $f(\cdot)$ is a function to extract embeddings. We choose two distinct margin values in our loss to avoid enforcing the same separation radius for both negative samples. This

allows more flexibility in the embedding space, preventing all negative instances from being pushed away uniformly, and instead adapting the separation based on their semantic or class-wise dissimilarity.

The overall federated learning algorithm is shown in Algorithm 4 that represents the FedQuad design. FedQuad remains applicable when only a small subset of clients participates in each federated learning round. We follow the standard FedAvg approach for model aggregation; each client maintains a local model, which is synchronised with the global model regularly and updated with the local models from the clients participating in that round. Table 5.3 and 5.2,

Input: Dataset D, Number of communication rounds T, number of clients N, local epochs E, B batch size, hyperparameters β, m_1, m_2
Output: Global model w^T
Server
Initialize global model w^0
for $t = 0, 1, \ldots, T-1$ **do**
  for $i = 1$ **to** N **do**
    Get global model w^t to update client C_i
    $w_i^t \leftarrow \textsc{trainClient}(i, w^t)$
  end
  $w^{t+1} \leftarrow \sum_{k=1}^{N} \frac{|D_i|}{|D|} w_k^t$
end
return w^T

Function $\textsc{trainClient}(i, w^t)$:

$w_i^t \leftarrow w^t$
for $e = 1$ **to** E **do**
  for *each batch* $b = \{x_i^a, x_i^p, x_i^{n1}, x_i^{n2}, y_i^a\}_{i<B}$ *from* $\mathcal{D}_i$ **do**
    $z_a \leftarrow f_{w_i^t}(x^a)$
    $z_{\mathrm{p}} \leftarrow f_{w_i^t}(x^p)$
    $z_{\mathrm{n1}} \leftarrow f_{w_i^t}(x^{n1})$
    $z_{\mathrm{n2}} \leftarrow f_{w_i^t}(x^{n2})$
    $\ell_{\mathrm{quad}*} \leftarrow \left[d\left(z_a, z_p\right)^2 - d\left(z_a, z_{n1}\right)^2 + m_1\right]_+ + \left[d\left(z_a, z_p\right)^2 - d\left(z_a, z_{n2}\right)^2 + m_2\right]_+$
    $\ell_{\mathrm{ce}} \leftarrow \mathrm{CrossEntropyLoss}(F_{w_i^t}(x^a), y^a)$
    $\ell \leftarrow \ell_{\mathrm{ce}} + \beta \cdot \ell_{\mathrm{quad}*}$
    $w_i^t \leftarrow w_i^t - \eta \cdot \nabla \ell$
  end
end
return w_i^t

Algorithm 4: FedQuad Framework

our version of QuadrupletFL and FedQuad, consistently outperforms the standard supervised federated metric learning baseline. Although the modified loss function introduces twice as two negative pairs compared to the traditional triplet loss, it achieves superior performance by effectively leveraging a richer set of negative relationships. While triplet loss focuses on a single

positive-negative pair at a time, incorporating multiple negative samples better facilitates the maximisation of inter-class variance, leading to more discriminative representations.

5.4 Experiment

We compare our proposed method against supervised metric learning approaches, including triplet loss, supervised contrastive loss, and quadruplet loss. To ensure a fair comparison, we implement and evaluate these losses within a federated learning setting. Additionally, we include a baseline supervised version trained on the entire dataset. Our experiments are conducted on the CIFAR10 and CIFAR100 datasets Krizhevsky, 2009.

5.4.1 Experimental Setup

Our model consists of a convolutional neural network (CNN) backbone followed by a fully connected layer to produce embeddings of a specified dimension. The CNN backbone comprises three convolutional blocks. Each block includes a 2D convolutional layer with a kernel size of 3×3, stride 1, and padding 1, followed by batch normalisation and a ReLU activation. The first two blocks conclude with a 2×2 max pooling layer to reduce spatial dimensions. The final convolutional block uses an adaptive average pooling layer to produce a fixed-size output of 1×1 spatial dimensions. The output of the convolutional backbone is flattened and passed through a fully connected layer that maps the 256-dimensional feature vector to the desired embedding dimension, which is 128 in our experiments. The forward pass applies the CNN layers, flattens the output, and then applies the linear layer to obtain the final embeddings. Additionally, a softmax layer is included solely for cross-entropy loss measurement. We make the code for our experiments publicly available to ensure reproducibility.[2] We use the Adam optimiser with a learning rate of 0.001 for all approaches. The Adam weight decay is set to 10^{-5}, and the momentum is fixed at 0.9. The batch size is 128. For all federated learning approaches, the number of local epochs is set to 5 unless otherwise specified. The number of communication rounds is set to 20 for CIFAR10 and CIFAR100, as additional rounds yield little to no accuracy improvement.

Following prior works McMahan et al., 2017, Q. Li et al., 2021, we employ a Dirichlet distribution to generate non-i.i.d. data partitions among clients, with concentration parameters $\alpha = 0.5$ and $\alpha = 0.3$ (lower α indicates highly skewed data distribution). This partitioning strategy results in some clients having relatively few or even no samples for certain classes. We evaluate scenarios with 10, 50, and 200 clients with many data distributions in table 5.1a for CIFAR10, table 5.1b for CIFAR100. The best β for reformulated quadrupled loss is 0.5, and for margin values $m_1 = 1.0$, $m_2 = 0.5$, which is shown in an ablation study Table 5.4.

[2] https://anonymous.4open.science/r/FedQuad-55C8/README.md

5.5 Results

We compare the proposed method, FedQuad, against several metric learning-based federated learning baselines under varying clients and data distribution settings, using the CIFAR10 and CIFAR100 datasets. As shown in Table 5.1, FedQuad consistently outperforms all baseline methods across different levels of data heterogeneity. Notably, Table 5.2 and 5.3 demonstrate that our method maintains high performance even under highly non i.i.d. distributions with many clients.

FedQuad explicitly maximises inter-class variance within each client's representation space, which enhances feature discrimination. This is particularly beneficial when evaluating the global model on test data, which typically includes samples from all clients. Even in cases where certain clients have not encountered specific class samples, FedQuad enables the global model to learn robust representations.

Method	**i.i.d.**	α **= 0.5**	α **= 0.3**
Supervised (Non-FL)			
Supervised	**82.57**	–	–
SupCon	76.67	–	–
Triplet	64.52	–	–
Quadruplet	67.72	–	–
Supervised FL			
FedAvg	81.26	77.34	74.05
SupConFL	71.06	69.46	68.07
TripletFL	59.67	60.77	40.42
QuadrupletFL	62.79	64.18	47.27
MOON	76.86	61.43	49.69
FedQuad	**82.35**	**80.83**	**80.13**

(a) CIFAR10 results (10 clients).

Method	**i.i.d.**	α **= 0.5**	α **= 0.3**
Supervised (Non-FL)			
Supervised	**50.96**	–	–
SupCon	37.33	–	–
Triplet	39.43	–	–
Quadruplet	40.33	–	–
Supervised FL			
FedAvg	51.33	47.56	44.32
SupConFL	34.79	36.21	36.72
TripletFL	36.29	35.26	33.01
QuadrupletFL	39.49	35.96	33.92
MOON	26.32	25.73	26.10
FedQuad	**51.27**	**50.64**	**48.55**

(b) CIFAR100 results (10 clients).

Table 5.1: Test accuracy (%) comparison across supervised and federated learning variants on CIFAR10 and CIFAR100 under varying data heterogeneity (α).

Tables 5.2 and 5.3 demonstrate that a larger number of clients can exacerbate data sparsity and increase diversity, which in turn hinders the performance of many existing methods. This effect is particularly evident on CIFAR100, which contains 100 classes with only 500 samples per class. When data is distributed among 200 clients, the number of samples per class per client becomes very limited, making it challenging to learn robust representations.

Despite this challenge, our method effectively handles both data imbalance and sparsity under non-i.i.d. data distributions. In contrast, MOON demonstrates substantially degraded performance under these conditions, highlighting the limitations of global model regularisation–based contrastive learning approaches in scenarios with a large number of clients and highly diverse data distributions. These findings indicate that MOON struggles to generalise and fails to learn robust representations when provided with serious data heterogeneity.

Method	**i.i.d.**			**α = 0.5**			**α = 0.3**		
	10C	50C	200C	10C	50C	200C	10C	50C	200C
FedAvg	79.86±0.28	70.53±0.09	61.46±0.44	76.85±0.86	68.61±0.46	59.82±0.21	74.03±1.37	67.6±0.67	58.57±0.76
SupConFL	71.55±0.77	50.27±0.08	46.92±0.51	68.88±3.33	49.75±0.05	46.38±0.41	69.07±1.32	49.53±0.21	46.29±0.03
TripletFL	74.25±0.40	61.23±0.10	54.86±0.62	64.60±3.68	60.48±0.58	54.25±0.30	55.53±3.10	57.97±0.25	54.46±0.23
QuadrupletFL	76.77±0.28	72.32±0.17	61.66±0.65	70.30±0.14	59.45±0.43	**61.12**±2.10	58.79±2.42	**68.68**±0.62	58.76±0.01
MOON	79.03±0.11	71.02±0.33	60.51±0.57	69.23±2.63	69.36±0.22	59.94±0.09	64.19±2.35	68.04±0.46	59.08±0.32
FedQuad	**82.37**±0.25	**72.37**±0.17	**63.02**±0.01	**80.76**±0.22	**69.80**±0.19	59.86±0.21	**79.45**±0.67	68.03±0.79	**59.46**±0.21

Table 5.2: Test accuracy (%) of different methods on CIFAR10 under varying data heterogeneity and client numbers. The bold values indicate the best results for each setting.

Method	**i.i.d.**			**α = 0.5**			**α = 0.3**		
	10C	50C	200C	10C	50C	200C	10C	50C	200C
FedAvg	**51.52±0.15**	35.23±0.17	24.86±0.84	47.75±0.31	**36.28±0.13**	24.33±0.13	44.95±0.44	**35.23±0.17**	23.84±0.21
SupConFL	35.05±0.21	27.28±0.56	20.22±0.21	36.00±0.21	27.01±0.58	19.34±0.11	36.41±0.22	26.87±0.32	19.01±0.18
TripletFL	36.56±0.19	27.56±0.12	23.83±0.29	35.10±0.47	30.89±4.56	23.63±0.74	33.96±0.96	28.42±0.15	24.26±0.15
QuadrupletFL	37.85±0.45	29.67±0.41	25.83±0.47	34.69±0.89	29.64±0.09	25.53±0.38	32.21±1.21	29.93±0.24	25.57±0.63
MOON	37.09±0.61	23.72±0.02	14.69±0.32	32.80±0.35	21.46±0.16	13.63±0.36	31.43±1.72	20.09±0.46	13.06±0.02
FedQuad	51.49±0.16	**36.59**±0.41	**26.07**±0.17	**50.77**±0.24	31.55±5.32	**26.95**±0.31	**50.56**±0.04	34.96±0.41	**26.90**±0.16

Table 5.3: Test accuracy (%) of different methods on CIFAR100 under varying data heterogeneity and client numbers. The bold values indicate the best results for each setting.

Figures 5.3 and 5.4 illustrate the ratio of inter-class to intra-class distances, along with the corresponding average inter-class distance values for both datasets. In the CIFAR10 experiments, across various data distribution settings, SupConFL demonstrates relatively weak performance, exhibiting poor discriminative ability between similar and dissimilar feature representations. A higher ratio indicates stronger class separability, which is observed in TripletFL with around 1.5. However, TripletFL also exhibits the highest average intra-class distance, suggesting that although inter-class separation is substantial, samples within the same class remain dispersed, leading to suboptimal intra-class compactness.

In contrast, FedQuad achieves the lowest intra-class distance, indicating that features of similar samples are tightly clustered. This balance between minimising intra-class variance and maintaining sufficient inter-class separation highlights the superior representational consistency of FedQuad compared to other methods.

The CIFAR100 figure 5.4 reveals that FedAvg exhibits the highest intra-class distance, indicating its limited ability to handle highly diverse datasets, particularly when the number of classes increases tenfold compared to CIFAR10. This suggests that FedAvg struggles to achieve effective feature separation under such complexity. However, when examining the pairwise distance distributions, FedAvg demonstrates slightly improved discrimination across certain class pairs.

In contrast, FedQuad consistently achieves a balanced inter-class to intra-class distance ratio across all three data distribution settings. Its notably lower intra-class distance highlights its superior capacity to cluster similar samples more compactly while maintaining clear separation

between distinct classes.

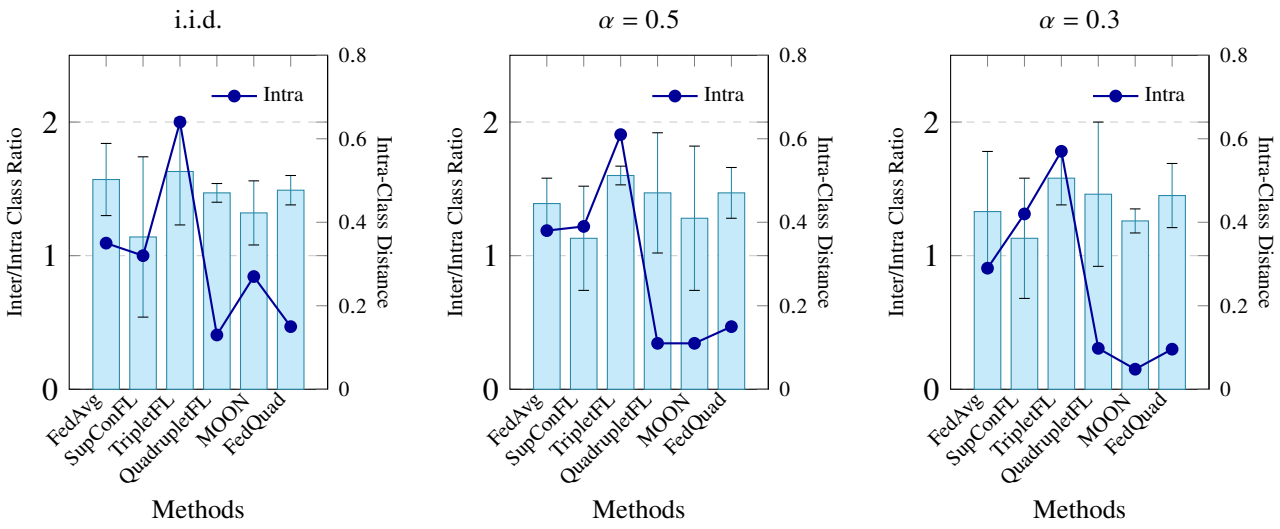


Figure 5.3: Inter/Intra-class ratio (↑ better) across federated learning methods on CIFAR10 (200 clients). Error bars indicate standard deviation over runs.

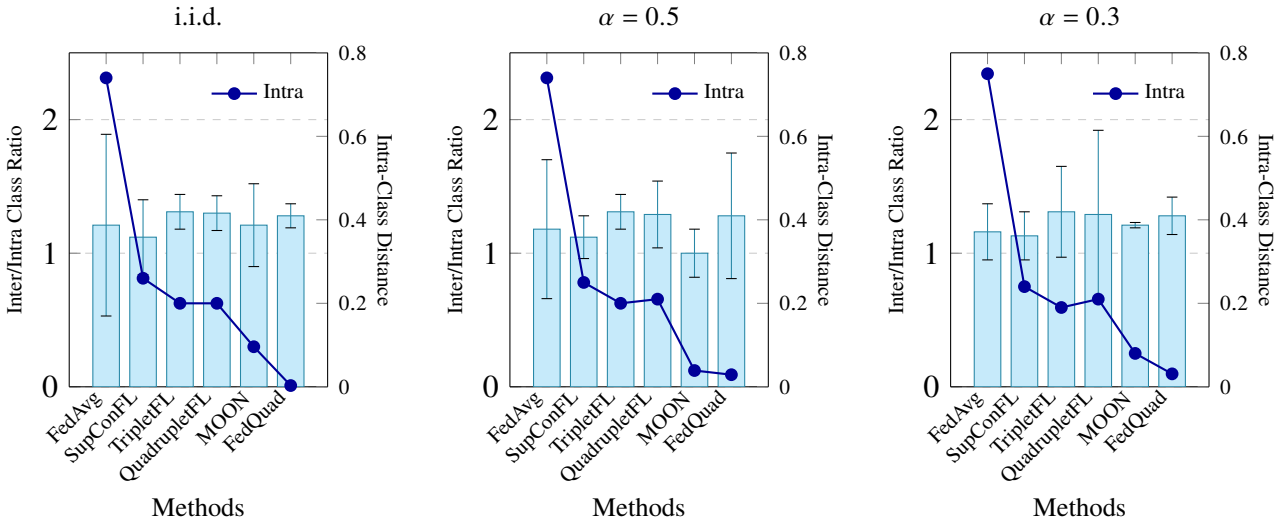
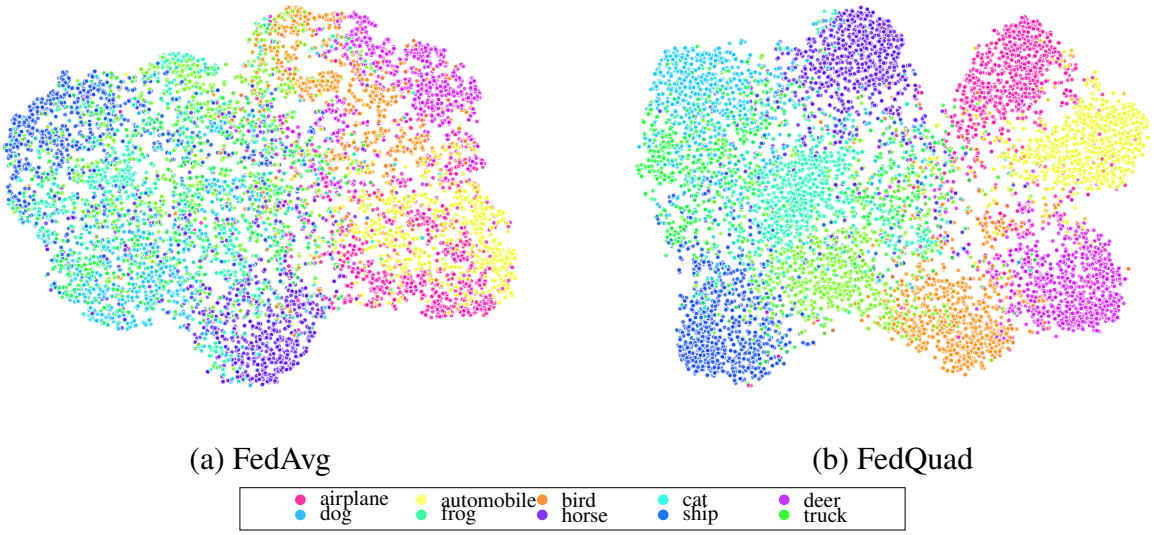


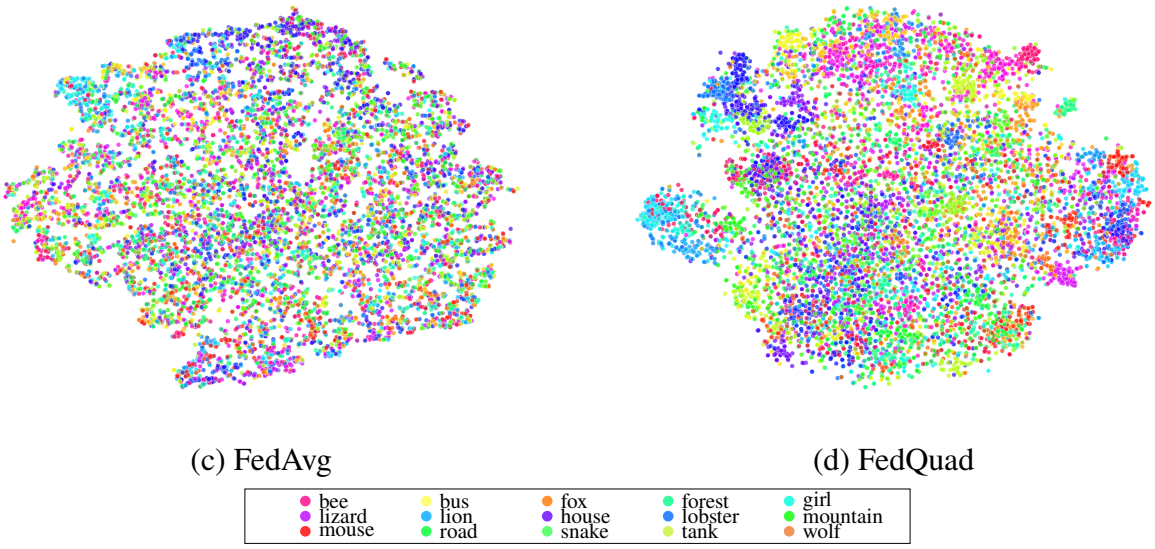
Figure 5.4: Inter/Intra-class ratio (↑ better) across federated learning methods on CIFAR100 (200 clients). Error bars indicate standard deviation over runs.

Figure 5.5 presents the global model embeddings for both datasets. Despite being at an early stage of training, t-SNE plots demonstrate that metric learning–based local training can capture discriminative and robust features, enabling effective separation of samples across classes.

The performance of the randomly selected client scenario is presented in Table 4.4, indicating significant data differences across training rounds. The random client selection approach, shown in Figure 5.6, is to select a subset of clients at each round for model aggregation, while the remaining clients do not update their local weights. Results show that increasing the total number of clients generally results in decreased accuracy, as the divergence between client data



(a) FedAvg (b) FedQuad

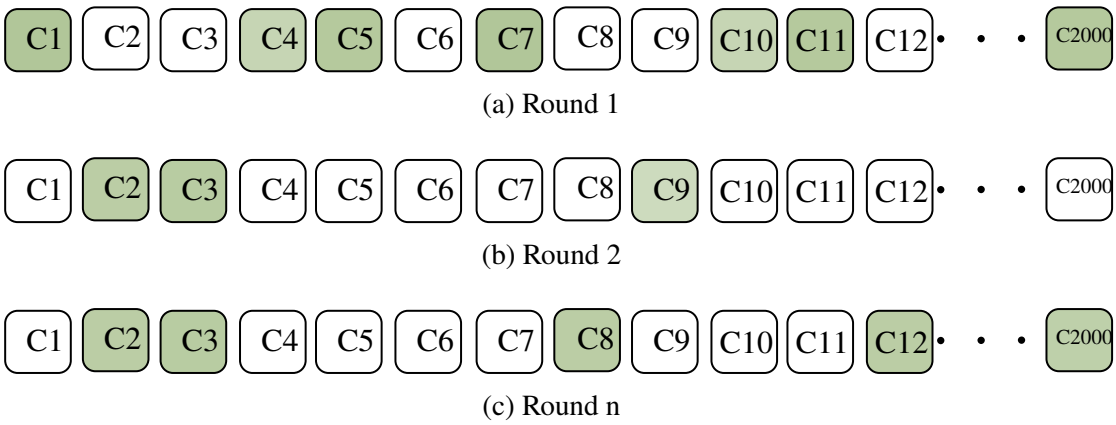


(c) FedAvg (d) FedQuad

Figure 5.5: t-SNE visualisation of learned representations at an early stage of training (Round 5) under a non-i.i.d. data distribution. The first row shows the global model's embeddings on CIFAR10, all ten classes. The second row presents embeddings for a subset of classes from CIFAR100.

Method	β	m_1	m_2	Accuracy (%)
FedQuad	0.5	1.0	1.0	71.51
FedQuad	0.5	1.0	0.5	**72.68**
FedQuad	1.0	1.0	0.5	70.4
FedQuad (without ℓ_{ce})	0.5	1.0	0.5	62.41
FedQuad (without ℓ_{ce})	0.5	2.0	0.5	60.99
FedQuad (without ℓ_{ce})	0.5	5.0	0.5	57.26
FedQuad (without ℓ_{ce})	0.5	1.0	1.0	56.42

Table 5.4: Ablation study on the effect of loss hyperparameters (β, m_1, m_2) in the proposed quadruplet loss, evaluated on CIFAR10 under an i.i.d. setting with 10 clients over 5 communication rounds.



(a) Round 1

(b) Round 2

(c) Round n

Figure 5.6: In each round, the central server samples a random subset of clients to participate. Only these selected clients perform local training and return updates, while non-selected clients remain idle for that round. This stochastic participation reduces communication cost and improves scalability.

distributions becomes more pronounced and the global model must aggregate highly inconsistent updates.

However, when the participation fraction increases, more clients contribute to each round, resulting in more stable aggregation and higher accuracy. This effect is most noticeable in the quadruplet-based federated learning method: Quadruplet FL achieves 57.62% accuracy, whereas FedQuad obtains 58.21%. These findings indicate metric-learning objectives, which explicitly optimise positive and negative pairwise distances, are more robust to low client participation rates

5.6 Discussion

Our experimental results show that FedQuad consistently enhances model generalisation in federated learning scenarios with a variety of data heterogeneity. FedQuad addresses one of the

Participation %	**Method**	CIFAR10		CIFAR100	
		200C	2000C	200C	2000C
1%	FedAvg	31.24±3.64	34.24±3.18	9.08±1.08	8.59±0.73
	SupCon-FL	44.29±0.27	43.67±0.44	16.84±0.2	17.21±0.28
	Triplet-FL	45.35±0.6	45.5±0.53	17.53±0.3	16.03±0.34
	Quadruplet-FL	52.96±0.23	50.61±0.29	16.01±0.54	15.92±0.53
	MOON	28.99±5.24	15.59±0.58	5.16±0.59	1.45±0.16
	FedQuad	**53.03 ±0.81**	**50.39±0.17**	**23.3±0.36**	**18.95±0.37**
5%	FedAvg	47.76±2.1	44.43±1.01	17.69±0.39	10.89±0.69
	SupCon-FL	45.44±0.34	43.60±0.55	17.95±0.25	17.32±0.2
	Triplet-FL	51.18±0.21	46.72±0.27	20.37±0.27	16.62±0.29
	Quadruplet-FL	57.62±0.29	51.75±0.23	18.81±0.24	16.97±0.04
	MOON	33.94±3.24	15.09±1.92	8.33±0.26	2.13±0.13
	FedQuad	**58.21±0.36**	**51.86±0.39**	**26.47±0.27**	**19.23±0.6**

Table 5.5: Method comparison on CIFAR10 and CIFAR100 with varying participation fraction and client scale.

most significant challenges in FL, representational collapse, which is common under non-i.i.d. conditions, by explicitly optimising the distances between positive and negative pairs. The strategy successfully increases inter-class variance, resulting in more stable and discriminative global representations without requiring raw data exchange among clients. Contrastive learning has shown potential in centralised settings, but its performance in federated learning is insignificant, especially under class-imbalanced distributions (e.g., Dirichlet $\alpha = 0.3$). We attribute this to the fundamental constraints of contrastive loss, which frequently results in mismatched or misaligned latent representations among clients. In contrast, our stochastic alignment approach provides flexibility to representation learning, allowing for more effective adaptation across varied client data. This leads to consistently higher performance, particularly under extreme non-i.i.d. conditions.

Traditional contrastive or triplet-based losses tend to collapse in cases with high class imbalance, leading to noisy or overlapping feature embeddings. FedQuad successfully maintains the structure of local representations while effectively aligning them with the global feature space. This is achieved without requiring direct computation of distances between negative pairs in each batch or any data exchange.

A key insight from our study is that achieving a balance between local discriminability and global consistency in federated learning requires deliberate loss function design. The proposed stochastic quadruplet formulation extends beyond traditional contrastive and triplet losses by introducing stronger negative constraints and finer-grained control over pairwise relations. This leads to embeddings that are significantly more robust to client-level distributional shifts. Such robustness is particularly valuable in low-resource settings or under severe class imbalance, where conventional federated learning methods often suffer from performance degradation due

to poorly aligned feature spaces.

While FedQuad demonstrates robust performance under various non-i.i.d. conditions, its applicability is limited in scenarios where clients possess only a small number of class labels such as two or three classes, or in binary classification settings. In such cases, negative sample mining becomes infeasible, thereby restricting the effectiveness of the quadruplet-based metric learning mechanism.

Our method focuses on positive versus two negatives simultaneously, allowing the model to push embeddings away from multiple negative directions while preserving similarity with positive pairs, a key strength that improves representation robustness. Furthermore, while our experiments on CIFAR10 and CIFAR100 demonstrate FedQuad's utility, extending evaluation to limited and domain-specific datasets such as medical imaging or user behaviour data would provide stronger evidence of the method's practical applicability and generalisability in real-world federated learning scenarios.

5.7 Conclusion

In this work, we introduced FedQuad, a federated learning framework based on metric learning, designed to address representational collapse caused by data heterogeneity among clients directly. Leveraging a stochastic quadruplet loss, FedQuad promotes lower intra-class variance and higher inter-class variance inside local feature spaces, thereby enhancing global representations without requiring access to raw client data. Furthermore, we provide an in-depth analysis of metric learning in federated settings, particularly under conditions where data is imbalanced and limited.

Comprehensive experiments on CIFAR10 and CIFAR100 across a range of non-i.i.d. scenarios demonstrate that FedQuad consistently outperforms existing baselines, especially in the presence of class imbalance and a large number of clients. These results underscore the promise of metric learning for local representation alignment and highlight the importance of structured embedding objectives in mitigating the effects of client divergence. This work establishes the way for future research, such as extending FedQuad to semi-supervised and unsupervised federated learning, enhancing hard negative mining strategies, and exploring local training objectives for robust representation alignment under heterogeneous distributions.

Chapter 6

Unsupervised Federated Partial Model Training

In this chapter, we investigate data silos and large client populations in federated and distributed learning settings, with a focus on highly heterogeneous data distributions. We focus on stochastic client participation, which better reflects real-world scenarios in which clients may temporarily join or drop out due to connectivity constraints, limited computational resources, or insufficient local data. We analyse how this periodic participation affects both the optimisation and global model convergence, and we evaluate how well learned representations generalise to unseen domains.

In addition, we evaluate a partial model federated learning setup in which a pretrained backbone remains frozen while a federated projection head is trained across clients. This design enables assessing whether robust, pretrained feature extractors can adjust for severe data heterogeneity, reduce the sensitivity of the global model to client dropout, and stabilise training. By decoupling feature extraction from the federated training process, we aim to determine to what extent fixed, high-quality representations improve robustness in the presence of diverse, imbalanced, and randomly available client.

6.1 Introduction

Federated learning (FL) has become a widely adopted paradigm for scenarios involving distributed data requirements, where collaborative model training is essential. A standard FL framework consists of three main stages: local training on each client, aggregation of model updates on a central server, and broadcasting of the updated global model back to clients. However, when the number of participating clients is large, data heterogeneity emerges as one of the most significant challenges. Each client holds data that includes its own patterns, environments, and characteristics, resulting in non-i.i.d. (non-independent and identically distributed) local datasets. As the number of clients grows into the hundreds or thousands, the global model

integrates diverse and imbalanced data sources, which commonly leads to client drift, unstable local updates, and slow global convergence.

These challenges become more pronounced in unsupervised settings, where clients operate without access to labelled data. Without supervision, clients rely on representation learning objectives, making them more sensitive to local distribution shifts and biases. Heterogeneous, unlabelled data easily causes feature spaces to diverge across clients, resulting in locally specialised embeddings that fail to generalise at the global level. Thus, achieving stable and robust representation learning under large-scale, distributed, non-i.i.d., and unlabelled conditions remains an open and unresolved problem in FL.

The literature proposes several approaches to address these challenges. Optimisation-oriented approaches seek to stabilise training and reduce client drift through strategies such as proximal regularisation (e.g. FedProx Sharma et al., 2022), stochastic variance reduction (e.g. SCAFFOLD Karimireddy et al., 2020), and server-side or adaptive momentum updates (e.g. FedOpt/FedAdam Reddi, Charles, et al., 2021, FedDyn Acar et al., 2021). Personalisation and model adaptation methods recognise that a single global model may not fit many clients; approaches such as Ditto T. Li et al., 2021, FedPer Arivazhagan et al., 2019, and pFedMe T Dinh et al., 2020 learn shared feature extractors with individual client model layers, while cluster-based approaches (e.g. IFCA Ghosh et al., 2020) group clients by distributional similarity to reduce cross-client interference.

Representation-based strategies, including contrastive, metric-based, and self-supervised objectives, seek to enforce encoder-level consistency across clients. Complementary approaches involving augmentation, distillation, or shared synthetic statistics (e.g. FedKD Lin et al., 2020, DS-FL Z. Zhu et al., 2021, FedGen Kang et al., 2025) aim to mitigate distribution gaps by approximating more i.i.d.-like conditions without violating privacy constraints.

While these directions address important aspects of data heterogeneity, standard FL assumes full model training on each client, which is computationally expensive and usually unrealistic for resource-limited or periodically connected devices in real-world deployments. Another line of research focuses on Partial Model Training (PMT) approaches (R. He et al., 2025), in which only a subset of the model parameters is optimised during federated training. These methods are inspired by a range of earlier ideas, including model pruning (FedMPR), selective parameter updating (e.g., FedSelect Tamirisa et al., 2024), layer-wise freezing and adaptation strategies such as FedBABU Oh et al., 2021, and dropout-based federated optimisation techniques Horvath et al., 2021, D. Wen et al., 2022. In general, PMT methods aim to reduce communication costs, improve training efficiency on resource-constrained clients, and address the effects of heterogeneous data by limiting the scope of federated updates to the most critical or adaptable parts of the model.

This leads to an important open research question: *How can we design learning frameworks that remain effective under high data heterogeneity across clients, particularly when client*

participation is random during global mode training?

To explore this question, this chapter investigates a lightweight and scalable federated learning approach. The method applies a pretrained DINOv3 network as a frozen backbone, while training a projection head across many clients. Traditional partial-model federated training, where different portions of the backbone may still be updated or selectively aggregated. In contrast to this, we separate feature extraction from the federated optimisation process: the entire backbone remains fixed throughout rounds and is never modified during model aggregation. Simply, the projection head is trained and updated in a distributed manner.

Freezing the backbone eliminates most of the computational and communication costs associated with full-model training, while preserving the robust, high-level representations learned by DINOv3 on large-scale vision data. Clients optimise the lightweight projection head using unsupervised contrastive learning on their local, unlabelled data, allowing practical adaptation to heterogeneous data distributions without requiring labels or large communication budgets. This design enables scalable, efficient, and robust representation learning even under severe client heterogeneity and partial participation.

A central focus of this chapter is its evaluation under large-scale federated conditions (cross-device), through stochastic client selection each round. This setup presents realistic deployments where client availability is unstable due to resource limitations, connectivity limitations, or temporal data access. Partial participation combined with non-i.i.d. data distributions typically undermines federated training.

This chapter addresses three main questions:

- To what extent can a frozen DINOv3 backbone support contrastive learning when only the projection head is updated in a federated environment?
- Can lightweight federated updates address the effects of non-i.i.d. and random client participation?
- Is this approach scalable to thousands of clients with diverse and unlabelled local data distributions?

In these investigations, we aim to demonstrate that high-quality pretrained backbones, combined with collaborative training, offer a promising approach to addressing core challenges in federated unsupervised learning. These challenges are data heterogeneity, partial federated model training and limited client availability.

6.2 Related Work

6.2.1 Federated Unsupervised Learning

Contrastive and representation-alignment-based methods are the most popular approaches in the unsupervised federated learning field. Recent methods, such as FedCA and FedSimCLR adapted centralised contrastive objectives to the federated setting by sharing a global encoder while clients compute local contrastive losses based on their own augmented samples. These approaches rely on instance-level discrimination, where positive pairs originate from augmented views of the same image and negatives are from other class samples within the batch. While effective on balanced local datasets, their performance degrades since class imbalance locally increases, and negative samples no longer follow a consistent semantic patterns across clients.

More papers address representation drift, the phenomenon in which local client encoders diverge over training due to highly heterogeneous data distributions. MOON Q. Li et al., 2021 introduced a representation-alignment loss that limited local updates' deviations from the previous global model, ensuring that local encoders remain close to a shared semantic anchor while shifting to their own data. Building on this idea, FedRCL (Seo et al., 2024) extends contrastive regularisation to a relational setting by comparing pairwise distances across local and global feature spaces, preserving latent space under non-i.i.d. conditions. FedCLF (Y. Tan et al., 2022) stabilises cross-client representations using feature-level alignment and cluster-aware contrastive objectives.

6.2.2 Self-supervised Learning

Self-supervised learning (SSL) has progressed extremely fast in recent years, allowing the learning of transferable representations without manual label annotation. Modern SSL approaches are broadly categorised into *discriminative* and *predictive* methods. Discriminative frameworks typically enforce invariance at the instance or cluster level. On the other hand, contrastive learning methods such as SimCLR (T. Chen, Kornblith, Norouzi, and Hinton, 2020) and MoCo (K. He et al., 2020) rely on large batch sizes or numerous negative samples to avoid collapse, assumptions that become infeasible on resource-constrained clients with limited and highly imbalanced non-i.i.d. Non-contrastive methods such as BYOL (Grill et al., 2020) and SimSiam (X. Chen and He, 2021) remove the dependency on negatives; however,they rely on batch statistics, architectural asymmetry, or momentum encoders for stability. Under heterogeneous data distributions, these approaches become unreliable, unstable and lead to inconsistent feature space and enhanced client drift.

Clustering-based approaches, including DeepCluster(Caron et al., 2018) and SwAV (Caron et al., 2020), assign samples to clusters and enforce consistency across cluster assignments. However, their balanced partitioning or balanced clusters break under federated non-i.i.d. set-

tings, where local clients may contain different or highly skewed class distributions, resulting in unstable or biased cluster assignments. Predictive approaches such as masked autoencoders (MAE) (K. He et al., 2022) or pretext-task learning (Gidaris et al., 2018) provide alternative training, however, they involve heavy encoder, decoder architectures or dense reconstruction tasks, which impose high computation and communication costs unsuitable for devices.

Teacher-student-based self-distillation frameworks provide a third group of these methods. DINO Caron et al., 2021 extends non-contrastive learning by leveraging momentum encoders to stabilise the training, yielding semantically varied features in centralised data settings. More recent large-scale variants such as DINOv2 (Oquab et al., 2024) and DINOv3 (Siméoni et al., 2025) enhance transferability through large curated datasets, refined regularisation, and improved training. Despite their performance in centralised training, these methods rely on predictable global model statistics, simple augmentations, and computational resources, making implementation in federated environments challenging. In addition, under non-i.i.d. data distributions, their teacher-student assumptions cause global model drift and degraded representation quality.

To sum up, discriminative, predictive, and distillation-based SSL methods (including the DINO versions) are highly useful in centralised data regimes; their dependence on large-scale data, many augmentations, and compute-heavy frameworks limits the applicability to federated settings. Thus, these limitations motivate the development of new SSL frameworks that remain stable under data heterogeneity, preserve discriminative feature space across random clients, and collaborate reliably under decentralised, label-scarce, and resource-constrained conditions.

6.2.3 Data Heterogeneity

Data distribution across clients in federated learning is commonly non-i.i.d., leading to significant statistical data heterogeneity. A wide range of methods has been proposed to address this challenge in recent years. Regularisation-based approaches, such as FedProx and MOON, aim to stabilise training by constraining local model updates. FedProx introduces a proximal term to keep local models closer to the global model, reducing divergence caused by heterogeneous client distributions. Also, MOON (Q. Li et al., 2021) applies a model-contrastive regularisation that penalises differences between local and global representations, encouraging a more consistent encoder across many clients. These methods reduce the mismatch between specific data distributions and improve global convergence under non-i.i.d. data conditions. More recent research directions, including FedDyn(Acar et al., 2021) and FedOpt (Reddi, Charles, et al., 2021), stabilise convergence via dynamic regularisation or adaptive server-side optimisation. A growing body of research in the field tackles data heterogeneity in the feature space. Model-contrastive learning (MOON) aligns local and global representations to reduce client drift, while FedRCL (Seo et al., 2024) sets relational or contrastive limitations to maintain accurate embedding space across clients. These methods are effective under label scarcity however, they still rely on similar augmentations and massive local computation.

6.3 Methodology

Our federated unsupervised learning framework includes two main components: local client training and global model aggregation. During local training, each client optimises a contrastive objective following the SimCLR paradigm, enabling the model to learn robust feature representations directly from its own unlabelled data. Rather than performing full model, end-to-end training, which is computationally expensive, we adopt a hybrid design in which a pretrained backbone is kept frozen and only a lightweight projection head is trained. This framework, illustrated in Figure 6.1, significantly reduces computational cost while improving model stability under partial client participation and non-i.i.d. regimes.

For feature extraction, we employ a pretrained DINOv3 model. DINOv3 is trained on large-scale, diverse image corpora using self-supervised objectives, producing robust, semantically correct, rich representations. By leveraging these high-quality features with federated settings, the projection head is efficiently fine-tuned with distributed heterogeneous data. In order to enable clients to adapt to their local data distributions without expensive backbone updates or labelled samples, this framework benefits from learned generalised features. We choose the DINOv3 model as the frozen pretrained backbone since it represents one of the more robust self-supervised visual encoders, producing generalisable and semantically correct features without relying on labelled data. DINO models, especially ViT-based variants, have been shown to learn global, object-centric representations, outperforming SOTA contrastive and supervised models in robustness, transferability, and generalisation across diverse visual domains.

These properties are valuable in federated environments, where data is highly heterogeneous, and clients have limited or unbalanced local datasets. By leveraging a pretrained DINOv3 encoder, trained on hundreds of millions of images, we transfer these representations to projection heads. Thus, this allows the federated frameworks to focus on training only a lightweight projection head (MLP), lowering computation and communication costs while maintaining optimal performance on downstream tasks.

6.3.1 Local Representation Learning via SimCLR

Each participating client receives the global model with the combination of a frozen DINOv3 encoder f_θ and a trainable projection head g_ϕ. Given an image $x \in \mathcal{D}_i$, from local unlabelled data, the client generates two augmented views,

$$\tilde{x}_1, \tilde{x}_2 = \mathcal{T}(x), \tag{6.1}$$

where $\mathcal{T}$ denotes the SimCLR augmentations (cropping, resizing, colour jittering, blurring, etc.). Both augmented views are passed through the frozen encoder to obtain embeddings as

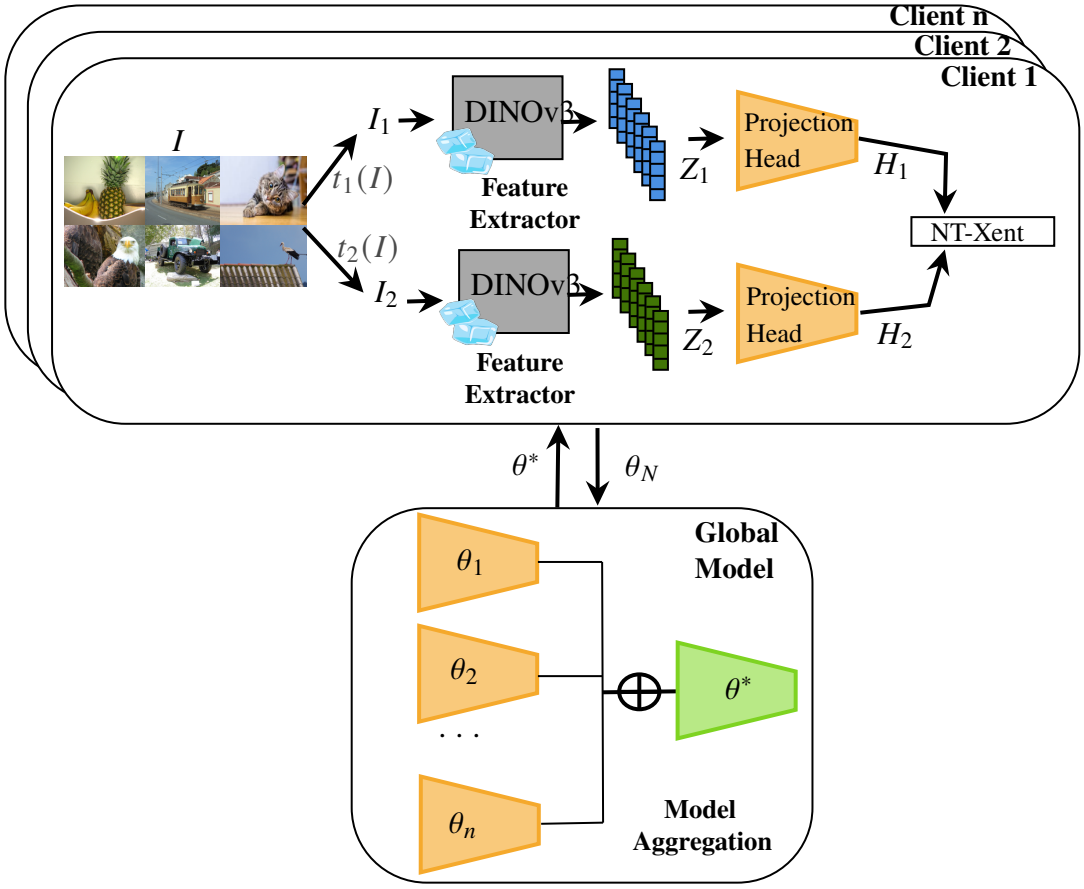


Figure 6.1: FedDinov3 framework, training with a frozen DINOv3 encoder and a 3-layer projection head. Two augmented views of the image are encoded into representations $\mathbf{t}_1, \mathbf{t}_2$, projected to contrastive embeddings $\mathbf{z}_1, \mathbf{z}_2$, and after projection head embeddings $\mathbf{h}_1, \mathbf{h}_2$ optimized with the NT-Xent loss. The global model aggregates client models using FedAvg.

input representations of the projection head.

$$\mathbf{z}_1 = f_\theta(\tilde{x}_1), \qquad \mathbf{z}_2 = f_\theta(\tilde{x}_2). \tag{6.2}$$

Following that, the projection head maps the input representations into a contrastive embedding space,

$$\mathbf{h}_1 = g_\phi(\mathbf{z}_1), \qquad \mathbf{h}_2 = g_\phi(\mathbf{z}_2). \tag{6.3}$$

The client's objective aims to minimise the NT-Xent loss,

$$\mathcal{L}_{\text{NT-Xent}} = -\log \frac{\exp(\text{sim}(\mathbf{h}_1, \mathbf{h}_2)/\tau)}{\sum_{k \neq 1} \exp(\text{sim}(\mathbf{h}_1, \mathbf{h}_k)/\tau)}, \tag{6.4}$$

where τ is the temperature and negatives are obtained from the batch. Most significantly, the backbone encoder f_θ remains frozen (parameters are not trainable), and only the parameters of the projection head g_ϕ are updated.

6.3.2 Federated Training Process

In typical FL frameworks, each communication round consists of three steps:

1. Broadcast The server sends the global model projection-head parameters ϕ^t to a subset of randomly selected clients. The DINOv3 encoder parameters θ remain stable during training.

2. Local Updates Each selected client performs multiple steps of stochastic gradient descent on its local dataset:

$$\phi_i^{t+1} \leftarrow \phi^t - \eta \nabla_\phi \mathcal{L}_{\text{NT-Xent}}(\mathcal{D}_i). \tag{6.5}$$

Since the pretrained encoder is frozen (no change in weights), local training becomes remarkably more memory-efficient. Table 6.1 shows that the frozen DINOv3 configuration trains the projection head (MLP), trainable parameters with full end-to-end training settings. Thus, this reduced-parameter training strategy is useful in non-i.i.d. federated learning, where prior work has shown that pre-trained networks show more remarkable stability under data heterogeneity (Y. Tan et al., 2022).

Limiting the local model complexity helps mitigate client drift and reduces the computational burden on resource-constrained devices. Furthermore, recent results from R. He et al. (2025) demonstrate that leveraging pre-trained embeddings yields more stable performance across a wide range of non-i.i.d. conditions, reinforcing the benefits of adopting a frozen backbone in federated self-supervised learning.

Model / Backbone	Parameters (Full)	Parameters (Frozen)
ResNet-18	$\approx$ 11.4M	428,416
ResNet-50	$\approx$ 23.9M	1,214,848
DINO v2 (ViT-S/14)	$\approx$ 21.7M	332,554
DINO v3 (ViT-S/16)	$\approx$ 29.0M	332,554

Table 6.1: Full trainable parameters of common backbones compared to the number of trainable parameters when the backbone is frozen and only the MLP projector (defined in Section 6.4.1) is trained.

3. Aggregation In our design, only the projection-head model parameters are sent back to the central server. Thus, the global model parameters are updated via the FedAvg aggregation algorithm,

$$\phi^{t+1} = \sum_{i \in C_t} \frac{n_i}{\sum_{j \in C_t} n_j} \phi_i^{t+1}, \tag{6.6}$$

where C_t represents the set of clients participating in each round t and n_i their training data size.

In realistic federated settings, client participation is unstable due to resource insufficiency and connection limitations. A simple participation matrix tracks which clients contribute in each

round. The proposed framework presents the following advantages:

- **Stability under non-i.i.d. :** Frozen encoder provides minimal cross-client drift.
- **Efficiency:** Instead of the full model training burden, partial model training with embeddings.
- **Label-free training:** Each local model is self-supervised.
- **High-quality features:** DINOv3 presents a fundamental and stable representational model, and federated optimisation of the projection head enables clients to calibrate these features with their local data.
- **Scalability:** Low communication and computation costs across many clients.

6.4 Evaluation

This section presents the evaluation criteria to compare representations learned by the proposed federated unsupervised SimCLR framework with recent works. Since the backbone encoder remains frozen, our evaluation focuses on analysing the *discriminative approaches*, *robustness*, and *transferability* of the feature space under severe non-i.i.d. federated data conditions.

6.4.1 Backbone Selection and Projection Head

Before implementing full model training-based federated experiments, we first evaluate multiple DINO variants (e.g., DINOv2, DINOv3 with ViT-S, ViT-B) to determine which encoder provides the most stable and transferable representations under federated data settings. Based on these fundamental experiments, we choose **DINOv3** as the frozen pretrained backbone due to its superior performance on many popular datasets like CIFAR10.

Following the encoder selection, we implement a lightweight but representative MLP, **projection head**, that is trained locally on each client. In our experiments, the projection head has a three-layer multilayer perceptron (MLP) with non-linear functions. The first layer maps the DINOv3 features (dimension d_{feat}) to a 512-dimensional latent space, followed by a LayerNorm operation, a GELU nonlinearity, and dropout with a rate of 0.2. The second layer transforms the 512-dimensional representation to 256 dimension vector and similarly applies LayerNorm with GELU, and dropout (rate 0.1). Finally, a linear layer maps the 256-dimensional hidden vector to the projection dimension used for contrastive learning loss. Thus, this network architecture ensures a simple reduction in feature dimensionality while preserving non-linear and regularisation effects.

In addition, this version of network architecture is highly flexible for adapting to many data scenarios with altered representations extracted by the DINOv3. Also, the presented algorithm benefits from affordable computation for large-scale data, cross-device or cross-silo cases.

6.4.2 Datasets and Augmentation Pipeline

We evaluate methods on three visual recognition benchmarks under many complex data scenarios: CIFAR10, which contains 10 classes with relatively low visual variability (low resolution, 32×32); CIFAR100, which contains 100 classes with higher intra-class diversity; and Tiny-ImageNet, a 200-class subset of large ImageNet that shows visual variation. To train the self-supervised projection head, we transformed each input image into two augmented views with various augmentations.

The data transformations include random resized cropping (cropping 20% of the image), random horizontal flipping, colour jittering, random grayscale transformation, and Gaussian blurring, followed by tensor conversion with ImageNet normalisation. This augmentations introduce the controlled variability required for contrastive learning, allowing the network models to learn robust representations. To evaluate representational robustness, we perform **cross-dataset evaluation**. After training the federated framework on CIFAR10, we test the learned global model on three datasets (CIFAR10, CIFAR100, Tiny-ImageNet). Our evaluation focuses on four key aspects of representation robustness: domain transferability, stability to unseen class distributions, the generalisation ability of the learned projection heads, and the effect of federated training on DINO-based representations. To evaluate methods' performance under realistic data heterogeneity, we provide two versions of label-skewed client datasets using Dirichlet partitioning: an imbalance and a non-i.i.d.

In the class imbalance scenario, Dirichlet splits the dataset without controlling the number of samples per client. In contrast, the balanced non-i.i.d. setting presents heterogeneous class proportions while preserving an equal number of samples per client. When the Dirichlet yields few samples for a class, samples are duplicated from other clients' data to ensure a similar dataset size. Thus, this strategy isolates the effect of label-distribution skew while preventing dataset-size imbalance in the analysis of federated contrastive learning behaviour.

6.5 Results

In this section, we present the experimental findings obtained from the proposed federated unsupervised SimCLR framework across multiple datasets. Our analysis mainly focuses on understanding how the frozen DINOv3 backbones, combined with a projection head, perform under severe non-i.i.d. distributions with(out) partial participations. We show these results on CIFAR10, CIFAR100, and Tiny-ImageNet, with domain generalisation experiments. Our methods evaluate performance with many metrics, such as linear probing accuracy, k-NN classification accuracy, and clustering.

When we achieve optimal supervised representation learning using the pretrained DINO variants in Table 6.2. Following that, DINOv3 performs as a more stable backbone network model for feature extraction when compared to earlier versions or other network architectures.

Method	Model	Proj. Head	top-1 acc.	FL settings (Clients, alpha, fraction)	Split
Supervised	DINOv2 (s-vit14)	1-linear	94.4	-	-
Supervised	DINOv3 (s-vit16)	1-linear	97.06	-	-
Supervised (FL)	DINOv2 (s-vit14)	1-linear	**85.97**	200C, 0.3, 5%	imbalance
Supervised (FL)	DINOv3 (s-vit16)	1-linear	69.56	200C, 0.3, 5%	imbalance
Supervised (FL)	DINOv2 (s-vit14)	1-linear	**90.92**	200C, 0.3, 5%	non-i.i.d.
Supervised (FL)	DINOv3 (s-vit16)	1-linear	84.01	200C, 0.3, 5%	non-i.i.d.
Supervised (FL)	DINOv2 (s-vit14)	1-linear	26.64	2000C, 0.3, 5%	imbalance
Supervised (FL)	DINOv3 (s-vit16)	1-linear	26.72	2000C, 0.3, 5%	imbalance
Supervised (FL)	DINOv2 (s-vit14)	1-linear	71.67	2000C, 0.3, 5%	non-i.i.d.
Supervised (FL)	DINOv3 (s-vit16)	1-linear	55.31	2000C, 0.3, 5%	non-i.i.d.

Table 6.2: Train head with CIFAR10. Frozen backbones (DINO versions)

Method	Model	Proj. Head	top-1 acc.	FL settings (Clients, alpha, fraction)	Split
Supervised	DINOv2 (s-vit14)	MLP	92.12	-	-
Supervised	DINOv3 (s-vit16)	MLP	95.97	-	-
Supervised (FL)	DINOv3 (s-vit16)	MLP	92.98	200C, 0.3, 5%	imbalance
Supervised (FL)	DINOv3 (s-vit16)	MLP	93.26	200C, 0.3, 5%	non-i.i.d.
Supervised (FL)	DINOv3 (s-vit16)	MLP	88.51	2000C, 0.3, 5%	imbalance
Supervised (FL)	DINOv3 (s-vit16)	MLP	91.89	2000C, 0.3, 5%	non-i.i.d.
Unsupervised (FT)	DINOv3 (s-vit16)	MLP	90.08	2000C, 0.3, 5%	imbalance
Unsupervised (FT)	DINOv3 (s-vit16)	MLP	**94.83**	2000C, 0.3, 5%	non-i.i.d.
Unsupervised (NO FT)	DINOv3 (s-vit16)	MLP	11.41	2000C, 0.3, 5%	imbalance
Unsupervised (NO FT)	DINOv3 (s-vit16)	MLP	12.40	2000C, 0.3, 5%	non-i.i.d.

Table 6.3: Train head with CIFAR10. Frozen backbones (DINOv3)

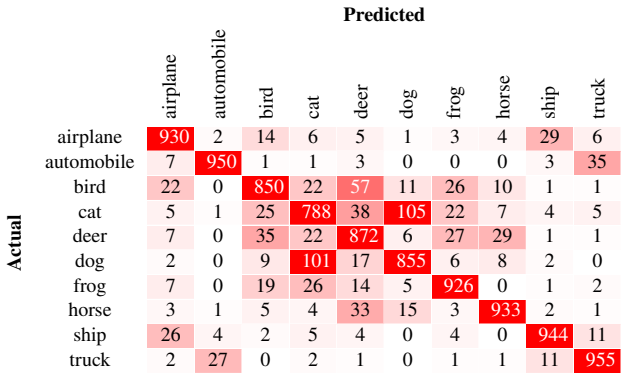
Table 6.2 shows that training a single linear layer-based classification causes underfitting, whereas a multi-layer perceptron is able to capture high-level representation. Based on these results, we show Table 6.3, with a three-layer MLP as the projection head, as described in the model section of this chapter.

Table 6.4 presents a comparison of three datasets evaluated under the federated metric-learning FedDINOv3 frameworks for image classification. The model is trained on CIFAR10 and then evaluated on several different datasets. The accuracy scores clearly show that federated partial model training is extremely successful on unseen domains. It explains that robust features generalise well to future, unseen domains. In the table, when we compare the accuracy of FedDINOv3 with FedQuad on Tiny-ImageNet, our method has notably more robust representations, achieving up to 60.89% accuracy with 200 clients with 1% client participation. Table 6.5 presents the representation quality evaluated on unseen domains. Although the federated partial model training is performed on the CIFAR10, the learned representations are evaluated on CIFAR100, CelebA-Gender, and Tiny-ImageNet. The k-NN classifier outperforms across these unseen domains, showing that the learned embeddings generalise beyond their training

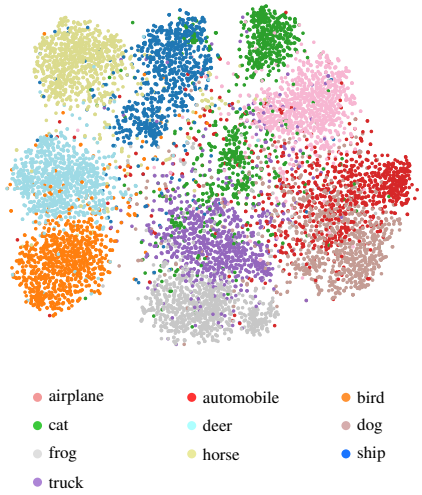
Participation %	Method	CIFAR10		CIFAR100		Tiny-ImageNet	
		200C	2000C	200C	2000C	200C	2000C
1%	FedAvg	31.24±3.64	34.24±3.18	9.08±1.08	8.59±0.73	5.11±0.23	3.97±0.32
	SupCon-FL	44.29±0.27	43.67±0.44	16.84±0.2	17.21±0.28	11.56±0.31	9.93±0.31
	Triplet-FL	45.35±0.6	45.5±0.53	17.53±0.3	16.03±0.34	12.47±0.51	10.57±0.30
	Quadruplet-FL	52.96±0.23	50.61±0.29	16.01±0.54	15.92±0.53	14.92±0.24	11.55±0.30
	MOON	28.99±5.24	15.59±0.58	5.16±0.59	1.45±0.16	4.80±0.09	3.97±0.34
	FedQuad	53.03 ±0.81	50.39±0.17	23.3±0.36	18.95±0.37	15.00±0.16	11.81±0.26
	FedDINOv3	**90.93±0.39**	**90.86**±0.18	**69.38**±0.22	**69.36**±0.08	**61.05**±0.08	**60.89**±0.14
5%	FedAvg	47.76±2.1	44.43±1.01	17.69±0.39	10.89±0.69	11.17±0.16	5.13±0.24
	SupCon-FL	45.44±0.34	43.60±0.55	17.95±0.25	17.32±0.2	12.10±0.14	10.04±0.32
	Triplet-FL	51.18±0.21	46.72±0.27	20.37±0.27	16.62±0.29	14.01±0.4	11.16±0.33
	Quadruplet-FL	57.62±0.29	51.75±0.23	18.81±0.24	16.97±0.04	14.89±0.12	11.81±0.27
	MOON	33.94±3.24	15.09±1.92	8.33±0.26	2.13±0.13	11.14±0.50	5.11±0.18
	FedQuad	58.21±0.36	51.86±0.39	26.47±0.27	19.23±0.6	15.25±0.07	12.20±0.39
	FedDINOv3	**90.89**±0.42	**92.33**±1.81	**69.42**±0.11	**68.81**±0.35	**60.99**±0.12	**60.86**±0.09

Table 6.4: Method comparison on CIFAR10, CIFAR100 and Tiny-ImageNet with varying participation fraction and client scale.

data distribution. In contrast, the k-means clustering results show limited generalisation. Also, these representations are inseparable for unsupervised clustering in more complex domains.



(a) Confusion matrix of model predictions.



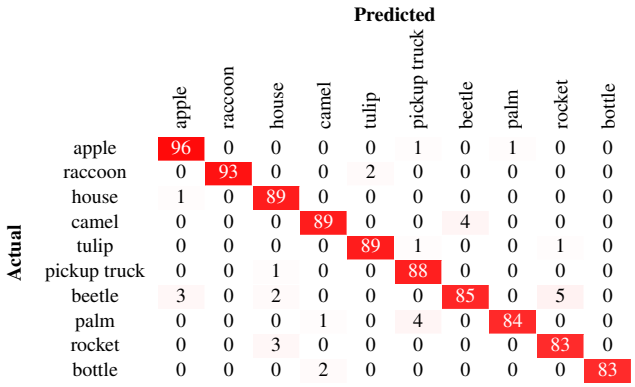
(b) t-SNE projection of the embedding space.

Figure 6.2: Illustration of the confusion matrix (after round 20) and t-SNE visualisation (after round 5) for the CIFAR10 dataset using FedDINOv3 representations. The model is trained on CIFAR10.

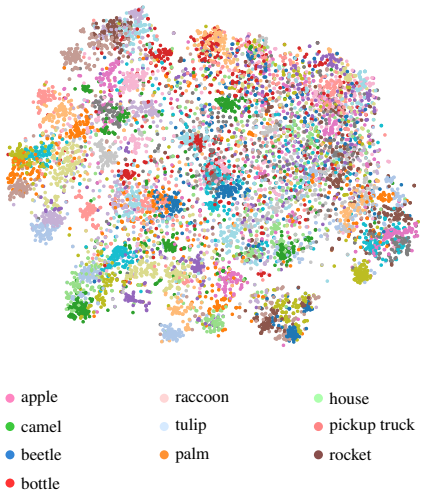
t-SNE plots and confusion matrices in these Figures 6.2,6.3, and 6.4 illustrate the evolution of representations across communication rounds. In these experiments, the best performance

Participation %	Data	Supervised	Unsupervised	
			k-NN	k-means
1%	CIFAR10	90.86±0.18	88.87±0.27	72.8±2.8
	CIFAR100	69.36±0.08	65.08±0.58	38.33±0.95
	Tiny-ImageNet	60.89±0.14	56.94±1.07	26.35±0.34
	CelebA-Gender	95.32±0.73	94.16±0.54	89.52±1.47
5%	CIFAR10	92.33±1.81	88.83±0.26	74.59±3.20
	CIFAR100	68.81±0.35	64.99±0.45	37.03±0.39
	Tiny-ImageNet	60.86±0.09	56.46±0.96	26.01±0.31
	CelebA-Gender	95.25±0.23	94.67±0.33	90.35±0.48

Table 6.5: We compare methods on CIFAR10, CIFAR100, and Tiny-ImageNet under varying client participation rates and client scales, evaluating both unsupervised and supervised training regimes. These results summarise a comprehensive set of experiments in which models are trained on CIFAR10 using 2,000 clients with participation rates of 0.5 and 0.1, and unseen target-domain datasets.



(a) Confusion matrix of model predictions.

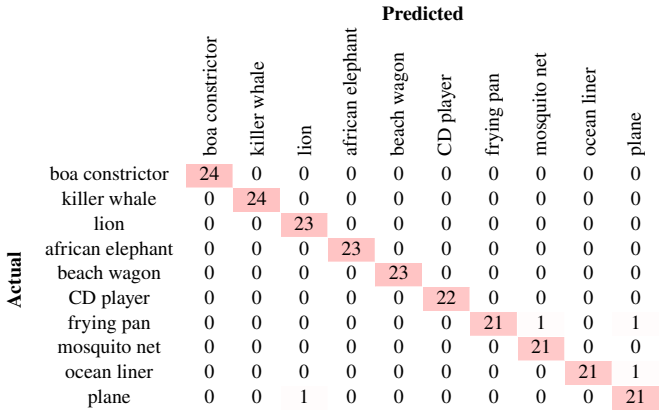


(b) t-SNE projection of the embedding space.

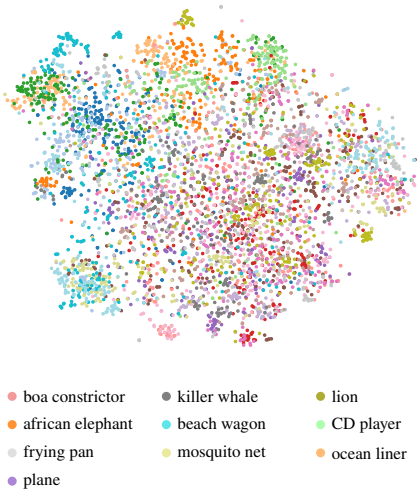
Figure 6.3: Illustration of the confusion matrix (after round 20) and t-SNE visualisation (after round 5) for the **CIFAR100** dataset using FedDINOv3 representations. The model is trained on CIFAR10. It shows the first 10 classes with the highest accuracy among the 100 CIFAR100 classes.

is measured on CIFAR10. In addition, t-SNE projections and confusion matrices show well-separated clusters and reliable class discrimination.

When evaluating on more complex data (Tiny-ImageNet), early-round t-SNE projections show that multiple classes overlap or partly collapse due to higher intra-class variance and domain shift. Despite this collapse in the embedding space, federated partial training leads to more compact representations, which is reflected in these confusion matrices that demonstrate



(a) Confusion matrix of model predictions.



(b) t-SNE projection of the embedding space.

Figure 6.4: Illustration of the confusion matrix (after round 20) and t-SNE visualisation (after round 5) for the **Tiny-ImageNet** dataset using FedDINOv3 representations. It shows the 10 classes with the highest accuracy among the 200 Tiny-ImageNet classes. The model is trained on CIFAR10.

classification accuracy over later rounds.

6.6 Discussion

This chapter investigates a partial model federated learning paradigm under highly diverse data distributions and large client populations. In realistic cross-device environments, only a small fraction of clients participate at each round, and full-model training becomes computationally expensive and usually impractical. In this chapter, our main aim is to investigate whether robust representation learning can be achieved without from scratch client model training.

In Chapter 5, we demonstrate that addressing representational collapse plays a critical role in addressing data heterogeneity and improving the robustness of learned representations. However, those methods relied on labelled data and full-model optimisation, which becomes impractical when scaling to thousands of clients. Also, federated training of deep encoders requires significant compute, communication, and memory resources, creating bottlenecks in large-scale non-i.i.d.

Motivated by recent advances in self-supervised vision models, especially the remarkable pattern-capturing performance of DINO models on many benchmarks. By leveraging a pretrained DINOv3 backbone as a feature extractor, clients optimise a small set of layers through contrastive representation learning on unlabelled local data. Thus, this approach considerably reduces computation cost per client while providing a stable global representation space.

Limitations. Despite this approach's robustness, the proposed methods have some limitations. First, the projection head has a limited capacity and cannot be fully accurate for local domain shifts caused by noisy data and image distortion. Second, freezing the backbone improves local training; the global model aggregation of projection heads can still introduce misalignment when clients have highly heterogeneous feature distributions.

Altogether, our main findings suggest that federated unsupervised learning with a frozen DINOv3 backbone provides a stable and computationally efficient baseline for many representation learning tasks under several heterogeneous data distributions. While this strategy cannot surpass end-to-end self-supervised training, it shows a remarkable balance between performance, representation robustness, and implementation efficiency. Also, this makes it advantageous for large-scale cross-device implementations, limited resource-based federated settings, and distributional environments where full-model training is impractical.

Future Work. A key observation from our experiments is that global model aggregation at every communication round can still be problematic under data heterogeneity, as the FedAvg assumes similar data distribution across clients' datasets. Thus, global model aggregation huge obstacle to representation learning in distributed systems. To address this limitation, future research may explore peer-to-peer representation learning paradigms such as swarm learning (Shammar et al., 2024) or gossip learning (Belenguer et al., 2025), which minimise a central model aggregator. Integrating these approaches into distributed self-supervision or unsupervised learning may yield more adaptive representation learning algorithms under realistic, highly non-i.i.d. data distributions.

6.7 Conclusion

This work explored federated self-supervised representation learning using a frozen DINOv3 backbone combined with a projection head. The proposed framework aims practical contrastive learning method in distributed settings while avoiding the computational cost of training large models on non-i.i.d. local data. By freezing the encoder layers and updating the projection head parameters, the process achieves efficiency, model stability, and generalisation.

The results demonstrate that robust features appear where robust embeddings from pretrained backbones. The frozen DINOv3 encoder outputs representative features, which reduces local model training computation and time. As a result, both standard Dirichlet and balanced Dirichlet partitions yield similar performance.

A comprehensive evaluation with linear probing, k-NN classification, clustering, silhouette scores, and prototype-based accuracy ensures that the learned representations preserve class separation. In summary, this study shows that federated self-supervision with a frozen encoder network is quite practical and robust for distributed learning. Also, it provides a reliable baseline

for federated SSL and a highly encouraging research direction for future work.

Chapter 7

Conclusion

This chapter reviews the key contributions, findings, and outcomes of the research. It focuses on developing representation learning methods in contexts indicated by diverse, limited, distributed, and unlabelled data scenarios. The chapter begins with a summary of this thesis's main contributions in relation to the research questions (mentioned in Chapter 1). Then, it discusses the limitations of the proposed approaches and the challenges encountered during the investigation. Finally, this section outlines potential research directions for the future that aim to extend and improve the current research.

7.1 Summary of Contributions

This thesis mainly investigated the challenges of visual representation learning from limited, distributed, and diverse data under both supervised and unsupervised settings. The studies conducted throughout the thesis addressed the main research questions, demonstrating how representation learning and federated approaches can be effectively applied. The thesis deeply investigates the research gaps of existing literature and presents four main approaches to handle many real-world situations. These contributions and the mentioned research questions are mentioned below:

- ***RQ1: How can we improve self-supervised contrastive learning methods under limited data conditions, where augmentations may introduce misleading (false positive/negative) views?***

  Chapter 3 presents an SSL method for representation learning under conditions of limited unlabelled data, optimising the self-supervised contrastive learning framework during training. In many existing SSL methods, the decision of dataset-specific augmentations is extremely time-consuming and requires extensive manual calibration. Also, most contrastive learning approaches rely heavily on augmented views of each image sample, but

there is no comprehensive evaluation of whether these augmentations are sufficiently robust or appropriate for effective representation learning.

In this chapter, we analyse the selection of training batches and categorise them as either "good" or "bad". Bad batches typically contain insufficient or low-quality augmented views of original images, which can hinder representation learning by introducing noisy or misleading features. To address this limitation, we reformulate data-sample similarity using the Fréchet ResNet Distance (FRD), which is a form of FID (Fréchet Inception Distance). FRD mainly represents the batch with properly transformed views. Lower FRD shows the batch contains semantically correct transformed views. As a result, our findings show that removing misleading augmented views during training allows the model to focus on more robust features and semantically correct representations, in the end improving performance within fully unsupervised learning frameworks.

- ***RQ2: How can covariate shift in federated learning be mitigated to enable robust and generalizable training when clients possess limited and non-overlapping datasets?***
  Chapter 4 addresses the problem of covariate shift in distributed and limited data environments. Covariate shift poses a significant challenge in federated learning scenarios, where data are distributed across multiple clients or devices, each local model possessing different local distributions and without the ability to share raw data. This heterogeneity usually leads to inadequate global model aggregation, suboptimal model performance and degraded representation learning.

  To tackle these issues, this chapter proposes a parameter selection–based federated approach that eliminates redundant or irrelevant weights. Also, the method benefits from features that are specific to each client, eliminating the obstacle of global model aggregation. The method reformulates the iterative magnitude pruning (IMP) algorithm into a round-wise framework, enabling incremental model sparsity across communication rounds rather than after each local training phase. Also, in order to prevent potential overfitting caused by parameter pruning locally, extra regularisation techniques such as noise injection and dropout layers are incorporated into the client models.

  Our experiments show that magnitude-based parameter pruning during federated training enhances models' (global and local) robustness and representation learning, when using typical model aggregation algorithms such as FedAvg. Similarly, in order to evaluate the proposed approach under more realistic conditions, a novel benchmark (CelebA-Gender) was introduced to compare existing FL methods and design more complex data distributions. Unlike traditional one-attribute settings (e.g., "smiling" vs. "not smiling"), this new dataset considers multiple facial attributes and groups facial factors to define heterogeneity across clients, under gender classification tasks.

- ***RQ3: What is the impact of the learning distance between samples within each local***

model on the global model's generalisation?

Chapter 5 investigates the problem of discriminating between samples during representation learning. This chapter demonstrates that minimising intra-class distances while maximising inter-class variance can effectively address representation collapse and data heterogeneity in FL. Data heterogeneity is a major challenge in federated settings, when each client can learn locally robust representations. The global model aggregation step usually leads to representation collapse in the global embedding space.

To overcome this issue, we propose a federated metric learning algorithm that enforces compact intra-class clusters and well-separated inter-class boundaries. Specifically, we introduce a reformulated quadruplet loss function that operates on four samples: an anchor, a positive, and two different negatives. Unlike the traditional quadruplet loss, which simultaneously encourages the positive to be close to the anchor and evaluates negative to negative distance, our formulation places particular emphasis on controlling the distances between the positive and each negative. Additionally, significant negative distances can be disadvantageous in federated settings: when client-specific feature spaces diverge extremely deeply, global model aggregation can lead to representational collapse. To solve this limitation, our modified loss controls negative distances to remain within a stable margin while enforcing positive alignment. Thus, the assumption benefits from improving cross-client consistency and preventing global space-based representational collapse.

Unlike the traditional triplet loss, which relies on a single positive–negative pair, the proposed loss increases the dissimilarity between the positive and two negative samples at the same time. This new formula promotes the same class discrimination and more robust representations. As a result, the global model learns to cluster similar samples more closely while pushing dissimilar samples further apart. Also, it leads to improved global representation quality after model aggregation, even under highly non-i.i.d data.

- ***RQ4: How can we design learning frameworks that remain effective under high data heterogeneity across clients, particularly when client participation is intermittent or partial during global aggregation?***
  Chapter 6 addresses the problem of data heterogeneity in cross-silo federated learning under small client participation rates, such as 1%, 5%. Training a full model from scratch locally is computationally expensive, longer in communication rounds, and highly sensitive to model aggregation-related effects. To avoid these limitations, we implement a partial model federated training strategy in which the pretrained backbone remains frozen while the projection head parameters are updated during federated optimisation. Thus, this approach significantly reduces computational cost, improves communication round

efficiency, and limits the effects of model drift.

In literature, several works point to the optimal performance of DINOv3 Siméoni et al., 2025 in image-based self-supervised representation learning, we employ a pretrained DINOv3 as a backbone and train only the lightweight projection head (MPL) to compute the NT-Xent loss. Beyond evaluating on the same domain, similar data distributions, we further compare the learned representations on unseen datasets, including more complex benchmarks such as CelebA-Gender and Tiny-ImageNet. As a conclusion, our findings demonstrate that without full local model training, partial model updates on top of a frozen backbone yield robust and transferable representations across diverse visual domains.

7.2 Limitations

The following section discusses the important limitations of this thesis, clarifying the assumptions and specific conditions that specify the applicability of the proposed methods.

- **Sample Scarcity in Federated Metric Learning**

  Federated metric learning faces fundamental limitations arising from label scarcity and heterogeneous data distributions across clients. In typical federated settings, each client holds only a small and non-uniform subset of the global dataset, frequently exhibiting severe class imbalance and missing labels. Metric learning objectives such as contrastive, triplet, and quadruplet losses depend on structured sample pairings, where an anchor, positive, and one or more negative samples optimise intra-class variance minimisation and inter-class variance maximisation within the embedding space.

  However, under conditions of label scarcity, the probability of developing valid and diverse positive or negative pairs causes performance degradation. Let each class c_i contain n_i labelled samples, and let there be C total classes in the dataset. The number of possible positive pairs within class c_i can be computed as:

$$N_{\text{pos}}^{(i)} = \binom{n_i}{2} = \frac{n_i(n_i - 1)}{2}. \tag{7.1}$$

  Similar to this, the number of negative pairs between classes c_i and c_j is represented by:

$$N_{\text{neg}}^{(i,j)} = n_i \cdot n_j. \tag{7.2}$$

  For the typical quadruplet loss, which involves one anchor a, one positive p, and two negatives n_1, n_2, the total number of possible quadruplets across each of the classes is like

this:

$$N_{\text{quad}} \approx \sum_{i=1}^{C} \binom{n_i}{2} \sum_{\substack{j=1 \\ j \neq i}}^{C} \binom{n_j}{2}. \tag{7.3}$$

Under federated data conditions, where n_i is limited for most clients, both $N_{\text{pos}}^{(i)}$ and N_{quad} rapidly approach zero. Thus, the optimisation of the metric objective becomes under-constrained, as there are insufficient positive and negative combinations to compute a gradient. This *integrative collapse* leads to unstable training, poor discrimination between classes, and a high risk of representational collapse in the global space.

Due to rigid privacy constraints, clients are not permitted to share raw data or labels, which prevents the global model generalisation of diverse positive and negative pairs. Each client relies on its local data, which frequently contains a limited number of classes and a high proportion of *hard negatives*, samples that are conceptually similar but semantically different. These hard negatives dominate the loss, change the inter-class distance, and hinder model convergence.

In our presented dataset, CelebA-Gender, the gender classification task (female vs. male) limits the applicability of certain metric-learning objectives. For example, federated metric learning methods such as the quadruplet loss or our method (FedQuad) require at least three specific classes to generate anchor–positive–negative pairs. Since CelebA-Gender contains two classes, it is not possible to create quadruplets, limiting the use of our FedQuad on this benchmark. Thus, this limitation highlights a shortcoming of metric-learning approaches when applied to binary or two-class scenarios.

In summary, label scarcity and classes fundamentally limit the theoretical scope of traditional metric learning objectives in centralised or decentralised learning environments. The insufficient number of pairs renders a triplet and quadruplet loss under highly imbalanced, distributed conditions. This motivates the development of reformulated metric objectives that remain robust when label density is not an obstacle.

- **Global Model Aggregation Considerations**

  Numerous model aggregation strategies have been proposed in the federated learning literature, including *FedAvg* McMahan et al., 2017, *FedProx* Sharma et al., 2022, and several adaptive or regularised variants such as *FedNova* and *FedOpt*Qi et al., 2024. These approaches differ in how they compute the global model parameters from locally updated client models, influencing convergence speed, generalisation capability, and stability under heterogeneous data distributions.

  In this thesis, we mainly focus on the *local training methodology* rather than on the specific concept of model aggregation. In order to provide a fair and consistent evaluation

of the proposed local representation learning approaches, each experiment utilises the standard but effective *Federated Averaging (FedAvg)* algorithm as the global aggregation rule. FedAvg performs simple parameter averaging across clients. In mathematical terms, K clients with local model parameters $\mathbf{w}_i^{(t)}$ at round t, and n_i samples per client, the global model parameters are computed as:

$$\mathbf{w}_g^{(t+1)} = \sum_{i=1}^{K} \frac{n_i}{N} \mathbf{w}_i^{(t)}, \quad \text{where } N = \sum_{i=1}^{K} n_i. \tag{7.4}$$

This model aggregation rule effectively computes a weighted mean of the client parameters, giving more influence to clients with larger local datasets. While simple and computationally efficient, FedAvg assumes that client updates are closely aligned in the parameter space. Under many heterogeneous situations, this assumption is usually neglected, leading to suboptimal convergence or model drift, as local models may diverge toward distinct minima.

In this thesis, we intentionally retain the standard FedAvg strategy to evaluate the effect of the proposed local training methodologies. Thus, the observed improvements are directly related to the local learning algorithms rather than to any advances in the aggregation step. Future extensions of this thesis may explore adaptive model aggregation strategies or personalised federated learning frameworks to improve the robustness of model updates under non-i.i.d. data distributions.

Another alternative direction may eliminate the global model aggregation step by implementing decentralised federated training approaches such as gossip learning or swarm learning Hegedűs et al., 2021. These methods rely on a peer-to-peer model exchange paradigm rather than a central server, and operate through predefined communication topologies such as ring, mesh, or random graph structures to transfer parameters across clients Song et al., 2022.

- **Data Bias in Gender Classification with Attribute Diversity**

  In Chapter 4, we introduce a gender classification dataset (CelebA-Gender) specifically designed to evaluate more complex and realistic federated learning scenarios. The dataset incorporates multiple different facial attributes, such as black hair, to represent diversity across samples and to simulate heterogeneous real-world data conditions. This design enables the understanding of how attribute-level variations influence local representation learning, global model convergence, and representational robustness in federated environments.

  The dataset is reproducible and can be extended by selecting alternative combinations of attributes, allowing future research to investigate additional aspects of data heterogeneity.

However, as discussed in Chapter 4, only a limited number of attribute configurations were explored in this research. Various attribute combinations may impact performance and inter-client variability in many ways. The main objective of this dataset is to demonstrate the influence of high attribute overlap within gender classification tasks and to assess its effect on model performance under non-i.i.d. distributions.

From a theoretical perspective, integrating attribute diversity introduces higher similar class variance, increasing the complexity of the underlying data manifold. In federated learning, diversity acts as a valuable benchmark for evaluating generalisation under covariate shifts and data heterogeneity. Diverse attribute combinations lead to biased feature distributions across clients, providing a more realistic and challenging heterogeneity to evaluate the robustness of aggregation strategies and local representation learning methodologies. Thus, this dataset not only facilitates performance evaluation under realistic data heterogeneity but also highlights the need for federated models that can effectively handle multi-attribute and imbalanced data scenarios (attribute overlaps).

- **Experimental Setup and Data Partitioning Strategy**

  In this thesis, we investigate the effect of limited data availability across multiple representation learning scenarios, including labelled, unlabelled, distributed labelled, and distributed unlabelled settings. Thus these scenarios are designed to broadly evaluate the representation robustness and adaptability of the proposed methods. The experiments are implemented on several popular benchmark datasets, including CIFAR10, CIFAR100, and Tiny-ImageNet, among others, to ensure both diversity and generalisation of the results.

  In the federated learning designs, the number of participating clients significantly influences the computational cost and communication efficiency. A larger number of clients requires sufficient memory and synchronisation, especially under full client participation. Therefore, we provide results with up to 200 clients under full participation and extend the evaluation to 2000 with 200 clients using partial participation (a subset of clients randomly selected in each communication round).

  To simulate limited and heterogeneous data conditions, we follow both manual and probabilistic data partitioning strategies. Label scarcity is introduced manually by reducing the number of labelled samples per client to correspond to real-world cases of supervised learning in distributed environments. In addition, data heterogeneity is controlled using a Dirichlet distribution-based partitioning, parameterised by the α. This α parameter defines the degree of non-i.i.d. data distribution across clients, where lower α values correspond to higher data diversity and high data heterogeneity. Specifically, we experiment with $\alpha \in [0.1, 0.6]$, enabling a controlled analysis of model performance under varying levels of distributional skewness.

  These experimental settings allow us to evaluate the effectiveness of the proposed methods

under diverse data conditions, including limited data, high client variability, and high non-i.i.d. data distributions.

- **Scalability to larger datasets or systems** This thesis focuses on developing a limited data-based framework for robust representation learning. While several approaches exist to increase the number of training samples such as data augmentation and generative models, these methods have inherent limitations. Data augmentations are often data-specific and mostly introduce synthetic data variations. On the other hand, generative models can produce synthetic data that may mislead the learning process, as the model can become biased toward the generated data distribution, finally limiting its ability to learn reliable and generalizable representations.

  Therefore, this thesis emphasises the development of methods that are inherently designed for limited data scenarios. In addition, the proposed framework remains scalable to larger datasets, with the cost of increased memory consumption and longer GPU training times. One of the main limitations of the proposed approach (Chapter 1) lies in the data augmentation-based pair construction process. Identifying reliable pairs that lead to lower Feature Representation Distance (FRD) scores can be computationally expensive and time-consuming, making batch construction inefficient. In contrast, other presented approaches may perform more efficiently on large-scale datasets. Additionally, metric learning-based method FedQuad require datasets with at least three classes to effectively learn distances (e.g., quadruplets). Thus, this requirement limits its applicability in scenarios where the number of classes is very small.

- **Sensitivity to dataset bias** While the proposed frameworks demonstrate optimal performance on the selected benchmark datasets, it remains sensitive to inherent dataset biases. In particular, the training and evaluation are mostly evaluated on widely used academic benchmarks such as CIFAR-10, CIFAR-100, and TinyImageNet. Although these datasets are considered standard in the literature, they may introduce selection bias, as real-world data is typically more diverse, noisy, and dynamically changing. In addition, this thesis does not evaluate the proposed methods on specific datasets such as medical images, biometric data, remote sensing, or fully synthetic data, which commonly exhibit significantly different statistical characteristics. As a result, the generalizability and scalability of the model to broader, real-world applications remain uncertain. To address these limitations, future work may incorporate a wider range of heterogeneous, real-world datasets. Evaluations may provide stronger evidence of the model's ability to maintain performance in high-variance and non-standardised data environments.

- **Privacy-Preserving Limitations in Federated Systems** A common misconception in distributed systems is that Federated Learning (FL) is inherently privacy-preserving. While FL effectively addresses challenges related to data locality, ownership, and decentralised

computation by ensuring that raw data remains on non-standardised devices. However, it does not provide a legal or firm guarantee of data privacy. A key limitation of FL arises from the communication of model updates, gradients, and model parameters between clients and the central model aggregator. These updates introduce non-trivial attacks, as they may decentralised encode sensitive information about the underlying local datasets. Recent studies have demonstrated that adversaries can exploit information through gradient inversion, model reconstruction, and inference attacks, reconstructing input data or inferring participation of specific samples. In this context, gradients can act as high-dimensional proxies for private data, undermining the fundamental data privacy assumptions of FL systems. Additionally, the current thesis does not incorporate standardised privacy guarantees. As a result, the proposed frameworks may be vulnerable in some sensitive domains such as healthcare or finance, where data confidentiality is critical. This limitation highlights a gap between theoretical distributed learning benefits and real-world privacy requirements.

To address these challenges, a critical direction for future research lies in the integration of Privacy-Enhancing Technologies (PETs) into federated learning. One promising approach is Differential Privacy (DP), which introduces calibrated noise to model updates to provide quantifiable privacy guarantees (Auñón et al., 2024). However, this comes at the cost of a trade-off between privacy and model performance, requiring careful tuning of the privacy budget. In addition, cryptographic techniques such as Secure Multi-Party Computation (SMPC, Byrd and Polychroniadou, 2020) and Homomorphic Encryption (Alqazzaz, 2025) offer extra privacy guarantees by enabling secure aggregation of model updates without exposing individual contributions. These methods ensure that the central server operates only on encrypted data, thereby reducing the risk of information leakage. Future work may focus on designing hybrid frameworks that balance privacy, efficiency, and model performance. In particular, integrating metric learning-based approaches (such as FedQuad) with privacy-preserving mechanisms presents an open research challenge, as the structure of contrastive or quadruplet losses may cause information leakage through embeddings. Thus, bridging the gap between data locality and complete data privacy requires a balanced approach that combines algorithmic, statistical, and cryptographic techniques, ensuring that federated learning systems are not only distributed but also secure and private.

7.3 Future Directions

In this thesis, we address the challenges associated with limited data in both traditional centralised learning and centralised federated learning scenarios. These cases are becoming important and widely discussed in the modern machine learning field, as data scarcity and distributed data without sharing raw data are common in many real-world applications. Domains such as remote sensing and medical imaging are difficult to collect, highly sensitive to network architectures,

and require expert annotations. Thus, developing representation learning algorithms that can operate effectively under constraints has become a critical research direction.

Our findings demonstrate that when the number of samples is limited and the data distribution is highly heterogeneous, due to unseen classes, class imbalance, or diverse client datasets. Our proposed methods are capable of maintaining robust representations, optimal performance and stable learning. These results indicate that the proposed approaches can effectively handle complex non-i.i.d. and data-scarce conditions that frequently occur in realistic distributed environments. However, FL still cannot guarantee data privacy or solve data leakage problems. An interesting and valuable future direction may be focusing on secure data privacy challenges for distributed systems.

7.3.1 Open Research Directions

In addition to these findings presented in this thesis, several promising research directions for future investigation exist. One potential direction is the exploration of *partial federated model aggregation* strategies. Unlike full local model training from scratch, partial model representation learning selectively updates subsets of each local model parameter or specific layers. This approach can significantly reduce communication costs and may preserve client-specific knowledge by preventing locally adapted representations. Investigating layer-wise or feature-level model training may enhance model personalisation while maintaining global coherence across heterogeneous clients.

Another promising research direction involves *synthetic data generation through latent-space federated learning*. In this paradigm, clients collaboratively learn a shared latent representation that captures the global data distribution without sharing raw samples. Using generative models such as variational autoencoders (VAEs) or diffusion-based frameworks, clients can generate synthetic data that approximates the global manifold. Also, this synthetic data can be used to augment local datasets or to facilitate more stable global model training. Latent-space generation techniques may be valuable in some domains where data are scarce or sensitive, including medical imaging, remote sensing, biometrics, autonomous vehicles, and industrial quality control.

Last but not least, research directions may be *serverless federated learning* or decentralised federated learning paradigms such as gossip learning or swarm learning. Unlike federated learning, which mainly relies on a central server for global model aggregation, gossip-based approaches enable peer-to-peer parameter exchanges, where each client iteratively communicates with its neighbours.

Integrating partial model aggregation, generative models collaboration, and gossip-based decentralisation, federated learning can significantly advance federated learning toward efficient communication and data accessibility problems. These directions represent key opportunities for the next generation of scalable distributed learning frameworks.

7.4 Applications of the Proposed Research

The methodologies developed in this thesis, which address representation collapse, data heterogeneity, and label scarcity in representation learning, have broad applicability across multiple domains. The following areas highlight some of the most relevant applications:

1. Medical and Healthcare Systems. Federated learning provides a powerful framework for collaborative model training across hospitals and medical institutions without sharing raw patient data or medical records. The proposed metric-based federated approaches can improve generalisation across different hospitals with varying patient details, imaging devices, or annotation quality. By addressing representation collapse and improving inter-class variance (client-specific samples), the methods can support diagnostic tasks such as disease detection, medical image classification, or pathology segmentation.

2. Remote Sensing and Environmental Monitoring. In the satellite and aerial remote sensing field, data collected by different camera sensors (multispectral or hyperspectral sensors) or over distinct geographic regions typically exhibit significant distributional shifts. These arise from variations in spatial resolution, spectral limitations, atmospheric conditions, and acquisition geometries across receivers such as Sentinel-2, Landsat-8, and WorldView-3 (X. Zhu et al., 2017, Yokoya et al., 2017). As a result, the same semantic category may appear visually distinct depending on environmental or sensor-specific factors, making robust representation learning methods is crucial.

For example, global ecological applications such as bird-habitat classification or wildlife monitoring face geographic and climatic uncertainty. Vegetation structure and canopy appearance differ across biomes, leading to notable domain shifts in species-habitat visualisations (Lyu et al., 2025, Rodrigues et al., 2019). For example, a model trained on European aerial surveys may suboptimally perform when transferred to tropical forests, savannas, or coastal parts, when the target species or habitats remain the same. Similar challenges arise in tasks such as land-cover classification, water quality classification and large-scale climate modelling, where labelled data are scarce but cross-domain heterogeneity is extremely high (Q. Wang et al., 2020, Sobue et al., 2021).

By focusing on robust representations, the proposed methods are well-suited to handling heterogeneity, enabling more reliable performance across multi-sensor, multi-regional, and label-limited remote-sensing scenarios.

3. Security and Biometric Recognition. The metric learning frameworks (FedQuad, TripletFL) developed in this thesis may apply to decentralised biometric authentication systems, where personal identity data has to remain private. The federated parameter selection method (FedMPR), can be useful for person tracking or face recognition systems. By improving the consistency

of same-class (same person) representations, allowing biometric network models to be trained without sharing raw sensitive data.

In summary, this thesis introduces several novel algorithms and methodologies that address scenarios where data are limited, unlabelled, distributed and highly heterogeneous. These contributions are broadly applicable to many real-world scenarios, providing useful solutions for datasets in which centralised data collection or decentralised representation learning.

Bibliography

Acar, D. A. E., Zhao, Y., Navarro, R. M., Mattina, M., Whatmough, P. N., & Saligrama, V. (2021). Federated Learning Based on Dynamic Regularisation. *International Conference on Learning Representations*.

Alqazzaz, A. (2025). Federated learning with homomorphic encryption: A privacy-preserving solution for smart cities. *International Journal of Computational Intelligence Systems*, *18*(1), 304.

Arivazhagan, M. G., Aggarwal, V., Singh, A. K., & Choudhary, S. (2019). Federated learning with personalisation layers. *arXiv preprint arXiv:1912.00818*.

Assran, M., Duval, Q., Misra, I., Bojanowski, P., Vincent, P., Rabbat, M., LeCun, Y., & Ballas, N. (2023). Self-supervised learning from images with a joint-embedding predictive architecture. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15619–15629.

Auñón, J., Hurtado-Ramírez, D., Porras-Díaz, L., Irigoyen-Peña, B., Rahmian, S., Al-Khazraji, Y., Soler-Garrido, J., & Kotsev, A. (2024). Evaluation and utilisation of privacy enhancing technologies—a data spaces perspective. *Data in Brief*, *55*, 110560.

Bachman, P., Hjelm, R. D., & Buchwalter, W. (2019). Learning representations by maximizing mutual information across views. *Advances in Neural Information Processing Systems*, *32*.

Balestriero, R., Ibrahim, M., Sobal, V., Morcos, A., Shekhar, S., Goldstein, T., Bordes, F., Bardes, A., Mialon, G., Tian, Y., et al. (2023). A cookbook of self-supervised learning. *arXiv preprint arXiv:2304.12210*.

Bandara, W. G. C., De Melo, C. M., & Patel, V. M. (2025). Guarding barlow twins against overfitting with mixed samples. *2025 IEEE International Conference on Advanced Visual and Signal-Based Systems*, 1–6.

Bardes, A., Ponce, J., & LeCun, Y. (2022). VICReg: Variance-Invariance-Covariance Regularization for Self-Supervised Learning. *International Conference on Learning Representations*.

Barron, J. T. (2019). A general and adaptive robust loss function. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4331–4339.

Belenguer, A., Pascual, J. A., & Navaridas, J. (2025). Glow a novel, flower-based simulated gossip learning strategy. *arXiv preprint arXiv:2501.10463*.

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation Learning: A Review and New Perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(8), 1798–1828.

Bonawitz, K., Eichner, H., Grieskamp, W., Huba, D., Ingerman, A., Ivanov, V., Kiddon, C., Konečnỳ, J., Mazzocchi, S., McMahan, B., et al. (2019). Towards federated learning at scale: System design. *Proceedings of machine Learning and systems*, *1*, 374–388.

Byrd, D., & Polychroniadou, A. (2020). Differentially private secure multi-party computation for federated learning in financial applications. *Proceedings of the first ACM international conference on AI in finance*, 1–9.

Caldas, S., Duddu, S. M. K., Wu, P., Li, T., Konečnỳ, J., McMahan, H. B., Smith, V., & Talwalkar, A. (2018). LEAF: A Benchmark for Federated Settings. *Workshop on Federated Learning for Data Privacy and Confidentiality*.

Caron, M., Bojanowski, P., Joulin, A., & Douze, M. (2018). Deep clustering for unsupervised learning of visual features. *Proceedings of the European conference on computer vision*, 132–149.

Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., & Joulin, A. (2020). Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems*, *33*, 9912–9924.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., & Joulin, A. (2021). Emerging Properties in Self-Supervised Vision Transformers. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9650–9660.

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A Simple Framework for Contrastive Learning of Visual Representations. *Proceedings of the 37th International Conference on Machine Learning*, 1597–1607.

Chen, T., Kornblith, S., Swersky, K., Norouzi, M., & Hinton, G. E. (2020). Big self-supervised models are strong semi-supervised learners. *Advances in Neural Information Processing Systems*, *33*, 22243–22255.

Chen, W., Chen, X., Zhang, J., & Huang, K. (2017). Beyond triplet loss: A deep quadruplet network for person re-identification. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 403–412.

Chen, X., & He, K. (2021). Exploring simple siamese representation learning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15750–15758.

Chen, X., Xie, S., & He, K. (2021). An empirical study of training self-supervised vision transformers. *ICCV*, 9640–9649.

Chuang, C.-Y., Hjelm, R. D., Wang, X., Vineet, V., Joshi, N., Torralba, A., Jegelka, S., & Song, Y. (2022). Robust contrastive learning against noisy views. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16670–16681.

Chuang, Y.-S., Li, Y., Wang, D., Yeh, C.-F., Lyu, K., Raghavendra, R., Glass, J., Huang, L., Weston, J., Zettlemoyer, L., et al. (2025). Meta clip 2: A worldwide scaling recipe. *arXiv preprint arXiv:2507.22062*.

Coates, A., Ng, A., & Lee, H. (2011). An Analysis of Single-Layer Networks in Unsupervised Feature Learning. *Artificial Intelligence and Statistics*, 215–223.

Cohen, T. S., & Welling, M. (2019). A general theory of equivariance and invariance in neural networks. *International Conference on Learning Representations*.

Cosentino, R., Shekkizhar, S., Soltanolkotabi, M., Avestimehr, S., & Ortega, A. (2022). The geometry of self-supervised learning models and its impact on transfer learning. *arXiv preprint arXiv:2209.08622*.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A Large-Scale Hierarchical Image Database. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 248–255.

Deng, L. (2012). The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, *29*(6), 141–142.

Deng, Y., Kamani, M. M., & Mahdavi, M. (2020). Adaptive personalised federated learning. *arXiv preprint arXiv:2003.13461*.

Dong, N., & Voiculescu, I. (2021). Federated contrastive learning for decentralised unlabeled medical images. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 378–387.

Ericsson, L., Gouk, H., Loy, C. C., & Hospedales, T. M. (2022). Self-supervised representation learning: Introduction, advances, and challenges. *IEEE Signal Processing Magazine*, *39*(3), 42–62.

Ermolov, A., Siarohin, A., Sangineto, E., & Sebe, N. (2021). Whitening for self-supervised representation learning. *International Conference on Machine Learning*, 3015–3024.

Fallah, A., Mokhtari, A., & Ozdaglar, A. (2020). Personalised Federated Learning with Theoretical Guarantees: A Model-Agnostic Meta-Learning Approach. *Advances in Neural Information Processing Systems*, *33*, 3557–3568.

Fang, X., Ye, M., & Du, B. (2025). Robust asymmetric heterogeneous federated learning with corrupted clients. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Fawzi, A., Samulowitz, H., Turaga, D., & Frossard, P. (2016). Adaptive data augmentation for image classification. *2016 IEEE international conference on image processing*, 3688–3692.

Fei-Fei, L., Fergus, R., & Perona, P. (2004). Learning Generative Visual Models from Few Training Examples: An Incremental Bayesian Approach Tested on 101 Object Categories. *Conference on Computer Vision and Pattern Recognition Workshop*, 178–178.

Frankle, J., & Carbin, M. (2019). The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks. *International Conference on Learning Representations*.

Gansekoele, A., Hess, E., & Bhulai, S. (2024). Meta-learning for federated face recognition in imbalanced data regimes. *2nd International Conference on Federated Learning Technologies and Applications*, 24–31.

Gao, L., Fu, H., Li, L., Chen, Y., Xu, M., & Xu, C.-Z. (2022). FedDC: Federated Learning with Non-iid Data via Local Drift Decoupling and Correction. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10112–10121.

General data protection regulation (gdpr), regulation (eu) 2016/679 [Accessed: 2025-10-31]. (n.d.).

Ghosh, A., Chung, J., Yin, D., & Ramchandran, K. (2020). An efficient framework for clustered federated learning. *Advances in Neural Information Processing Systems*, *33*, 19586–19597.

Gidaris, S., Singh, P., & Komodakis, N. (2018). Unsupervised representation learning by predicting image rotations. *International Conference on Learning Representations*.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al. (2020). Bootstrap Your Own Latent: A New Approach to Self-Supervised Learning. *Advances in Neural Information Processing Systems*, *33*, 21271–21284.

Gu, Z., Shi, J., Yang, Y., & He, L. (2023). Defending against adversarial attacks in federated learning on metric learning model. *2023 IEEE 22nd International Conference on Trust, Security and Privacy in Computing and Communications*, 197–206.

Han, B., Zhang, S., Shi, X., & Reichstein, M. (2024). Bridging remote sensors with multisensor geospatial foundation models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 27852–27862.

Han, S., Pool, J., Tran, J., & Dally, W. (2015). Learning Both Weights and Connections for Efficient Neural Network. *Advances in Neural Information Processing Systems*, *28*.

Hard, A., Rao, K., Mathews, R., Ramaswamy, S., Beaufays, F., Augenstein, S., Eichner, H., Kiddon, C., & Ramage, D. (2018). Federated learning for mobile keyboard prediction. *arXiv preprint arXiv:1811.03604*.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., & Girshick, R. (2022). Masked autoencoders are scalable vision learners. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16000–16009.

He, K., Fan, H., Wu, Y., Xie, S., & Girshick, R. (2020). Momentum Contrast for Unsupervised Visual Representation Learning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9729–9738.

He, R., Tong, K., Fang, D., Sun, H., Zeng, Z., Li, H., Chen, T., & Zhuang, H. (2025). Afl: A single-round analytic approach for federated learning with pre-trained models. *Proceedings of the Computer Vision and Pattern Recognition Conference*, 4988–4998.

Health insurance portability and accountability act of 1996 (hipaa) [Accessed: 2025-10-31]. (n.d.).

Hegedűs, I., Danner, G., & Jelasity, M. (2021). Decentralized learning works: An empirical comparison of gossip learning and federated learning. *Journal of Parallel and Distributed Computing*, *148*, 109–124.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., & Hochreiter, S. (2017). GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. *Advances in Neural Information Processing Systems*, *30*.

Hoffer, E., & Ailon, N. (2015). Deep Metric Learning Using Triplet Network. *Similarity-Based Pattern Recognition Workshop*, 84–92.

Horvath, S., Laskaridis, S., Almeida, M., Leontiadis, I., Venieris, S., & Lane, N. (2021). Fjord: Fair and accurate federated learning under heterogeneous targets with ordered dropout. *Advances in Neural Information Processing Systems*, *34*, 12876–12889.

Hu, Q., Wang, X., Hu, W., & Qi, G.-J. (2021). AdCo: Adversarial Contrast for Efficient Learning of Unsupervised Representations from Self-Trained Negative Adversaries. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1074–1083.

Huang, C., Chen, X., Zhang, Y., & Wang, H. (2024). Fedcrl: Personalised federated learning with contrastive shared representations for label heterogeneity in non-iid data. *arXiv preprint arXiv:2404.17916*.

Hubel, D. H., & Wiesel, T. N. (1959). Receptive fields of single neurones in the cat's striate cortex. *The Journal of Physiology*, *148*(3), 574.

Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., & Madry, A. (2019). Adversarial examples are not bugs, they are features. *Advances in Neural Information Processing Systems*, *32*.

Jaiswal, A., Babu, A. R., Zadeh, M. Z., Banerjee, D., & Makedon, F. (2020). A Survey on Contrastive Self-Supervised Learning. *Technologies*, *9*(1), 2.

Jayasumana, S., Ramalingam, S., Veit, A., Glasner, D., Chakrabarti, A., & Kumar, S. (2024). Rethinking FID: Towards a Better Evaluation Metric for Image Generation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9307–9315.

Jeong, W., Yoon, J.-G., & Hwang, S. (2021). Federated self-supervised learning: An application to privacy-preserving visual representation learning. *IEEE Conference on Computer Vision and Pattern Recognition Workshops*.

Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q., Sung, Y.-H., Li, Z., & Duerig, T. (2021). Scaling up visual and vision-language representation learning with noisy text supervision. *International Conference on Machine Learning*, 4904–4916.

Jia, Y., Zhang, X., Hu, H., Choo, K.-K. R., Qi, L., Xu, X., Beheshti, A., & Dou, W. (2024). DapperFL: Domain Adaptive Federated Learning with Model Fusion Pruning for Edge Devices. *Advances in Neural Information Processing Systems*, *37*, 13099–13123.

Jiang, L., Ma, L., & Yang, G. (2025). Shadow defense against gradient inversion attack in federated learning. *Medical Image Analysis*, *105*, 103673.

Jiang, Y., Wang, S., Valls, V., Ko, B. J., Lee, W.-H., Leung, K. K., & Tassiulas, L. (2022). Model Pruning Enables Efficient Federated Learning on Edge Devices. *IEEE Transactions on Neural Networks and Learning Systems*, *34*(12), 10374–10386.

Jimenez-Gutierrez, D. M., Hassanzadeh, M., Anagnostopoulos, A., Chatzigiannakis, I., & Vitaletti, A. (2025). A thorough assessment of the non-iid data impact in federated learning. *arXiv preprint arXiv:2503.17070*.

Johnson, J. E., Sundaresan, S., Daylan, T., et al. (2020). RotNet: Fast and Scalable Estimation of Stellar Rotation Periods Using Convolutional Neural Networks. *Advances in Neural Information Processing Systems*.

Kairouz, P., McMahan, H. B., Avent, B., et al. (2021). Advances and open problems in federated learning. *Foundations and Trends in Machine Learning*, *14*(1–2), 1–210.

Kang, H., Cha, S., & Kang, J. (2025). Gefl: Model-agnostic federated learning with generative models. *IEEE Transactions on Mobile Computing*.

Karimireddy, S. P., Kale, S., Mohri, M., Reddi, S. J., Stich, S. U., & Suresh, A. T. (2020). SCAFFOLD: Stochastic Controlled Averaging for On-Device Federated Learning. *Proceedings of the 37th International Conference on Machine Learning*.

Kawaguchi, K., Deng, Z., Ji, X., & Huang, J. (2023). How does information bottleneck help deep learning? *International Conference on Machine Learning*, 16049–16096.

Khosla, P., Teterwak, P., Wang, C., et al. (2020). Supervised Contrastive Learning. *Advances in Neural Information Processing Systems*, *33*, 18661–18673.

Koch, G., Zemel, R., Salakhutdinov, R., et al. (2015). Siamese neural networks for one-shot image recognition. *International Conference on Machine Learning deep Learning workshop*, *2*(1), 1–30.

Krizhevsky, A. (2009). Learning Multiple Layers of Features from Tiny Images. *Master's Thesis, University of Toronto*.

Kurtuluş, E., Li, Z., Dauphin, Y., & Cubuk, E. D. (2023). Tied-augment: Controlling representation similarity improves data augmentation. *International Conference on Machine Learning*, 17994–18007.

Larsson, G., Maire, M., & Shakhnarovich, G. (2017). Colourization as a Proxy Task for Visual Understanding. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6874–6883.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-Based Learning Applied to Document Recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324.

Lee, G., & Choi, D. (2024). Regularising and Aggregating Clients with Class Distribution for Personalised Federated Learning. *arXiv preprint arXiv:2406.07800*.

Li, A., Sun, J., Wang, B., Duan, L., Li, S., Chen, Y., & Li, H. (2020). Lotteryfl: Personalized and communication-efficient federated learning with lottery ticket hypothesis on non-iid datasets. *arXiv preprint arXiv:2008.03371*.

Li, Q., He, B., & Song, D. (2021). Model-Contrastive Federated Learning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10713–10722.

Li, S., Deng, W., & Du, J. (2017). Reliable Crowdsourcing and Deep Locality-Preserving Learning for Expression Recognition in the Wild. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2852–2861.

Li, T., Hu, S., Beirami, A., & Smith, V. (2021). Ditto: Fair and robust federated learning through personalisation. *International Conference on Machine Learning*, 6357–6368.

Li, X., Liu, M., Sun, S., Wang, Y., Jiang, H., & Jiang, X. (2023). Fedtrip: A resource-efficient federated learning method with triplet regularisation. *2023 IEEE International Parallel and Distributed Processing Symposium*, 809–819.

Liang, P. P., Liu, T., Ziyin, L., Allen, N. B., Auerbach, R. P., Brent, D., Salakhutdinov, R., & Morency, L.-P. (2020). Think Locally, Act Globally: Federated Learning with Local and Global Representations. *Workshop on Federated Learning for Data Privacy and Confidentiality*.

Lin, T., Kong, L., Stich, S. U., & Jaggi, M. (2020). Ensemble distillation for robust model fusion in federated learning. *Advances in Neural Information Processing Systems*, *33*, 2351–2363.

Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., & Song, L. (2016). Large-margin softmax loss for convolutional neural networks. *Proceedings of the 33rd International Conference on Machine Learning*, 507–516.

Liu, Y., Sun, R., et al. (2021). No one left behind: Real-world semi-supervised federated learning with fedmix. *International Conference on Machine Learning*.

Liu, Z., Luo, P., Wang, X., & Tang, X. (2015). Deep Learning Face Attributes in the Wild. *Proceedings of the IEEE International Conference on Computer Vision*, 3730–3738.

Liu, Z., Luo, P., Wang, X., & Tang, X. (2018). Large-scale celebfaces attributes (celeba) dataset. *Retrieved August*, *15*(2018), 11.

Louizos, C., Reisser, M., & Korzhenkov, D. (2024). A mutual information perspective on federated contrastive learning. *International Conference on Learning Representations*.

Lu, J., Zhang, H., Zhou, P., Wang, X., Wang, C., & Wu, D. O. (2024). Fedlaw: Value-aware federated learning with individual fairness and coalition stability. *IEEE Transactions on Emerging Topics in Computational Intelligence*, *9*(1), 1049–1062.

Lyu, S., Zhao, Q., Zhou, Z., Li, M., Zhou, Y., Yao, D., Cheng, G., Zhou, H., & Shi, Z. (2025). Deep learning based domain adaptation methods in remote sensing: A comprehensive survey. *IEEE Geoscience and Remote Sensing Magazine*.

Marathe, A., Walambe, R., & Kotecha, K. (2022). In rain or shine: Understanding and overcoming dataset bias for improving robustness against weather corruptions for autonomous vehicles. *arXiv preprint arXiv:2204.01062*.

McMahan, B., Moore, E., Ramage, D., Hampson, S., & Aguera y Arcas, B. (2017). Communication-Efficient Learning of Deep Networks from Decentralised Data. *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, 1273–1282.

Minka, T. (2000). Estimating a dirichlet distribution.

Mishra, S., Shah, A., Bansal, A., Jagannatha, A., Sharma, A., Jacobs, D., & Krishnan, D. (2022). Object-aware cropping for self-supervised learning. *Transactions on Machine Learning Research*.

Mora, A., Bujari, A., & Bellavista, P. (2024). Enhancing Generalisation in Federated Learning with Heterogeneous Data: A Comparative Literature Review. *Future Generation Computer Systems*.

Movshovitz-Attias, Y., Toshev, A., Leung, S. J., Ioffe, S., & Singh, S. (2017). No fuss distance metric learning using proxies. *Proceedings of the IEEE International Conference on Computer Vision*, 360–368.

Mu, X., Shen, Y., Cheng, K., Geng, X., Fu, J., Zhang, T., & Zhang, Z. (2023). Fedproc: Prototypical contrastive federated learning on non-iid data. *Future Generation Computer Systems*, *143*, 93–104.

Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., & Ng, A. Y. (2011). Reading Digits in Natural Images with Unsupervised Feature Learning. *Workshop on Deep Learning and Unsupervised Feature Learning*, 4.

Nilsback, M.-E., & Zisserman, A. (2008). Automated Flower Classification over a Large Number of Classes. *Indian Conference on Computer Vision, Graphics and Image Processing*, 722–729.

Noroozi, M., & Favaro, P. (2016). Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles. *European Conference on Computer Vision*, 69–84.

Oh, J., Kim, S., & Yun, S.-Y. (2021). FedBABU: Towards Enhanced Representation for Federated Image Classification. *International Conference on Learning Representations*.

Oquab, M., Darcet, T., Moutakanni, T., et al. (2024). DINOv2: Learning Robust Visual Features Without Supervision. *Transactions on Machine Learning Research*.

Ozbulak, U., Lee, H. J., Boga, B., Anzaku, E. T., Park, H., Van Messem, A., De Neve, W., & Vankerschaver, J. (2023). Know Your Self-Supervised Learning: A Survey on Image-Based Generative and Discriminative Training. *Transactions on Machine Learning Research*.

Pang, B., Zhang, Y., Li, Y., Cai, J., & Lu, C. (2022). Unsupervised Visual Representation Learning by Synchronous Momentum Grouping. *European Conference on Computer Vision*, 265–282.

Park, H., Hosseini, H., & Yun, S. (2021). Federated learning with metric loss. *International Workshop on Federated Learning for User Privacy and Data Confidentiality in Conjunction with ICML 2021*.

Park, S., Lee, S., Kim, B., & Hwang, S. J. (2025). Fedrand: Enhancing privacy in federated learning with randomized lora subparameter updates. *arXiv preprint arXiv:2503.07216*.

Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., & Efros, A. A. (2016). Context encoders: Feature learning by inpainting. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2536–2544.

Qi, P., Chiaro, D., Guzzo, A., Ianni, M., Fortino, G., & Piccialli, F. (2024). Model aggregation techniques in federated learning: A comprehensive survey. *Future Generation Computer Systems*, *150*, 272–293.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. *Proceedings of the 38th International Conference on Machine Learning*, 8748–8763.

Reddi, S., Charles, Z., et al. (2021). Adaptive federated optimization. *International Conference on Learning Representations*.

Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H. R., Albarqouni, S., Bakas, S., Galtier, M. N., Landman, B. A., Maier-Hein, O. S., Klaus, Sheller Micah, S. R. M., Trask Andrew, X. D., & Baust Maximilian, C. M. J. (2020). The future of digital health with federated learning. *NPJ digital medicine*, *3*(1), 119.

Robinson, J., Chuang, C.-Y., Sra, S., & Jegelka, S. (2021). Contrastive Learning with Hard Negative Samples. *International Conference on Learning Representations*.

Rodrigues, A., et al. (2019). Deep learning for species distribution modelling. *Ecological Informatics*.

Sattler, F., Müller, S., Markert, T., & Samek, W. (2020). Clustered federated learning: Model-agnostic distributed multi-task optimization. *International Conference on Artificial Intelligence and Statistics*.

Schroff, F., Kalenichenko, D., & Philbin, J. (2015). FaceNet: A Unified Embedding for Face Recognition and Clustering. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 815–823.

Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Mitchell Wortsman, S. P., Kundurthy Srivatsa, C. K., Schmidt Ludwig, K. R., & Jenia., J. (2022). Laion-5b: An open large-scale dataset for training next-generation image-text models. *Advances in Neural Information Processing Systems*, *35*, 25278–25294.

Seo, S., Kim, J., Kim, G., & Han, B. (2024). Relaxed contrastive learning for federated learning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12279–12288.

Shammar, E., Cui, X., & Al-qaness, M. A. (2024). Swarm learning: A survey of concepts, applications, and trends. *arXiv preprint arXiv:2405.00556*.

Shao, H., Liu, C., Li, X., & Zhong, D. (2023). Privacy preserving palmprint recognition via federated metric learning. *IEEE Transactions on Information Forensics and Security*, *19*, 878–891.

Sharma, P., Panda, R., Joshi, G., & Varshney, P. (2022). Federated Minimax Optimisation: Improved Convergence Analyses and Algorithms. *Proceedings of the 39th International Conference on Machine Learning*, 19683–19730.

Siméoni, O., Vo, H. V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., Massa, F., Haziza, D., Wehrstedt, L., Wang, J., Darcet, T., Moutakanni, T., Sentana, L., Roberts, C., Vedaldi, A., ... Bojanowski, P. (2025). DINOv3.

Sobue, S., et al. (2021). Earth observation for climate information. *Remote Sensing*.

Sohn, K. (2016). Improved deep metric learning with multi-class n-pair loss objective. *Advances in Neural Information Processing Systems*, *29*.

Song, Z., Li, W., Jin, K., Shi, L., Yan, M., Yin, W., & Yuan, K. (2022). Communication-efficient topologies for decentralized learning with $o(1)$ consensus rate. *Advances in Neural Information Processing Systems*, *35*, 1073–1085.

T Dinh, C., Tran, N., & Nguyen, J. (2020). Personalised federated learning with moreau envelopes. *Advances in Neural Information Processing Systems*, *33*, 21394–21405.

Tamirisa, R., Xie, C., Bao, W., Zhou, A., Arel, R., & Shamsian, A. (2024). FedSelect: Personalised Federated Learning with Customised Selection of Parameters for Fine-Tuning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23985–23994.

Tan, A. Z., Yu, H., Cui, L., & Yang, Q. (2022). Towards Personalized Federated Learning. *IEEE Transactions on Neural Networks and Learning Systems*, *34*(12), 9587–9603.

Tan, Y., Long, G., Ma, J., Liu, L., Zhou, T., & Jiang, J. (2022). Federated learning from pre-trained models: A contrastive learning approach. *Advances in Neural Information Processing Systems*, *35*, 19332–19344.

Tao, C., Wang, H., Zhu, X., Dong, J., Song, S., Huang, G., & Dai, J. (2022). Exploring the equivalence of siamese self-supervised learning via a unified gradient framework. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14431–14440.

Tian, Y., Sun, C., Poole, B., Krishnan, D., Schmid, C., & Isola, P. (2020). What makes for good views for contrastive learning? *Advances in Neural Information Processing Systems*, *33*, 6827–6839.

Tian, Y., Ke, X., Tao, Z., Ding, S., Xu, F., Li, Q., Han, H., Zhong, S., & Fu, X. (2022). Privacy-preserving and robust federated deep metric learning. *2022 IEEE/ACM 30th International Symposium on Quality of Service*, 1–11.

Turing, A. M. (2007). Computing machinery and intelligence. In *Parsing the turing test: Philosophical and methodological issues in the quest for the thinking computer* (pp. 23–65). Springer.

van den Oord, A., Li, Y., & Vinyals, O. (2018). Representation Learning with Contrastive Predictive Coding. *arXiv preprint arXiv:1807.03748*.

Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., Manzagol, P.-A., & Bottou, L. (2010). Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*, *11*(12), 3371–3408.

Wang, H., Wang, Y., Zhou, Z., Ji, X., Gong, D., Zhou, J., Li, Z., & Liu, W. (2018). Cosface: Large margin cosine loss for deep face recognition. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5265–5274.

Wang, Q., et al. (2020). Deep learning for land cover classification and mapping. *Remote Sensing*.

Wang, T., & Isola, P. (2021). Understanding contrastive representation learning through alignment and uniformity on the hypersphere. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6382–6390.

Wang, X., Huang, Y., Zeng, D., & Qi, G.-J. (2023). CaCo: Both Positive and Negative Samples Are Directly Learnable via Cooperative-Adversarial Contrastive Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Welle, M. C., Poklukar, P., & Kragic, D. (2021). Batch Curation for Unsupervised Contrastive Representation Learning. *International Conference on Machine Learning*.

Wen, D., Jeon, K.-J., & Huang, K. (2022). Federated Dropout: A Simple Approach for Enabling Federated Learning on Resource-Constrained Devices. *IEEE Wireless Communications Letters*, *11*(5), 923–927.

Wen, Y., Zhang, K., Li, Z., & Qiao, Y. (2016). A discriminative feature learning approach for deep face recognition. *European Conference on Computer Vision*, 499–515.

Wu, J., Atito, S., Feng, Z., Mo, S., Kitler, J., & Awais, M. (2024). Rethinking positive pairs in contrastive learning. *arXiv preprint arXiv:2410.18200*.

Wu, Z., Xiong, Y., Yu, S. X., & Lin, D. (2018). Unsupervised Feature Learning via Non-Parametric Instance Discrimination. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3733–3742.

Xiao, H., Rasul, K., & Vollgraf, R. (2017). Fashion-MNIST: A Novel Image Dataset for Benchmarking Machine Learning Algorithms. *arXiv preprint arXiv:1708.07747*.

Xu, H., Fang, J., Zhang, X., Xie, L., Wang, X., Dai, W., Xiong, H., & Tian, Q. (2022). Bag of Instances Aggregation Boosts Self-Supervised Distillation. *International Conference on Learning Representations*.

Xuan, H., Stylianou, A., Liu, X., & Pless, R. (2020). Hard negative examples are hard, but useful. *European conference on computer vision*, 126–142.

Yokoya, N., et al. (2017). Hyperspectral and multispectral data fusion: A comparative review. *Remote Sensing*.

Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., & Wu, Y. (2022). Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*.

Yuan, L., Chen, D., Chen, Y.-L., Codella, N., Dai, X., Gao, J., Hu, H., Huang, X., Li, B., Li, C., et al. (2021). Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*.

Yuan, Y., Chen, W., Yang, Y., & Wang, Z. (2020). In defence of the triplet loss again: Learning robust person re-identification with fast approximated triplet loss and label distillation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 354–355.

Zbontar, J., Jing, L., Misra, I., LeCun, Y., & Deny, S. (2021). Barlow twins: Self-supervised learning via redundancy reduction. *International Conference on Machine Learning*, 12310–12320.

Zheng, M., You, S., Wang, F., Qian, C., Zhang, C., Wang, X., & Xu, C. (2021). ReSSL: Relational Self-Supervised Learning with Weak Augmentation. *Advances in Neural Information Processing Systems*, *34*, 2543–2555.

Zhu, H., Xu, J., Liu, S., & Jin, Y. (2021). Federated Learning on Non-iid Data: A Survey. *Neurocomputing*, *465*, 371–390.

Zhu, W., Liu, J., & Huang, Y. (2023). Hnssl: Hard negative-based self-supervised learning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4777–4786.

Zhu, X., et al. (2017). Deep learning in remote sensing: A comprehensive review. *IEEE Geoscience and Remote Sensing Magazine*.

Zhu, Z., Hong, J., & Zhou, J. (2021). Data-free knowledge distillation for heterogeneous federated learning. *International Conference on Machine Learning*, 12878–12889.

Appendix

Reproducibility Details

The computational experiments were performed on my workstation supplied with an NVIDIA GeForce RTX 3090 Ti (24 GB VRAM capacity) and 32 GB RAM with 16 cores, Ubuntu/Linux operating system. The software environment was managed via Conda, Spyder IDE and the used libraries are shown in Table 2.

To ensure the reproducibility of these results presented in Table 1, random seeds of [e.g., 11, 42, 123] are used for multiple operations, including weight initialisation, data shuffling, and train-test splitting.

Component	**Setting**	**Goal**
Random Seed(s)	11, 17, 27, 42, 123	Baseline consistency
cuDNN Deterministic	True	Prevents algorithm-based variance
cuDNN Benchmark	False	Disables performance-based algorithm switching
Multi-GPU Support	manual_seed_all	Ensures safety for future scaling

Table 1: Detailed hyperparameters and environment specifications for experimental reproduction.

For the CIFAR-10, the average epoch time ranged between 1 and 1.5 minutes (with batch size 128), resulting in a total training time of approximately 3.5 to 5 hours for 200 epochs in the experiments reported in Chapter 3. In contrast, training on the ImageNet requires slightly longer computation, training time, typically taking several days to complete due to the significantly larger dataset size.

For the federated learning experiments described in Chapter 4, with 10 clients, a training round is around 10–15 minutes, while 20 communication rounds resulted in a total training time of about 6–7 hours when using the CIFAR-10. In Chapter 5, the metric learning method (FedQuad) shows improved computational efficiency. In this setting, one round is approximately 4–5 minutes, and 20 rounds of training take approximately 1.5–2 hours for configurations involving a few clients. It is noted that these reported runtimes correspond only to the training and accuracy evaluation stages. Additional processes, such as t-SNE visualisation, confusion matrix plots, detailed plotting, and specific evaluation metrics, are not included in these measurements. For the partial model federated training (FedDINOv3), using the same experimental federated settings, the total training time remained around 1.5–2 hours when scaling to a larger number of clients.

Chapter	Library	Min. Version	Role
3,4,5,6	torch	2.1.0	Core tensor computations
3,4,5,6	torchvision	0.18.0	Image processing, dataset handling
3,4,5,6	numpy	1.25.0	Linear algebra, array manipulations
3,4,5,6	pandas	2.1.0	Dataframe management, CSV handling
3,4,5,6	scikit-learn	1.3.0	Data splitting, evaluation metrics
3,4,5,6	scipy	1.12.0	Signal processing, statistical functions
4	h5py	3.9.0	Large-scale dataset storage (HDF5)
4	pillow	10.0.0	Images and basic manipulations
3,4,5,6	matplotlib	3.8.0	Static plotting, loss curves
3,4,5,6	seaborn	0.12.2	Statistical data visualization
6	timm	Latest Accessing SOTA	pre-trained Vision models

Table 2: Software environment and libraries list for each chapter in the thesis.

CelebA-Gender Dataset

Our assumption in constructing the new dataset is that combining multiple attributes to control the degree of overlap and balance across clients' data distributions introduces greater task complexity. Thus, this setup allows us to carefully evaluate the performance of federated learning models under more realistic and heterogeneous data conditions, where client distributions are diverse but exhibit basic similarities. These scenarios present a more significant challenge for model aggregation algorithms, providing deeper insight into their robustness and generalisation capabilities.

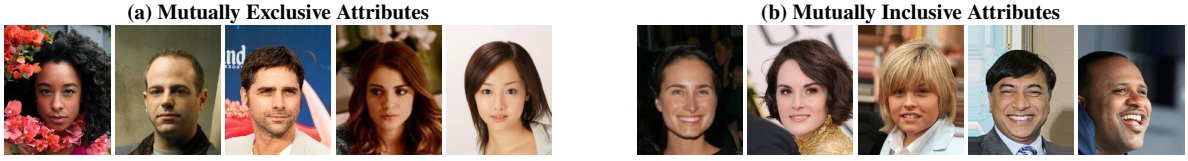
In this paper Gansekoele et al., 2024, proposes an approach to handling non-i.i.d. settings through attribute-specific data partitioning for CelebA data (one attribute, binary classification). In contrast, our work focuses on genders as a classification task, where data heterogeneity arises from the presence or absence of multiple attributes. Therefore, this benchmark enables the construction of more diverse client datasets, instead of relying on the dominance of any single attribute.

CelebA-Gender dataset consists of approximately 46,330 images at a resolution of 178×218 under female/male classes for mutually exclusive and inclusive. Dataset reconstructed with attributes from two to seven. In our experiments, we mostly use the five-attribute-based gender dataset. These attributes are *Black Hair, Smiling, High Cheekbones, Attractive, Mouth Slightly Open*. While multiple attributes can be defined, we focus on five to ensure sufficient overlap per sample. Increasing the number of target attributes reduces the likelihood of their co-occurrence, shrinking the sample space to select and limiting the effectiveness of federated learning.

We show in Chapter 3 that FRD or FID scores can determine whether samples are good or bad. Thus, we evaluate the similarity between client datasets using the Fréchet Inception Distance (FID) (Heusel et al., 2017), where a higher FID score indicates dissimilarity, corresponding to a higher covariate shift among clients. To obtain a more comprehensive understanding of client-by-client data distributional differences, we employ the CLIP-based Maximum Mean Dis-

crepancy (CMMD) metric Jayasumana et al., 2024. Unlike FID scores, CMMD leverages CLIP embeddings to capture high-level features and computes distances using a Gaussian RBF kernel, offering a more discriminative measurement to define data similarity. In both metrics, lower FID and CMMD values correspond to similarity between client data distributions.

When analysing the t-SNE plots in Figures 8, the distribution of mutually exclusive samples with five attributes shows minimal overlap, indicating a more complex data distribution compared to the two-attribute plots shown in Figure 11. Thus, this demonstrates that we can construct datasets with complexity not only in terms of label scarcity but also based on multiple attributes. Therefore, these findings provide a more realistic dataset for evaluating federated methods in many applications. Figure 1 shows one attribute (*Smiling*), figure 2 has two attributes (*Smiling,*



Each sample has NO *Smiling* attribute.

Each sample includes only 1 (*Smiling*) as its specified attribute.

Figure 1: (1 Attribute) Samples from the CelebA-Gender dataset illustrating mutually exclusive (a) and inclusive (b) attribute combinations across gender classes.

Mouth Slightly Open) image samples, figure 3 depicts three attributes (*Smiling, Mouth Slightly Open, High Cheekbones*), figure 4 shows four attributes (*Smiling, Mouth Slightly Open, High Cheekbones, Black Hair*), and figure 5 shows five attributes (*Smiling, Mouth Slightly Open, High Cheekbones, Black Hair, Attractive*) mutually exclusive and inclusive attributes.

Mutually Exclusive: Each sample is annotated with only one target attribute and never with multiple attributes together, yielding a high covariate shift (high-CS) scenario across clients. When each client's dataset contains samples annotated with only one attribute, the diversity of visual or semantic features within that client's data becomes highly constrained. As a result, the underlying input feature distribution differs significantly across clients, although the basic label space remains the same.

Mutually Inclusive Each sample contains specified attributes concurrently, resulting in a low covariate shift (low-CS) condition. In this case, the visual characteristics of the data are well-balanced across clients. Each client obtains samples that collectively represent the full range of attribute variations present in the global dataset. Therefore, the input distributions of different clients are more closely aligned, reducing client drift and eliminating the effects of feature imbalance.

We observe that samples in the mutually inclusive (e.g., clients 1–5) exhibit high visual similarity, reflecting overlapping attribute distributions across clients. In contrast, in the mutually exclusive case, where each client receives distinct subsets of data, the client similarity is

(a) Mutually Exclusive Attributes

Each sample is annotated with only one specified attribute.

(b) Mutually Inclusive Attributes

Each sample has 2 specified attributes.

Figure 2: (2 Attribute) Samples from the CelebA-Gender dataset illustrating mutually exclusive (a) and inclusive (b) attribute combinations across gender classes.

(a) Mutually Exclusive Attributes

Each sample is annotated with only one specified attribute.

(b) Mutually Inclusive Attributes

Each sample has 3 specified attributes.

Figure 3: (3 Attribute) Samples from the CelebA-Gender dataset illustrating mutually exclusive (a) and inclusive (b) attribute combinations across gender classes.

(a) Mutually Exclusive Attributes

Each sample is annotated with only one specified attribute.

(b) Mutually Inclusive Attributes

Each sample has 4 specified attributes.

Figure 4: (4 Attribute) Samples from the CelebA-Gender dataset illustrating mutually exclusive (a) and inclusive (b) attribute combinations across gender classes.

(a) Mutually Exclusive Attributes

Each sample is annotated with only one specified attribute.

(b) Mutually Inclusive Attributes

Each sample has 5 specified attributes.

Figure 5: (5 Attribute) Samples from the CelebA-Gender dataset illustrating mutually exclusive (a) and inclusive (b) attribute combinations across gender classes.

Dataset Attributes	**1**	**2**	**3**	**4**	**5**
1	0	1.9845	2.5237	4.0715	8.0375
2	1.9845	0	0.7359	2.8336	7.9505
3	2.5237	0.7359	0	2.25	7.6385
4	4.0715	2.8336	2.25	0	5.7606
5	8.0375	7.9505	7.6385	5.7606	0

Table 3: FID (ME) scores, among 2 clients' data distribution.

Dataset Attributes	**1**	**2**	**3**	**4**	**5**
1	0	1.6796	2.9084	12.1902	13.9753
2	1.6796	0	1.6581	12.5319	14.5646
3	2.9084	1.6581	0	11.5614	13.7768
4	12.1902	12.5319	11.5614	0	3.6587
5	13.9753	14.5646	13.7768	3.6587	0

Table 4: FID (MI) scores, among 2 clients' data distribution.

significantly decreased, leading to highly heterogeneous data and high covariate shift.

When the covariate shift among datasets is minimal, the datasets demonstrate higher similarity, as reflected by lower FID scores. Thus, this analysis underscores the impact of covariate shifts on dataset similarity in federated settings. Increasing the number of attributes in these cases with high covariate shift reduces the probability of sample selection. Since the dataset becomes sparsely distributed in data space. For example, selecting seven attributes narrows the desirable dataset to highly specific samples, and adding more attributes limits this subset. Therefore, we limited the attribute selection to seven attributes, since including a higher number of attributes yields an insufficient sample size.

When comparing the FID scores reported in Tables 3 and 4, we observe that datasets created with a larger number of attributes yield notably higher FID values compared to those with fewer attributes. This indicates lower similarity among client datasets, reflecting complex data heterogeneity. In other words, when more attributes are combined, the effective data subspace becomes increasingly limited, and each client's samples occupy a smaller, more specific part of the data points. Thus, these samples become more isolated from the global data distribution, leading to complex covariate shift situations and reduced client data overlaps.

Table 5 shows a comparison between mutually exclusive and mutually inclusive CelebA-Gender data subsets across multiple attributes, based on their FID scores. When comparing these datasets with five attributes, the mutually exclusive dataset exhibits a higher FID, indicating a difference between the datasets. *"Smiling vs. Not Smiling"* attributes are shown in Figure 1 as traditional data partitioning for CelebA data. In summary, these results highlight that the CelebA-Gender dataset is significantly more complex than the simpler one-attribute-based usage in previous studies. Figure 6 demonstrates the number of attributes for both female

		ME				
		1	2	3	4	5
MI	1	14.223	13.5451	14.2916	14.9454	16.989
	2	15.1052	14.762	15.5018	15.9695	17.6062
	3	16.6699	16.3822	17.1808	17.8266	19.676
	4	21.6963	20.8346	21.5909	25.3815	27.8452
	5	24.1501	22.7483	23.4875	27.5644	32.2518

Table 5: Attribute-by-attribute mutually exclusive versus mutually inclusive data distribution similarity, measured using FID.

and male, highlighting the attributes with the highest occurrence for each gender. Figure 6 illustrates pairs of attributes that co-occur within the same image. Thus, both male and female samples reach their highest frequency around eight or nine attributes. However, identifying an optimal combination of attributes with the highest number of samples for both genders is quite challenging. For example, selecting any three attributes out of forty attributes, we obtain 9,880 possible combinations. In order to find an efficient and fast solution, we centre our attribute selection on the frequency distribution of attributes within each gender.

To build our gender dataset, firstly, we select the attributes with the highest number of samples for each gender. For example, *"No Beard"* attribute has the largest number of samples among female images. When we aim to create mutually inclusive subsets, this provides a large set of samples. However, we need mutually exclusive subsets; the available samples become seriously limited. Since *"No Beard"* attribute includes over 120k female samples, excluding this attribute (to create a mutually exclusive set) results in losing nearly half of the available data points. Thus, we decided on a much smaller subset when creating datasets based on the attributes with the highest frequencies. To address this limitation, we assume that starting from an attribute with an average frequency across both genders provides flexibility to select new attributes. It allows to obtain a higher possible number of samples for the gender dataset. After running several experiments, we decided that the highest number of samples is obtained when starting with *"Smiling"* attribute, which includes 63,871 female and 33,379 male samples. Following the same process, we selected additional attributes with high sample frequencies for both genders and evaluated different combinations of these attributes. Our main goal is to maximise the total number of samples in the dataset. Following this process, we observed that choosing seven attributes yields an extremely limited number of mutually inclusive samples (e.g., only 318 or 148 samples in total), which is insufficient for local model training. Therefore, we limited the attribute selection to five attributes to maintain a reasonable sample size. After determining the attributes listed in Table 6, we avoided those with notably low sample quantities for females, such as *"Bald"*: (17 samples), *"Goatee"*: (4 samples), and *"Moustache"*: (3 samples), since they provide insufficient data for balanced analysis.

In order to reach the highest possible attribute co-occurrence for a sample and maximise the

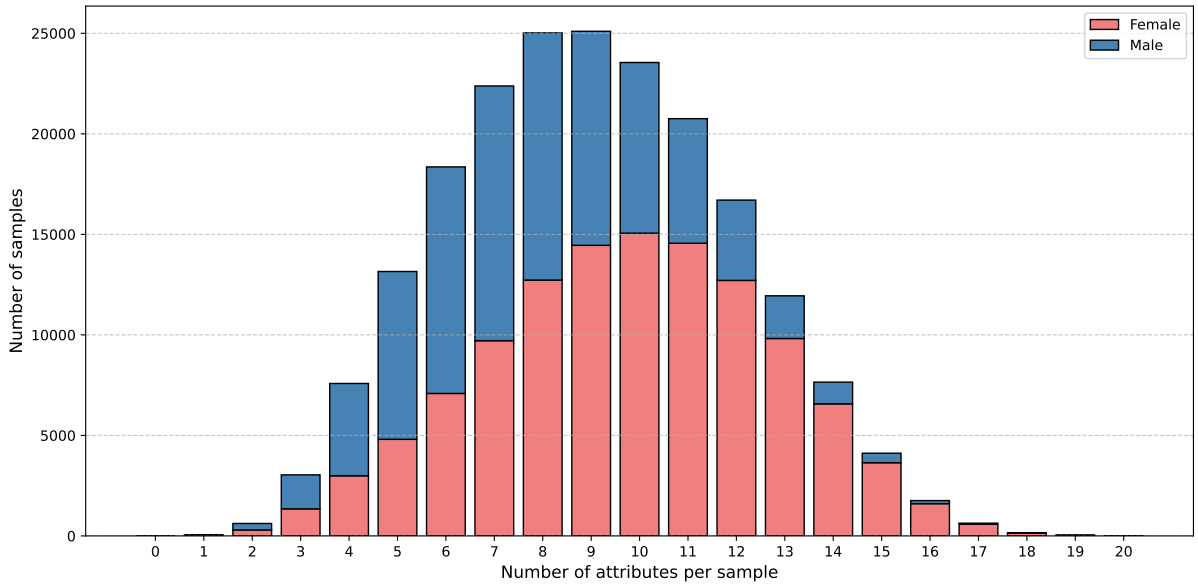


Figure 6: The number of attributes per gender reflects the frequency distribution of samples with different attribute counts, such as the number of samples containing five attributes.

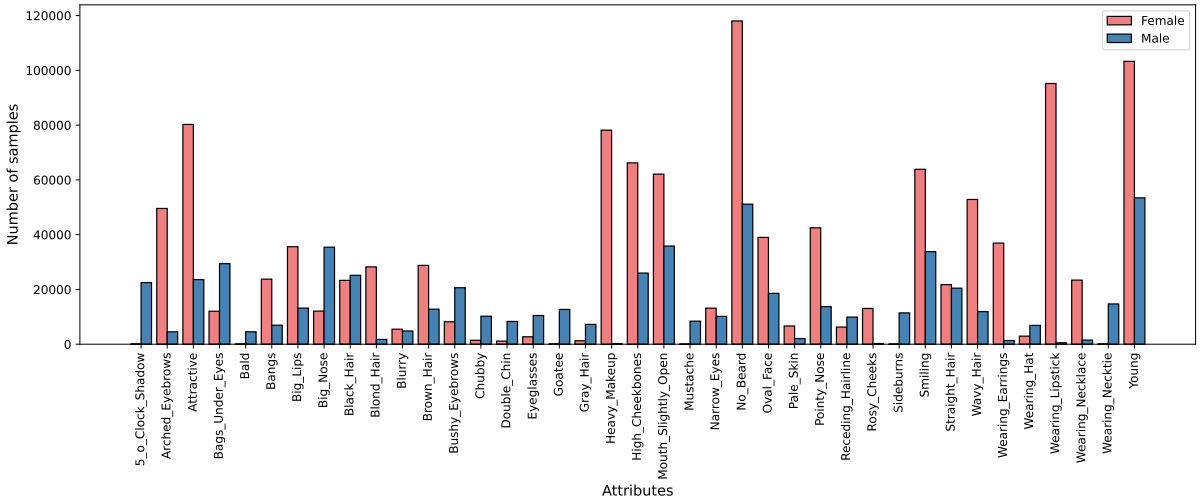


Figure 7: Distribution of attribute frequencies across both genders.

Attribute	**Female**	**Male**
5 o Clock Shadow	20	22496
Arched Eyebrows	49577	4513
Attractive	80254	23579
Bags Under Eyes	12042	29404
Bald	17	4530
Bangs	23759	6950
Big Lips	35606	13179
Big Nose	12085	35431
Black Hair	23316	25156
Blond Hair	28234	1749
Blurry	5479	4833
Brown Hair	28784	12788
Bushy Eyebrows	8181	20622
Chubby	1440	10223
Double Chin	1144	8315
Eyeglasses	2715	10478
Goatee	13	12703
Gray Hair	1264	7235
Heavy Makeup	78156	234
High Cheekbones	66212	25977
Mouth Slightly Open	62096	35846
Mustache	3	8414
Narrow Eyes	13167	10162
No Beard	118026	51132
Oval Face	38997	18570
Pale Skin	6645	2056
Pointy Nose	42498	13712
Receding Hairline	6250	9913
Rosy Cheeks	13061	254
Sideburns	11	11438
Smiling	63871	33798
Straight Hair	21751	20471
Wavy Hair	52852	11892
Wearing Earrings	36927	1349
Wearing Hat	2947	6871
Wearing Lipstick	95192	523
Wearing Necklace	23406	1507
Wearing Necktie	35	14697
Young	103287	53447

Table 6: The list of CelebA attributes and the number of samples for each of the 40 attributes.

Attribute	Female	Male
Mouth Slightly Open, Smiling	**41312**	**40027**
Mouth Slightly Open, Attractive	17631	34475
Mouth Slightly Open, Black Hair	44735	33882
Mouth Slightly Open, High Cheekbones	36855	41337
Mouth Slightly Open, Oval Face	39963	39159
Mouth Slightly Open, Pointy Nose	36027	39732
Mouth Slightly Open, Straight Hair	45533	36725

Table 7: Number of samples (2 attributes at a time) in CelebA-Gender data (ME).

Attribute	Female	Male
Mouth Slightly Open, Smiling, Attractive	14157	28939
Mouth Slightly Open, Smiling, Black Hair	33180	27843
Mouth Slightly Open, Smiling, High Cheekbones	**34541**	**37139**
Mouth Slightly Open, Smiling, Oval Face	32194	33387
Mouth Slightly Open, Smiling, Pointy Nose	26921	32862
Mouth Slightly Open, Smiling, Straight Hair	33491	30422

Table 8: Number of samples (3 attributes at a time) in CelebA-Gender data (ME).

Attribute	Female	Male
Mouth Slightly Open, Smiling, High Cheekbones, Attractive	11973	26532
Mouth Slightly Open, Smiling, High Cheekbones, Black Hair	**28046**	**25963**
Mouth Slightly Open, Smiling, High Cheekbones, Oval Face	27548	31375
Mouth Slightly Open, Smiling, High Cheekbones, Pointy Nose	22697	30275
Mouth Slightly Open, Smiling, High Cheekbones, Straight Hair	27863	28172

Table 9: Number of samples (4 attributes at a time) in CelebA-Gender data (ME).

Attribute	Female	Male
Mouth Slightly Open, Smiling, High Cheekbones, Black Hair, Attractive	**9751**	**19483**
Mouth Slightly Open, Smiling, High Cheekbones, Black Hair, Oval Face	22531	22283
Mouth Slightly Open, Smiling, High Cheekbones, Black Hair, Pointy Nose	18315	21282
Mouth Slightly Open, Smiling, High Cheekbones, Black Hair, Straight Hair	22924	20907

Table 10: Number of samples (5 attributes at a time) in CelebA-Gender data (ME).

Attribute	Female	Male
Mouth Slightly Open, Smiling, High Cheekbones, Black Hair, Attractive, Oval Face	8650	17117
Mouth Slightly Open, Smiling, High Cheekbones, Black Hair, Attractive, Pointy Nose	7337	16594
Mouth Slightly Open, Smiling, High Cheekbones, Black Hair, Attractive, Straight Hair	**8025**	**16266**

Table 11: Number of samples (6 attributes at a time) in CelebA-Gender data (ME).

Attribute	**Female**	**Male**
Mouth Slightly Open, Smiling, High Cheekbones, Black Hair, Attractive, Straight Hair, Oval Face	7145	14297
Mouth Slightly Open, Smiling, High Cheekbones, Black Hair, Attractive, Straight Hair, Pointy Nose	6034	13932

Table 12: Number of samples (7 attributes at a time) in CelebA-Gender data (ME).

number of samples in our dataset, we repeat the same process with mutually inclusive sample selection, ensuring that each attribute's presence or absence is properly represented. When

Attribute	**Female**	**Male**
Mouth Slightly Open, Smiling	**49114**	**25237**
Mouth Slightly Open, Attractive	41816	9466
Mouth Slightly Open, Black Hair	11982	10450
Mouth Slightly Open, High Cheekbones	46998	18726
Mouth Slightly Open, Oval Face	22891	9141
Mouth Slightly Open, Pointy Nose	22456	4856
Mouth Slightly Open, Straight Hair	11215	8608

Table 13: Number of samples (2 attributes at a time) in CelebA-Gender data (MI).

Attribute	**Female**	**Male**
Mouth Slightly Open, Smiling, Attractive	34885	8344
Mouth Slightly Open, Smiling, Black Hair	9776	7758
Mouth Slightly Open, Smiling, High Cheekbones	**44925**	**17168**
Mouth Slightly Open, Smiling, Oval Face	20115	7257
Mouth Slightly Open, Smiling, Pointy Nose	18062	3631
Mouth Slightly Open, Smiling, Straight Hair	9081	6532

Table 14: Number of samples (3 attributes at a time) in CelebA-Gender data (MI).

analysing table 6, we observe that both mutually exclusive and inclusive lead to smaller subsets where the number of attributes increases. Increasing the number of attributes produces findings in a smaller subset of samples. For example, starting with a single female attribute, the sample count is 63,871; however, the sample size drops to 9,751 when combining five attributes in a mutually exclusive scenario in table 10. In the mutually inclusive case in this table 16, the sample count decreases from 33,798 to 2,315 since the number of attributes increases from one to five.

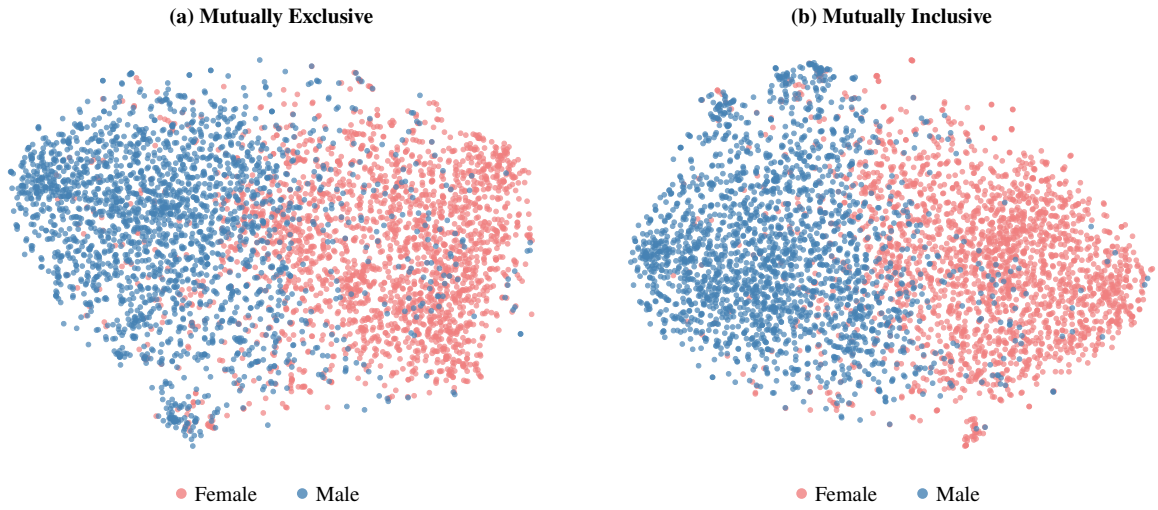
The tables and t-SNE visualisations clearly show that increasing the number of common attributes per sample significantly reduces the size of the selection CelebA subsets. Since the co-occurrence of multiple attributes in one sample is quite minimal, typical methods rely on one attribute at a time. However, the real world is more complex, and we cannot evaluate the dataset (celebA) by selecting one attribute.

Attribute	**Female**	**Male**
Mouth Slightly Open, Smiling, High Cheekbones, Attractive	32385	5454
Mouth Slightly Open, Smiling, High Cheekbones, Black Hair	**9172**	**5538**
Mouth Slightly Open, Smiling, High Cheekbones, Oval Face	18877	5487
Mouth Slightly Open, Smiling, High Cheekbones, Pointy Nose	16814	2170
Mouth Slightly Open, Smiling, High Cheekbones, Straight Hair	8210	4471

Table 15: Number of samples (4 attributes at a time) in CelebA-Gender data (MI).

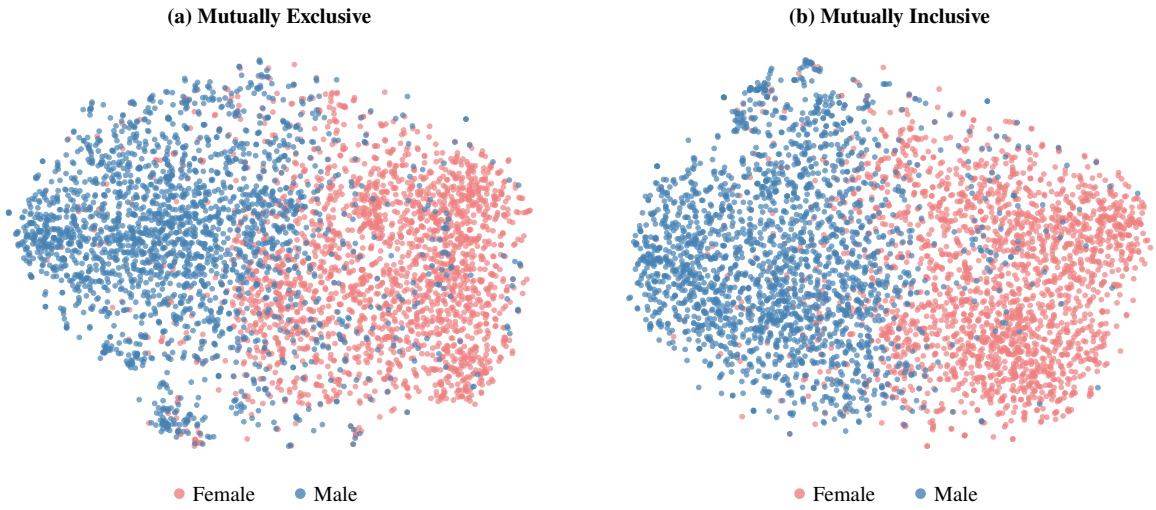
Attribute	**Female**	**Male**
Mouth Slightly Open, Smiling, High Cheekbones, Black Hair, Attractive	**6333**	**2315**
Mouth Slightly Open, Smiling, High Cheekbones, Black Hair, Oval Face	4200	1857
Mouth Slightly Open, Smiling, High Cheekbones, Black Hair, Pointy Nose	2801	647
Mouth Slightly Open, Smiling, High Cheekbones, Black Hair, Straight Hair	1998	2048

Table 16: Number of samples (5 attributes at a time) in CelebA-Gender data (MI).



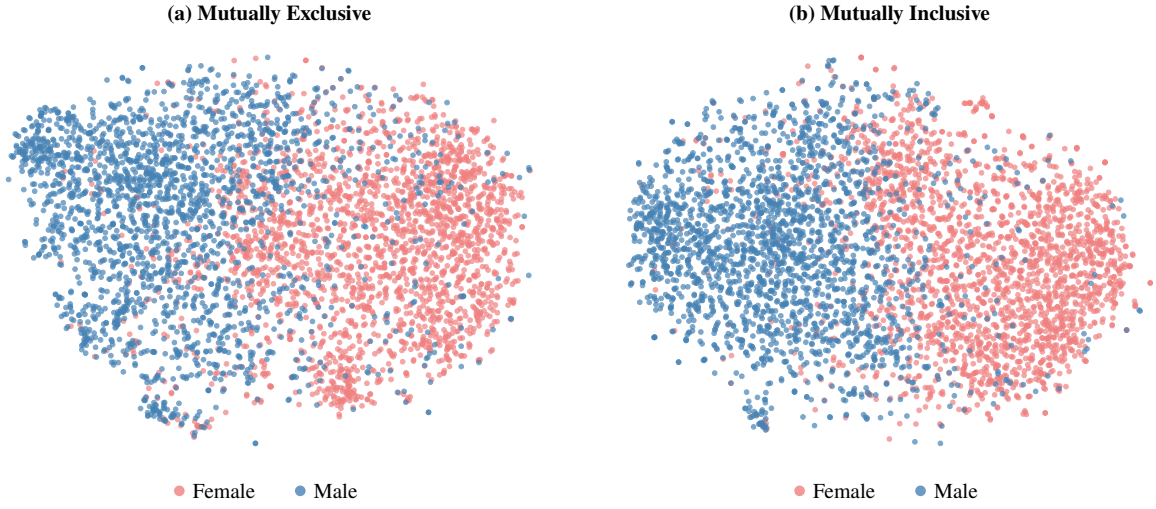
Each sample possesses only one specified attribute. Each sample possesses 2 specified attributes.

Figure 8: t-SNE plots features for 2 attribute data representation for mutually inclusive and exclusive.



Each sample possesses only one specified attribute. Each sample possesses 3 specified attributes.

Figure 9: t-SNE plots features for 3 attribute data representation for mutually inclusive and exclusive.



Each sample possesses only one specified attribute. Each sample possesses 4 specified attributes.

Figure 10: t-SNE plots features for 4 attribute data representation for mutually inclusive and exclusive.

Each sample possesses only one specified attribute. Each sample possesses 5 specified attributes.

Figure 11: t-SNE plots features for 5 attribute data representation for mutually inclusive and exclusive.

Each sample possesses only one specified attribute. Each sample possesses all 6 specified attributes.

Figure 12: Samples from the CelebA-Gender dataset illustrating mutually exclusive (a) and inclusive (b) attribute combinations across gender classes.

Attribute	Female	Male
Mouth Slightly Open, Smiling, High Cheekbones, Black Hair, Attractive, Oval Face	3363	848
Mouth Slightly Open, Smiling, High Cheekbones, Black Hair, Attractive, Pointy Nose	2368	374
Mouth Slightly Open, Smiling, High Cheekbones, Black Hair, Attractive, Straight Hair	**1432**	**1009**

Table 17: Number of samples (6 attributes at a time) in CelebA-Gender data (MI).

(a) Mutually Exclusive Attributes — Each sample possesses only one specified attribute.

(b) Mutually Inclusive Attributes — Each sample possesses all 7 specified attributes.

Figure 13: Samples from the CelebA-Gender dataset illustrating mutually exclusive (a) and inclusive (b) attribute combinations across gender classes.

Attribute	Female	Male
Mouth Slightly Open, Smiling, High Cheekbones, Black Hair, Attractive, Straight Hair, Oval Face	827	318
Mouth Slightly Open, Smiling, High Cheekbones, Black Hair, Attractive, Straight Hair, Pointy Nose	461	148

Table 18: Number of samples (7 attributes at a time) in CelebA-Gender data (MI).